



Shed Light on Unconscious Disclosure and Consumption of Information in the Digital Age.

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik des
Karlsruher Institut für Technologie (KIT)
genehmigte

Dissertation

von

Stephan Cornelius Escher (Dipl.-Medieninf.)
geb. in Schorndorf

1. Referent:

Prof. Dr. Thorsten Strufe

2. Referentin:

Prof. Dr. habil. Simone Fischer-Hübner

Tag der mündlichen Prüfung: 12.02.2025

Abstract

The technological evolution of computers, their networking, and the digitization of information revolutionized the exchange of information fundamentally. Nowadays, almost anyone can easily receive, create, modify and distribute information with almost unlimited reach. The resulting democratized flow of information and the elimination of spatial separation and temporal boundaries enabled a bunch of new possibilities for individuals and societies. However, like any preceding upheaval in human communication, these changes cause also new challenges and issues that must be learned to deal with. On one hand, the unlimited amount of digital information available and the simultaneous creation and dissemination of misleading, false, influencing, and malicious content makes it difficult to assess and verify the credibility of received information. The impacts of such malicious information range from serious issues for individuals to societies. On the other hand, the interactive way of digital information exchange discloses a lot of (privacy-sensitive) information about the user. Information collected and analyzed is used to provide services, personalize information flows, or even to detect and prevent the spread of malicious information and behavior. However, the path between new achievements, safety, and freedom is very narrow. Today's digital information exchange can simultaneously be used for global surveillance of individuals on an unprecedented scale. In the worst case, complete surveillance leads to repression of minorities and unwelcome opinions as well as the establishment of self-censorship, and thus undermines freedom of speech an essential human right and the foundation of modern democracies. Overall, digital information exchange enabled surveillance and manipulation at a low cost. The best solution to these problems would, of course, be systematic prevention. Basically, however, systemic measures of both problems are opposed to each other. The more data is disclosed, the more possible surveillance; the less, the less control over shared information. Hence, in practice, new and old technological developments and their systemic measures are always subject to negotiation processes between safety, freedom, and utility. Thus, existing systemic measures cannot completely protect users from the mentioned issues of digital information exchange.

Transparency and education concerning the consumption and unconscious disclosure of digital information and thus increased awareness of end users is, therefore, an important supplement. On the one hand, to fill the gaps of systemic measures and, on the other hand, to empower users and societies to (co-)determine the negotiation processes themselves - and thus to counteract the new power asymmetries as well as to become part of the solution. This work aims to reveal such gaps for different use cases; to develop transparency solutions for identified gaps; and to evaluate the impact and efficacy of developed solutions. First, we analyze how the traditional exchange of analog information has changed with digitization in terms of verification and unconscious disclosure of information, and develop a transparency tool for the exchange of analog printed documents. Afterward, we investigate the field of new digital developments, with a focus on the IoT. In particular, the integration of small sensors into any physical objects is increasingly blurring the boundary between the digital and analog worlds, and let the information transfer disappear more and more unconsciously in the background. Thus, on one hand, we investigate in detail the change in mobility (connected driving) and, on the other hand, at the impact on bystanders who unconsciously disclose data through surrounding recording sensors without even being an active part of the system. For the latter, we additionally develop and evaluate a transparency solution. In the last part, we investigate the state of news consumption in the German-speaking population and develop and evaluate a solution for contextualizing information to support the assessment of news in social networks.

Contents

1	Introduction	4
2	Digital Information Exchange - A Background	11
2.1	Unconscious Disclosure of Information	11
2.1.1	User Privacy Perceptions and Behavior	13
2.1.2	Countermeasures	15
2.2	Digital Information Consumption	17
2.2.1	User Perceptions and Behavior	18
2.2.2	Countermeasures	19
2.3	Transparency Enhancing Technologies	21
2.3.1	Privacy TETs	21
2.3.2	Assessment TETs	24
2.3.3	Preliminaries	26
3	Analog Information Exchange in the Digital Age	27
3.1	Customary Printer Technologies and their Functionality	29
3.1.1	Laser Printer	29
3.1.2	Inkjet Printer	29
3.1.3	Halftoning	30
3.2	Intrinsic Printer Signatures and their Robustness	31
3.2.1	Existing Intrinsic Signatures for Identification	32
3.2.2	Halftoning as Intrinsic Signature of EP Printer Models	33
3.2.3	Discussion: Impacts of Intrinsic Signatures on Privacy and Verification	38
3.3	Extrinsic Signatures of Printed Documents - Tracking Dots	39
3.3.1	Investigation of Tracking Dots - Background and Approaches	40
3.3.2	Results of Tracking Dot Pattern Analyses	45
3.3.3	An Anonymisation Approach against Surveillance through Tracking Dots	55
3.3.4	Towards a Declaration TET - The DEDA Toolkit	59
3.4	Concluding Remarks & Summary regarding Analog Information Exchange	61

4	Information Disclosure while using or being surrounded by Digital Devices	63
4.1	Privacy-Utility Trade-Off during Connected Driving	64
4.1.1	Pseudonymization of V2X Communication - Related Work	65
4.1.2	The European Pseudonym Scheme	67
4.1.3	Experimental Setup for Evaluating the Scheme	69
4.1.4	Results of the Analysis	73
4.1.5	Transparency for Connected Driving	76
4.1.6	Concluding remarks regarding Connected Driving	77
4.2	Transparency for Bystanders in IoT regarding audiovisual Recordings	78
4.2.1	Bystander Perceptions regarding IoT	78
4.2.2	Towards Transparency: A Background	79
4.2.3	Designing an Audit TET for Bystander in IoT	82
4.2.4	Analysis of Bystander Perceptions regarding the Issue and the Audit TET	88
4.2.5	Concluding Remarks regarding Bystanders in the IoT	99
5	Digital Information Consumption and its Risks for Democratic Societies	100
5.1	News Consumption within the German-Speaking Twitter Community	101
5.1.1	Preliminaries and Related Work	102
5.1.2	Acquisition and Enrichment of the Twitter Data Set	105
5.1.3	The German-Speaking Twitter Community (GTC)	110
5.1.4	News Consumption within the GTC	117
5.1.5	Limitations	124
5.1.6	Concluding Remarks regarding News Consumption within the GTC	125
5.2	Nunti-Score: Supporting users in the Assessment of News Article Previews in OSNs	127
5.2.1	Counter the Impacts of Misinformation - Related Work	128
5.2.2	Assisting the Crowd	129
5.2.3	Nunti-Score Analysis - Hypotheses and Experiment	132
5.2.4	Impact of the Nunti-Score - Experiment Results	137
5.2.5	Concluding Remarks & Discussion of the Nunti-Score	142
6	Discussion and Conclusion	144
7	Thanks	148
8	Declaration of Authorship	149
I	Appendix to Analog Information Exchange	150
A	Benchmarking Toolkits for Perceptual Image Hashing Algorithms	150
A.1	Perceptual Image Hashing	151
A.2	MIHBS - Benchmarking PIH Algorithms	152
A.3	Twizzle - a Multi-Purpose Benchmarking Framework	156
B	Re-Investigation of Banding as Intrinsic EP-Printer Signature	159
C	Manufacturer Statements concerning Tracking Dots	163
D	User Perceptions regarding Tracking Dots	163

E	A new Type of Tracking Dots, a unified Data Set and other Open Challenges	164
II	Appendix to Information Exchange in the IoT	168
A	Theoretical Considerations regarding the V2X EU-Strategy	168
B	Bridging V2X Pseudonym Changes with additional Identifiers	169
C	Open Challenges & Future Work regarding Transparency for Bystander	171
III	Survey Structures	174
A	Analysis of the Privacy Paradox for Mobile and Web Applications	174
B	Nunti-Score - News Items and Distribution	178
	Bibliography	181

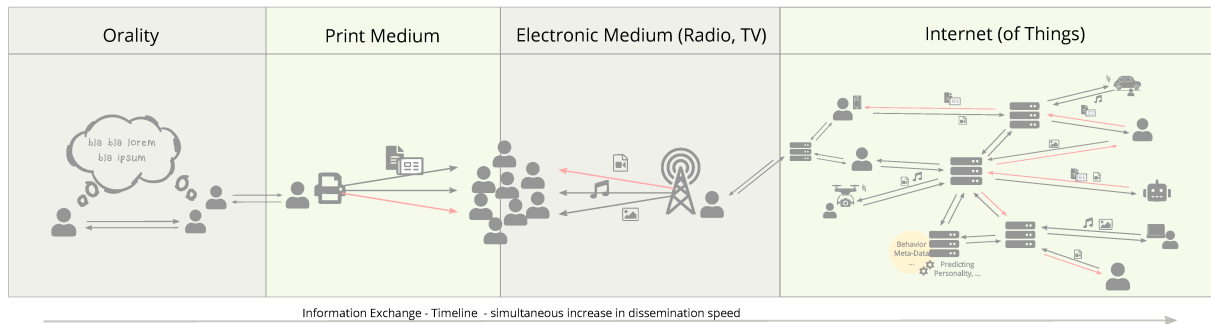
Introduction

Since the beginning of human history, the way information is exchanged among each other has undergone a continuous transformation. From the arise of speech and the accompanying oral transmission and creation of information, e.g. in form of stories, myths, songs, or traditions, to the development of scripture and thus the writing down of oral traditions enabled the exchange of complex information from one generation to the next [346].

The later evolution of printing paved the way for mass (re-)production and distribution of information, knowledge, and ideas and their accessibility to all levels of society. The associated change in the exchange of information enabled cultural, economic, and scientific upheavals and formed the basis for today's societies [125]. However, the creation and dissemination of information, e.g. via books and newspapers, later supplemented by electronic information sources, such as radios or televisions, was overall reserved for only a few. The interactive distribution of information, e.g. via telephones or letters, remained limited in their reach.

The technological evolution of computers today, their networking, and the digitization of information eliminate this asymmetry and thus revolutionize the exchange of information once again fundamentally. Nowadays, almost anyone can easily receive, create, modify and distribute information with almost unlimited reach. The only requirement: a digital device with access to the internet and distribution platforms, like online social networks. The resulting democratized flow of information and the elimination of spatial separation and temporal boundaries – within seconds information can be sent over the entire globe – enable a bunch of new possibilities for individuals and societies. Basically, the dream of "knowledge for all" is being realized through the constant and immediate availability of and the simple and inexpensive access to a myriad of information. However, like any preceding upheaval in human communication, these changes bring new challenges and issues that must be learned to deal with.

The sheer amount of digital information available as well as the democratization of its content, meaning that anyone can create and distribute, complicates the **assessment and verification of received information**, e.g. regarding its credibility [310, 483]. The different presentations of this information through a multitude of possible information channels, the possible diverse accesses to these channels, and the



rapid technological developments amplify the difficulty of assessment and the general understanding of digital information flows. Moreover, available information is far from consisting solely of investigative, researched, and legitimate content and its provision and distribution is not only altruistic but strongly driven by self-interests. Various actors deliberately exploit the uncertainty of end users to influence their opinions or behavior, in particular, due to commercial and/or ideological motives [499]. Hence, the digital information flow integrated into all areas of life and used in almost every daily situation is mixed with misleading, false, influencing, and fraudulent information. Of course, manipulation and deception is not a new invention of the digital world, as, e.g., the Grimm brothers' fairy tale "Little Red Riding Hood" beautifully describes. In the area of digital information consumption, however, the impacts on individuals and societies are far greater, due to the ease of creating, modifying, and disseminating information, its high speed of distribution, and almost unlimited reach, as well as the possibility of automation.

In addition, the establishment of digital information exchange also gives rise to new **power asymmetries**, no longer between information producers and consumers but between end users and providers of the centralized digital infrastructure, its devices, and services. These enable the digital flow of information, but thus also have sovereignty over it and can view, influence, and change it.

This is desirable on the one hand because these providers support to process the wealth of information in a meaningful way. In addition, they can detect and filter horrible, illegal, or malicious information and prevent their impacts. On the other hand, selective filtering of information constantly pushes the boundaries between influence, censorship, and safety, and could lead to restrictions on freedom of speech¹. Moreover, service providers of the internet are mainly driven by commercial interests [509]. Thus, the selective choice of information presented is not primarily based on its quality but on its commercial profit². Due to the prevailing financing model of the internet, which is no longer based on direct payment for a service but on click rates and views, information that matches the user's interest and attracts attention is played out first. This influence exacerbates the problem of assessing information and at the same time creates a new one: the surveillance of end users.

The interactive way of digital information exchange enables not only the analysis of the actual sent/received information content but in addition a lot of **privacy** invasive metadata, e.g. who communicates how, when, with whom, as well as the possibility to infer more information from it. The information collected is used to provide, improve, and secure online services, personalize information flows, or even to detect and prevent the spread of malicious information. At the same time, however, today's digital information

¹https://www.nrk.no/osloogviken/xl/tiktok-doesn_t-show-the-war-in-ukraine-to-russian-users-1.15921522

²<https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-60-minutes-polarizing-divisive-content/>

exchange can be used for global surveillance of individuals on an unprecedented scale. Thereby, the rapid progress of digitization and the ever closer symbiosis between the analog and digital worlds make it more difficult to understand information flows and to know what information is being disclosed and when. Thus, a lot of (sensitive) **information is disclosed unconsciously**. Nowadays >60% of Americans think that they "can't go through daily life without being tracked" by companies or governments, at the same time they are concerned about data usage and feel they have no control [27]. Moreover, this directly affects the first issue of verifying information. The more data is disclosed, the better manipulative information can be personalized and targeted, which increases its impact [219, 452].

In summary, digital information exchange allows surveillance and manipulation at a low cost. The best solution to these problems would be to deal with them systemically so that users of digital systems are not confronted with them at all. In the case of the distribution of malicious and manipulative information, detection algorithms help to block their delivery even before it reaches the user. Verification and authentication methods help to verify legitimate information. Systemic methods of data minimization and data avoidance, so-called Privacy Enhancing Technologies (PETs), are used to preserve the privacy of users when using digital systems.

However, basically, systemic measures of both problems are opposed to each other. The more data is disclosed, the more possible surveillance; the less, the less control over shared information. Hence, in practice, new and old technological developments and their systematic measures are always subject to negotiation processes between



safety (secure digital systems, law enforcement, verification of information, prevention of malicious information), freedom (informational self-determination, protection of personal data, prevention of utter surveillance, freedom of speech), and utility (ensuring that services function sensibly and effectively). Thus, systemic measures cannot completely protect users from the problems of digital information exchange.

Transparency and education concerning the consumption and unconscious disclosure of digital information and thus increased **awareness** and **literacy** of end users is, therefore, an important addition to overcoming these issues. On the one hand, to fill the gaps of systemic measures and, on the other hand, to empower users and societies to (co-)determine the negotiation processes themselves - and thus to counteract the new power asymmetries as well as to become part of the solution.

In this work, we therefore want to contribute to a step towards more transparency and understanding of information flows. To this end, we investigated three specific use cases and examine possible gaps in systemic solutions to find out if there is a need for transparency. For identified gaps we conceptualized and developed transparency approaches in order to support user understanding and awareness within these use cases. In addition, we investigated whether the problems identified are even perceived as such by users from the German cultural sphere and whether the transparency concepts developed are considered a sensible solution. In the following we describe these three use cases and overall the outline of this thesis.

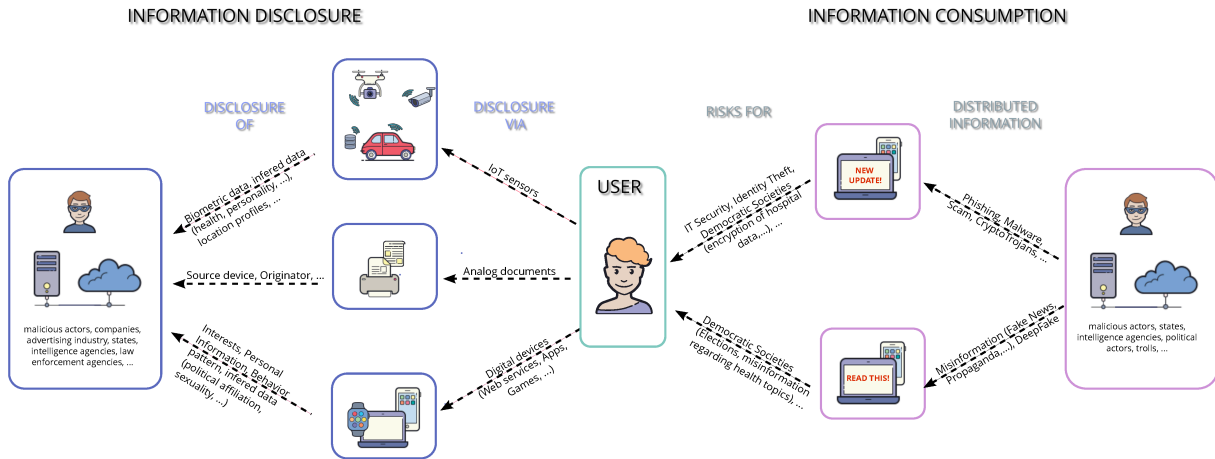


Figure 1.1: Overview of application areas.

Contributions and Outline

Contrary to popular belief, the paperless office has not yet established itself [55]. Therefore, in the first part (Chapter 3) we investigate the unconscious disclosure of information during the **exchange of traditional analog documents**. In particular, we analyzed how technologies for verifying printed documents have changed as a result of digitization and whether this has created a gap between security and privacy.

In the first part of this chapter we investigate *passive printer forensic* approaches, which attempt to extract specific artifacts from printed documents. These artifacts are used as identification features, called *intrinsic signatures*, of specific printer technology, brand, model, or the device itself. Due to digital advances, nowadays such approaches can be automated and used with commercially available recording devices such as scanners or cameras, making forensic examination cheap, easy and accessible to everyone. However, this not only allows the forensic investigation of crimes and verification of documents, but potentially also the arbitrary tracking of authors of analog documents. While **investigating** such approaches, we found that its difficult to find **intrinsic signatures** that are robust against a variety of influences of the printing process and at the same time can identify the specific source device. In most cases, the identification rate only reaches up to the printer model. Overall, this makes passive forensic analysis more difficult but simultaneously prevents the analog document from being traced back to the creator without suspicion and additional information.

The situation is different with a technique called machine identification codes or tracking dots. With this approach, information about the printing process and the exact source device is embedded **extrinsically** into almost every color laser print out while printing, invisible to the human eye. By simply reading out the information, documents can be verified or the producer can be revealed. What is a valuable tool from a forensic point of view, the integration of the exact source device is a severe limitation to the privacy of the users, resulting in a discrepancy between privacy and security. However, in this approach, the embedding of the information, who is responsible, what specific information is integrated, and how it is coded is non-transparent and partly unknown. Thus, we created a data set of 1515 print outs, **investigated** embedded **tracking dots** and developed an extraction algorithm. Overall, we found six different tracking dot patterns, decoded their structure and partly their integrated information. Based on our investigations we developed a **transparency tool** called **deda**. Transparency on one hand for verification of information and on the other hand for privacy to make transparent what information the own printer discloses. In

addition, we developed an anonymization approach to defeat arbitrary tracking and embed it within the deda toolkit. Finally, we briefly analyzed the awareness of users regarding the integration of tracking dots, whether it is perceived as a problem, as well as the need for transparency.

In Chapter 4, we investigate future challenges of digital information exchange. In particular, the establishment of the **Internet of Things** (IoT) and its impacts on the unconscious disclosure of information.

First, we investigate the developments of future mobility. Here, Cooperative Intelligent Transportation Systems (C-ITS) will enable wireless communication between vehicles and the surrounding transport infrastructure, with the objective of safer and more efficient road traffic. Vehicles will regularly broadcast data such as position, speed, or direction to all receivers in the vicinity. In order to verify that the information is not faulty or manipulated, common systems rely on the authentication of sent messages through digital signatures and certificates. With this, however, each vehicle would be directly traceable based on its broadcasted messages, enabling movement profiling and thus limiting location-based privacy. Overall, evolving a privacy vs security gap. Proposals to solve this issue, are fundamentally based on pseudonymization. A vehicle is issued with several certificates, which it changes according to a specific strategy in order to make tracking more difficult. Thereby, it's crucial to use an effective and robust pseudonym-changing scheme as the location privacy of vehicles relies strongly on it. Therefore, we **analyzed a pseudonym change strategy** that is recommended by the **European C-ITS** platform and has a good chance to be included in a future European standard. By simulating a realistic urban traffic scenario within Luxembourg, and applying and attacking the pseudonym change strategy, we could evaluate the effectiveness of the scheme. Our results suggest that the introduction of C-ITS, even with the pseudonym scheme, enables the tracking of vehicles, and thus will decrease location privacy in the future. Overall, the gap between security and privacy is weakened but not solved, which should at least be communicated transparently to the user.

Next to wireless sensors, also visual sensors, like cameras, will be integrated within future vehicles, to enable (semi-)autonomous driving and reacting to the environment. However, the establishment of nearly invisible audiovisual sensors, such as those in smart cars and other devices like smart speakers, or augmented reality glasses, affect not only device owners, but everyone who enters the recording radius of them. Currently, it is nearly impossible for such uninvolved users or bystander to identify surrounding smart devices and corresponding data handling, which leads to foreign control of the bystander's recorded data. Therefore, in the second part of this chapter, we have developed and prototypically implemented a **transparency concept for bystanders**, to enable insights into surrounding, audiovisual smart devices in everyday life and their privacy implications. Further, we conducted a semi-structured interview to analyse whether and how the bystander issue is perceived by participants of the German culture as well as whether our transparency approach is considered as a sensible solution. We verified that there is a high demand for transparency, which, however, is very individual and context-dependent. Moreover, our solution helped to satisfy the bystanders' desire for transparency.

In the last part (Chapter 5), we investigate changes in information consumption due to digitization, especially in the area of **news consumption within social media**. Here, X formerly Twitter frequently is claimed to be a platform for the dissemination of news, with high volumes of campaigning and populism. This claim coincides with the growth of audiences who embrace social media as their primary news

source. Effects like reduction of political education, misinformation, or ideological segregation then arguably represent a major risk for democratic societies. We set out to **investigate** the situation within a comprehensive population with its political system and media market, the **German-speaking Twitter community**. Collecting the entirety of German tweets, we analyze what types of content are disseminated and what types of actors are involved. Our data covers ~ 77 million tweets from ~ 7 million users, collected within 2 months, involving the European Parliament election 2019. It contains observable artifacts corresponding to major political events and exposes clear, artificial dissemination structures around news outlets and populist parties from the political right wing. The results also indicate that not only political actors but content providers, too, have established highly influential profiles that are heavily engaged in political discourse.

The rise of fake news and populism online increases the demand for tools to counter misinformation and better inform users about the credibility of posts on social media. Effective information with high user acceptance is anticipated to increase competence in assessing credibility and assigning trust to posts encountered online. Misjudgment due to the increasing difficulty to evaluate the credibility of news sources can lead to societal risks, even to the point of threatening democratic societies, given the increasingly incidental information acquisition of citizens on the internet. Thus, we finally developed a **transparency concept for the assessment of information**, which enriches news article previews with context information. It visualizes the information quality of linked news articles with a rating, which is based on automatically extracted background information. We investigate the utility and acceptance of this approach. Based on results from a quantitative online experiment with 455 participants, we obtained indications that users were better able to judge the credibility of news articles in OSNs with higher certainty. Additional feedback confirmed that transparency and comprehensibility of the rating were fundamental for its acceptance.

List of Publications

- Jan Ludwig Reubold, Stephan Cornelius Escher, Christian Wresnegger, and Thorsten Strufe. *How to protect the public opinion against new types of bots?*. In International Conference on Big Data, 2022.
- Jan Reubold, Stephan Escher, Johannes Pflugmacher, and Thorsten Strufe. *Dissecting chirping patterns of invasive Tweeter flocks in the German Twitter forest*. In Online Social Networks and Media (OSNEM), 2022.
- Stephan Escher, Katrin Etzrodt, Benjamin Weller, Stefan Köpsell, and Thorsten Strufe. *Transparency for Bystanders in IoT regarding audiovisual Recordings*. In 6th International Workshop on Security, Privacy and Trust in the Internet of Things (SPT-IoT) In conjunction with IEEE Percom, 2022.
- Stephan Escher, Markus Sontowski, Knut Berling, Stefan Köpsell and Thorsten Strufe. *How well can your car be tracked: Analysis of the European C-ITS pseudonym scheme*. In IEEE 93rd Vehicular Technology Conference (VTC-Spring), 2021.
- Stephan Escher, Patrick Teufert, Lukas Hain, and Thorsten Strufe. *You've got nothing on me! Privacy Friendly Face Recognition Reloaded*. In Proceedings of the 3rd International Workshop on Multimedia Privacy and Security (MPS), 2020.
- Stephan Escher, Patrick Teufert, Robin Hermann, and Thorsten Strufe. *Twizzle - A Multi-Purpose Benchmarking Framework for Semantic Comparisons of Multimedia Object Pairs*. In Proceedings of the 3rd International Workshop on Multimedia Privacy and Security (MPS), 2020.

- Johannes Pflugmacher, Stephan Escher, Jan Reubold, Thorsten Strufe. *The German-Speaking Twitter Community Reference Data Set*. In *Proceedings of the 12th International Workshop on Hot Topics in Pervasive Mobile and Online Social Networking (HotPOST)*, 2020.
- Stephan Escher, Benjamin Weller, Stefan Köpsell, and Thorsten Strufe. *Towards Transparency in the Internet of Things*. In *Annual Privacy Forum (APF)*, 2020.
- Timo Richter*, Stephan Escher*, Dagmar Schönfeld, Thorsten Strufe. *Forensic Analysis and Anonymisation of Printed Documents*. In *Proceedings of 6th ACM Workshop on Information Hiding and Multimedia Security (IH& MMSec)*, 2018.
- Stephan C. Escher, Jan L. Reubold, Richard Kwasnicki, Joachim Scharloth, Lutz M. Hagen and Thorsten Strufe. *Towards Automated Contextualization of News Articles*. In *MIS2: Misinformation and Misbehavior Mining on the Web, WSDM Workshops*, 2018.
- Jan Reubold, Stephan Escher, and Thorsten Strufe. *The Latent Behavior Space – A Vector Space for Time-Series Data*. In *Proceedings of the Workshop on Time-Series at the International Conference of Machine Learning (ICML)*, 2017.
- Stephan Escher, Thorsten Strufe. *Robustness analysis of a passive printer identification scheme for halftone images*. In *IEEE International Conference on Image Processing (ICIP)*, 2017.
- Stephan Escher, Stefan Köpsell. *Durchführung eines integrierten Anti-Phishing-Trainings*. In *Sicherheit in vernetzten Systemen: 23. DFN-CERT Konferenz*. DFN-CERT, 2016.

Collaborations

While I am the sole author of this thesis, its content is the result of extensive discussions with students, coworkers, collaborators, and my supervisor Prof. Dr. Thorsten Strufe as well as Dr. Stefan Köpsell. The printer forensic part was accompanied by Thomas Gloe and Jakob Hasse from the forensics company Dence. The DEDA Toolkit as well as the yellow dot dataset have been used and extended by students that i worked with, namely Timo Richter, Julius Wenzel, and Anita Fritzsche. Large parts of the extraction component was implemented by Timo Richter. For the extended dataset, the artist Wolfgang Plöger helped with his open call. Robin Herrmann and Anika Borchmann helped to build the benchmarking tools and analyze perceptual image hashing algorithms. In the IoT area, Markus Sontowski, Richard Stosch, and Knut Behrling have been involved in the analysis of V2X pseudonymization. The development of the IoT bystander TET was supported by Benjamin Weller, as well as communication scientist Katrin Etzrodt for the qualitative analysis. The survey regarding web and mobile usage, tracking knowledge, and used countermeasures was realized together with Thuy Nga Thi Pham. The whole misinformation area was in close cooperation and collaboration with Jan Reubold. The awesome discussions with the ITAS Deepfake team namely Jutta Jahnel, Reinhard Heil, and Olivia Hägele were also very helpful in this area. Various metrics and algorithms regarding the extraction and analysis of Twitter data have been implemented by Johannes Pflugmacher. The development and evaluation of the Nunti score was supported by Felix Kienhöfer and Anne Trinkaus as well as Prof. Dr. Sebastian Pannasch for setting up the psychological experiment.

I use these collaborative results in agreement with my collaborators. To indicate that part of the results presented in this thesis are the outcome of collaborations, I use the pronoun ‘we’ instead of ‘I’ throughout the remainder of this thesis.

2

Digital Information Exchange - A Background

In this chapter, we want to describe the motivation and background regarding the issue of unconscious disclosure as well as regarding verifying information that arises from the development of digital information exchange. We describe technical and regulatory approaches which try to tackle the arising problems and their gaps. Further, we describe related work concerning the perceptions and behavior of end-users regarding the issues. Finally, we highlight solutions of transparency and awareness related to this issues.

2.1 Unconscious Disclosure of Information

Whereas 35 years ago a simple census in germany was accompanied by protests from citizens for reasons of data protection¹, the today's integration of digital information flows into all areas of life takes place largely without controversy. Of course, compared to a census, the digital disclosure of information also offers a bunch of new possibilities and developments. Simultaneously, however, these new possibilities also reveal far more sensitive data. Nowadays, when using digital, networked systems, a lot of information is exchanged between the communication partners as well as with the necessary service providers of these systems and the digital infrastructure. Disclosed data could thereby be categorized as explicit, implicit and predicted data [329].

Explicitly exchanged information refer to conscious sent/received information content, intentionally provided by the user. This includes, for example, social media posts, e-mail and messaging content, registration information, search queries, user profile information, purchases made in e-commerce services, or location data for navigation services.

In addition, the interactive nature of digital information exchange leads to a rather unconscious disclosure of *implicit* data. This includes, for example, meta-data of the communication such as IP address (incl. geolocation), time stamps, url referrers² (where do I come from), user device characteristics, integrated identifiers such as cookies or advertiser ids, or meta-data of the content, like Exif data of images [167]. Further implicit information are disclosed intrinsically, like biometric data within exchanged multimedia

¹<https://www.bpb.de/politik/hintergrund-aktuell/248750/volkszaehlung-1987-22-05-2017>

²<https://www.eff.org/deeplinks/2015/01/healthcare.gov-sends-personal-data>

files, or behavioral data and habits through communication characteristics (e.g. who communicates with whom, which services/domains are visited). With the progressive digitalization of all areas of life, the implicit disclosure of information is also increasing. This is particular evident in the establishment of the Internet of Things (IoT), i.e., the continuous integration of tiny, unobtrusive, and networked sensors in all kinds of physical objects that measure and monitor their environment. Additional, rather unconscious, disclosed information ranges from location data through mobile device sensors, health data through wearable sensors, biometric data through the use or surrounding of audiovisual smart devices, to motion data through tactile applications [200].

At the same time, big data and advances in machine learning are increasing the possibilities for *predicting* and *inferring* further information from disclosed explicit and implicit data. Sensitive demographic attributes, like age, gender, or ethnicity of a user could be predicted based on search queries [43]; browsing behavior [213]; used language [387], likes, or public attributes in social media [43, 72]; smartphone settings [389]; network traffic [275]; or biometric data in images and videos [101]. Moreover, highly sensitive information such as personality characteristics [234, 488], political and religious views [253, 43], health conditions [265, 409] (e.g. depression out of instagram photos [370]), sexuality [470], or mental states and emotions [493] could be predicted, for example based on motion behavior [254], biometric data [253], appliance data [442], or communication behavior [416]. And, overall, disclosed information can be used to track and identify users even if they are not directly authenticated at the used online services [58]. Collected and predicted data is used in ways that can impact our reality [410]. On one hand, information is collected and analyzed to provide and improve online services and functionalities, personalize information flows, or even to detect and prevent the spread of malicious information/behavior and other crimes [505, 99]. On the other hand, today's digital information exchange can simultaneously be used for global surveillance of individuals on an unprecedented scale. In the process, a very lucrative data-driven economy has built up, often named surveillance capitalism [509]. In 2019 alone, the value of data sold in EU countries amounted to 75.3 billion euros³. On the basis of personalized profiles, content and in particular advertising is targeted at users [473, 413]. Collected and predicted data is also forwarded and sold to third parties for this purpose, advertising spaces are sold in real-time (Real Time Bidding) based on the user's characteristics⁴. This enables the usage of online services without having to pay real money, but also entails new risks, such as price steering or discrimination [201, 413], assessed financial credibility (based on predicted data which does not have to be correct⁵), risk assessment for insurances, impacts on the job market [58], or impacts of social scoring ideas⁶. Moreover, not only companies but also governments are interested in such information. This is shown in various surveillance programs of secret services⁷, data retention of different countries⁸, up to the scoring systems of citizens [148]. In the worst case, complete surveillance leads to repression of minorities and unwelcome opinions⁹, as well as the establishment of self-censorship, and thus undermines freedom of speech an essential human right and the foundation of modern democracies. Moreover, the loss of such recorded data leads to the possibility that third (malicious)

³<https://www.iwd.de/artikel/der-datenmarkt-waechst-rasant-482082>

⁴<https://netzpolitik.org/2023/microsofts-datenmarktplatz-xandr-das-sind-650-000-kategorien-in-die-uns-die-online-werbeindustrie-einsortiert>

⁵<https://algorithmwatch.org/en/schufa-a-black-box-openschufa-results-published>

⁶<https://www.engadget.com/2020-01-17-your-online-activity-effectively-social-credit-score-airbnb.html>

⁷<https://www.theguardian.com/us-news/the-nsa-files>, <https://heise.de/-/7157752>

⁸<https://netzpolitik.org/2019/vorratsdatenspeicherung-in-europa-wo-sie-in-kraft-ist-und-was-die-eu-plant>

⁹<https://www.theguardian.com/world/2013/sep/08/shi-tao-china-frees-yahoo>

actors can misuse the data for their own purposes¹⁰. Impacts range from manipulation, influence, extortion, identity theft, to targeted attacks like spear phishing. This is for example demonstrated by the Cambridge Analytica data scandal. Based on illegally acquired Facebook data, psychological profiles of 87 million users were predicted and used, i.a., for the US election campaign in 2016 to influence (swing) voters in their voting decisions [452, 218].

2.1.1 User Privacy Perceptions and Behavior

Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say.

Edward Snowden

Next to well known unconcerned statements of users, like the 'I have nothing to hide' argument [237, 411], which could be explained by a lack of understanding of potential consequences [371, 404, 465], studies show that in general internet users are concerned about their disclosure of personal information and have a need to protect their privacy [294, 91, 90, 93, 463, 333, 24]. In mid 2019, the Pew Research Center, for example, conducted a survey with 4.272 U.S. adults, where a majority reported to be concerned about the personal data collection and usage by companies (79%) or the government (64%) [27]. Moreover, they think that they 'can't go through daily life without being tracked' by them. However, their behavior within digital environments often does not reflect these concerns. This contradiction is referred to in the literature as *privacy paradox* [6, 339, 37].

There exist different theories which try to explain this controversy [251]. The *privacy calculus* theory states that individuals weigh the risks and losses from information disclosure versus expected benefits. The corresponding behavior is the result of this privacy trade-off. I.e., users decide to use a service/application; respectively to disclose personal information; when expected gains exceed expected losses of privacy. Considering the right decision parameters, e.g. social interaction rewards, the disclosing behavior becomes more comprehensible and thus is no longer paradoxical. Benefit considerations range e.g. from time or monetary savings, self-enhancement, maintaining social relationships, trying out novelties, to simply the need for entertainment and pleasure [463, 251]. Risk considerations depend for example on the disclosed data type, trust towards the service/application, social norms, the specific situation (e.g. private vs public spaces), or the purpose of data processing [463, 333, 453].

A survey regarding web and mobile usage, tracking knowledge and used countermeasures, which we conducted in October 2019 with 137 participants (Appendix A), shows for example different perceptions regarding the disclosure of one and the same data type (location) used for different application purposes (Fig. 2.1). While one application purpose is felt very useful the other is felt rather creepy. Overall, the trade-off decision if and when someone feels (un-)comfortable regarding personal data disclosure is, through many decision parameters, very individual and context dependent [474, 333, 251, 5, 117], as well as different across cultures [322]. However, if asked in general, nowadays, users felt that potential risks of data collection outweigh the benefits [27], which is also reflected in our survey (Fig. 2.2).

The individual trade-off decision is additionally influenced by *social factors*; e.g. 'belonging to a group outweighs the risk of disclosing sensitive data' [463]; i.e. to maintain my social life, for some decisions, I

¹⁰<https://tcn.ch/3Q1PgOz>, <https://tcn.ch/45LV8nS>

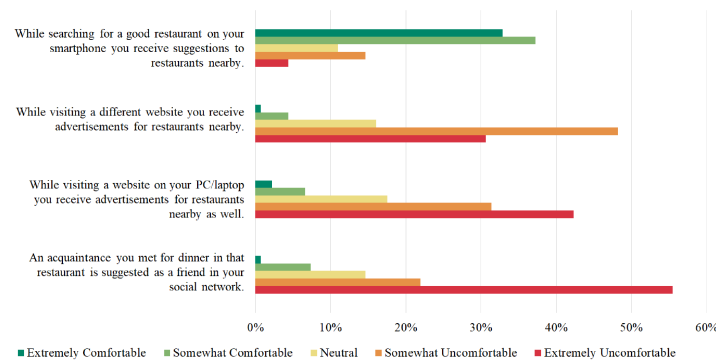


Figure 2.1: User perceptions regarding different application purposes of location disclosure. 137 participants, mainly aged 21-30 years (76.64%)

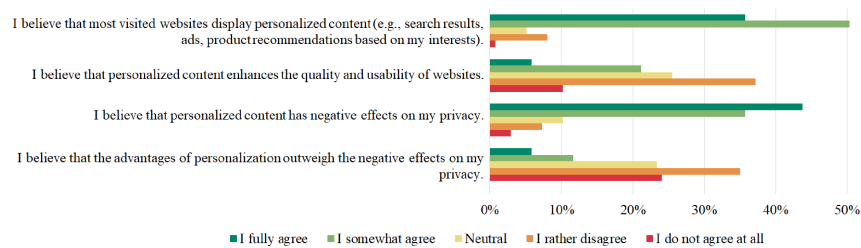


Figure 2.2: User perceptions regarding personalization of web content and its impacts on privacy. 137 participants, mainly aged 21-30 years (76.64%)

have to disclose data, despite my privacy concerns [251].

Another explanation for the privacy paradox is, that the trade-off decision in practice is, compared to self-reported concerns in surveys, influenced by *heuristics and cognitive biases*. This includes for example optimism bias (I'm not so affected by privacy risks as others - underestimation of risks), affect heuristics (decisions based on affective impressions) or benefit heuristics (overestimation of benefits).

In addition, the decision of many people is based on **limited knowledge and misconceptions** [237, 444, 463, 371, 404, 465, 474, 490, 384]. These include a lack of general understanding of digital information flows, the concrete disclosure of information and their sensitivity, as well as regarding possible solutions and tools [379, 193, 490, 24]. Due to the lack of knowledge about these areas, it becomes very difficult for end users to make correct assessments of the associated risks [237, 379]. Incorrect user mental models can also lead to certain privacy fatalism; in the sense: 'it will be monitored anyway'; 'nothing is 100% safe' [379, 171, 3]. As a result, possible solutions are considered unnecessary.

Overall, misconceptions and limited knowledge are reinforced by a large **information asymmetry** which exists between app/service providers and the user [251]. This is manifested by a vast collection, processing and storing of personal information on the one site and little to no knowledge about these practises on the users' site due to missing transparency [278]. Using an online service or application, it is for the user not really clear which information is collected, how and for what it is used, processed, shared, and stored. Thus, in practice, trade-off decisions between risk and benefit are not balanced and therefore no real informed consent takes place. Overall, this reflects the need for solutions to increase transparency.

2.1.2 Countermeasures

This extensive data disclosure in the use of digital systems and its impacts, have lead to an increased demand for privacy protection mechanisms. Thereby, technical privacy protection mechanisms aim primarily at confidentiality, i.e., data minimization and the protection of personal and sensitive data against unauthorised access and use, without disrupting service functionality [457]. Such technical approaches are summarized under the term **Privacy Enhancing Technologies** (PETs). Examples include generalization, obfuscation or pseudonymization of *direct* (IP, face), or *quasi* (age, gender) *identifiers* within data sets [200]; cryptographic encryption of information to protect its content from unauthorized access by providers or actors in the digital infrastructure; anonymous decentralized peer-to-peer (P2P) [374], or mix networks [382] to prevent the disclosure of transmission information; or cryptographic algorithms for privacy-friendly processing of information on a pseudomized/anonymized basis, like differential privacy [120], homomorphic encryption [4], or zero-knowledge proofs [66] to ensure information processing and thus services while maintaining data privacy.

In practice, however, effective data protection mechanisms are seldom used within internet services. Data protection mechanisms are often in conflict with commercial interests and, on the other hand, they tend to make applications more complex and computationally intensive. Basically, any PET poses a trade-off between privacy and functionality (privacy-utility trade-off). The fewer concrete data points, the lower the functionality, the lower the protection, the easier the re-identification [109].

Next to technical solutions, regulatory approaches regarding the collection and processing of personal data were devised in order to protect user's informational self-determination rights. Inspired by the work of Westin [480], the first influential data protection law was the US Privacy Act¹¹, introduced in 1974 [267]. In 1995, the EU Directive 95/46/EC [144] was passed, which i.a. defines that personal data of individuals may not be collected without their *explicit consent*. In 2016, this was replaced by the General Data Protection Regulation (GDPR) [145], implemented in all EU states by 2018. Here, i.a., the following definitions were introduced¹²:

'*Personal data* means any information relating to an identified or identifiable natural person (*Data Subject (DS)*); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier [...]'. *Data Controller (DC)* 'means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data [...]'. *Data Processor (DP)* 'means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.'

The GDPR specifically requires that when personal data is collected, the data subject must be provided with information about the data controller as well as insight into the data processing and storage. This includes for example the purposes of the recording, the transfer of personal data to third parties or countries, retention time, or whether automated decision-making is involved. In addition, the data subject is granted the right to access, rectify or delete collected personal data. Thereby, the transmission of this information should take place in a 'concise, transparent, intelligible and easily accessible form, using clear and plain language'. This leads to the fact that at least users in the EU are generally informed about

¹¹<https://www.justice.gov/opcl/privacy-act-1974>

¹²<https://gdpr-info.eu/art-4-gdpr/>

the data collection and its processing before the use of digital services or applications via *privacy policies*. Hence, in theory they are enabled to make an informed decision whether they want agree with the reported data collection / processing and want to use this service.

However, in practice users rarely tend to read privacy policies, mainly because these are long and complex, basically legal documents [429, 358, 363, 149, 342]. Overall, they are difficult to grasp and time-consuming due to their scope [279]. Back in 2008, McDonald and Cranor [304] calculated that it would take 40 minutes per day to read all policies of visited websites. Nowadays the number of services has strongly increased and data is shared with third parties with their own additional policies. Thus, users give explicit consent to data collection without being really informed about it, to use the services functionality [420, 342]. Even if users have the option to explicitly opt-in/opt-out to data collection and sharing, implemented e.g. via cookie banners, they are deceived by questionable visual designs, called dark pattern or deceptive designs, to lead to decisions that are contrary to the actual informational self-determination [288, 341, 302, 292]. The poor impact of the current implementation of transparency through cookie banners was also reflected in our survey (Fig. 2.3). Even though regulative proposals (Digital Services Act [94]) trying to fix this issue, the regulation is rather lagging behind.

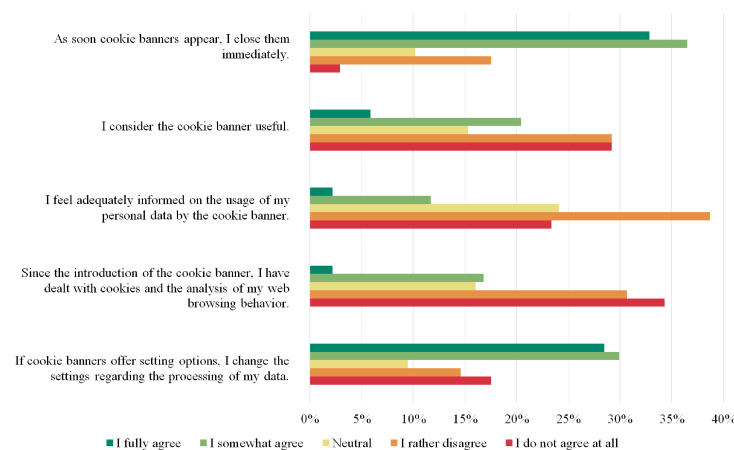


Figure 2.3: Users perceptions and dealing with cookie banners. 137 participants, mainly aged 21-30 years (76.64%)

In addition, some work questions whether the GDPR may not be concise enough to deal with new developments, especially with the complexity of the Internet of Things (IoT). Due to the number of stakeholders and their complex relationship amongst each other, concerns about the GDPR were noted that the framework for responsibilities may not be sufficiently defined [468, 280]. This is shown esp. in the relationship between data controller and data processor.

Furthermore, privacy policies show only intended/reported data collection practices and not the actual data disclosure. After approval, the disclosure of data mainly disappears in the background and remains rather invisible to the normal user. Malicious behavior of services, e.g. if the user has fallen for dark patterns or contrary behavior to policies [343], is not noticeable.

Overall, even though regulatory approaches provide a suitable baseline for achieving informed consent before information is disclosed, as well as the possibility to view and delete disclosed data, the **information asymmetry** between user and service remains largely in place.

2.2 Digital Information Consumption

In addition to its impact on privacy, the advance of digitization has also drastically changed the way we consume information. With a variety of different devices, like PCs, laptops, smartphones, -speaker, -glasses or other wearables, nowadays an almost infinite amount of digital information can be accessed at any time and from anywhere. This includes for example scientific and factual knowledge, political or institutional statements, philosophies of faith and life, tutorials, documentations, entertainment, private conversation, daily news, financial or health advice, opinions, reviews or ratings. Such information could be consumed through a multitude of possible information channels, like websites, search engines, instant messaging, social media, gaming, or e-commerce platforms, VoIP, or E-Mail. However, due to the rapid technological developments, the enormous amount of information, the increased speed of distribution, the variety of channels and representations (e.g. text, memes [498], images, or videos), as well as the democratization of its content, meaning that anyone can create and distribute, lead to the fact that the general understanding of information flows as well as the assessment of information credibility becomes very difficult [483, 310]. Moreover, available information is far from consisting solely of investigative, researched and legit content, and its provision and distribution is not only altruistic, but strongly driven by self-interests. Various actors deliberately exploit the uncertainty of end users to influence their opinions or behavior, in particular due to commercial and/or ideological motives [499]. Hence, the digital information flow integrated into all areas of life and used in almost every daily situation is mixed with misleading, false, influencing and malicious information.

The types of malicious information users are confronted with today are manifold. Advertising-Spam [157], spoofed product ratings, or clickbait are usually distributed with the aim of (hidden) marketing products or services of any kind. Social Engineering actors, like Phisher, spread malicious information with forged legitimate identities and contents to influence the behavior of users, e.g. to install malware, to release sensitive information or simply money [220, 224]. Objectives range from extortion and identity theft to espionage and the sale of sensitive data, or all at once. When users react to this information, the resulting damage is enormous not only for individuals but also for companies and civil society [60, 61, 62, 119]. Nowadays, German companies receive an average of two pieces of malicious information for every legitimate piece of information, taking only the email channel into account [169]. Overall, this problem will increase rather than decrease in the future [217].

Using similar techniques, other actors, like criminals, organisations, governments, activists, or even trolls, spread misinformation — from conspiracy theories, biased or on-sided information to rumors, hoaxes or propaganda — to influence human minds, to generate monetary profits¹³, to cause harm to a person or organization, or even for fun or passion [499, 238]. Again, impacts range from serious issues for individuals, to companies [32], to societies (e.g. impacts on democratic elections [45, 15], crisis situations like the Covid-19 pandemic [227, 76], or the promotion of acts of physical violence¹⁴ [308]), and overall promote polarization and ideological division.

Thereby, the individual disclosure of more and more information (see Chapter 2.1) has a strong influence on the possibilities of manipulation and deception. With enough knowledge about victims, malicious information can be played out targeted and personalized to increase efficacy [219, 68]. Visible, e.g., also within the results of an own study [134], where we simulated a spear-phishing e-mail attack (targeted

¹³<https://money.cnn.com/interactive/media/the-macedonia-story>

¹⁴<http://wapo.st/2hfxKqk>

content) at the technical university of Dresden. Of the 4348 employees contacted over $\frac{1}{4}$ fell for it, mainly within the first 2 hours (see Fig. 2.4). In the area of news consumption, nowadays almost 50% of internet users are worried about missing out important information or challenging viewpoints due to personalization of the content [336].

In addition, available disclosed information and advances in machine learning enable new methods of information creation. Deep learning algorithms [317, 56, 436] simplify the automatic manipulation as well as synthetic generation of (high-quality) information that is virtually indistinguishable from other content¹⁵. In particular, the possibility of synthetic creation of audiovisual information from people with content they have never produced, often named deep fakes [317, 233, 441], facilitates targeted manipulation attempts. For example to destabilize or denounce political opponents¹⁶, influence user opinions, or to trick users into performing malicious actions or revealing sensible information [478]. Overall, such techniques will contribute to the intensification of the assessment issue in terms of quality and quantity. Moreover, digitization enables the simplification and acceleration of the distribution of information through automation by means of (social) bot networks [425, 100, 422].

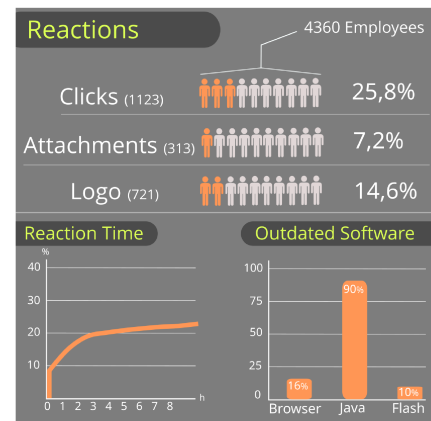


Figure 2.4: Spear-phishing attack results.

2.2.1 User Perceptions and Behavior

Ten intense years on the Internet had taught Agnes a lot. For example, that every published opinion, no matter where it came from, was met with contradiction because someone always had a different opinion. Moreover, fiction and truth merged, so that almost no one could distinguish one from the other. In the end, people almost always stuck to what they wished, completely detached from science and probability. Which led to something like the following intellectual exchange: A: "Consequently, it is scientifically proven that ..." B: "Shut the fuck up, asshole."

Jona Jonasson, The Prophet and the Idiot, 2022

The main reason of misinformation success, such as phishing, is baiting and deceiving users. Factors that influence how vulnerable users react to such attacks are, for example, age, gender, and technical background of the recipient, and the design and content of the misinformation [219, 48, 402, 344]. Depending on the content, simple curiosity leads to reacting to the attacks [35]. Personal character traits also play a role, especially the conscientiousness of the user [196]. In the area of IT security there is also a lack of interest and motivation [261], as well as a lack of risk awareness [112]. Simply not knowing about phishing increases the success significantly [118]. However, even with a known attack pattern, the users assessment criteria are not appropriate [48, 112, 263, 17]. A fundamental problem is that this type

¹⁵<https://arstechnica.com/information-technology/2022/12/thanks-to-ai-its-probably-time-to-take-your-photos-off-the-internet>
e.g.: <https://thispersondoesnotexist.com>, <https://podcast.ai>

¹⁶<https://www.bbc.com/news/technology-60780142>

of misinformation is composed of elements of everyday social interaction. Therefore, risks are often not seen in the automatic behaviors of everyday life, e.g. during work [261].

Similar to phishing, users are nowadays confronted with political content and news information rather by chance during their everyday lives. In 2020, 70% of Germans obtained their news online, 37% within social networks (OSNs) [209]. Online news consumption offers unprecedented access to a huge variety of news sources [160]. Thereby, OSNs are used as a tool to deal with the constant flood of information from various sources [355]. Within OSNs, however, news content is mixed with all kinds of information. This results in a form of news consumption where news articles are considered as part of the feed viewed for entertainment rather than being accessed in a dedicated manner. This is called *incidental news consumption* [49], which can i.a. lead to lower levels of political learning [178]. Moravec et al. [323] found that users tend to exhibit a 'hedonistic mindset' in this type of news consumption, where entertainment is the primary focus and critical engagement with the content is foregone. Thus, engagement with news content is superficial and is seen more as part of social interaction within the network [160]. Moreover, many users only read the headlines and summaries of the news information [181], feel thereby better informed than they actually are [21], and share news content based on these summaries without having read the actual content [323]. Mixed with misleading, biased, false and polarizing information, users struggle to assess the credibility of information [258, 337]. Nowadays 64% of social media users are concerned about what is real and fake news [336]. Similar to phishing, used assessment criteria are not appropriate [310]. Instead of using the concrete content or source of information [114, 222], indicators like the user who distributed the information [421], social metrics like the number of likes [28], or used images influence decisions [198].

How users deal with misinformation depends, i.a., on characteristics such as educational background, frequency of news consumption, and age [15, 397, 188]. In addition, cognitive skills such as analytical thinking [354] and ideological priors and beliefs play a role. Latter include factors such as naive realism (users tend to believe that their own perception of reality is the only correct one), confirmation bias (users prefer information that confirms their existing beliefs), or normative influence theory (users consume and distribute information based on their community norms due to the need for social acceptance and affirmation) [407].

2.2.2 Countermeasures

One solution may be the **verification of legitimate information** in the flood of information. Identification and authentication approaches help to verify users which distribute information within digital services, like OSNs [392]. Cryptographic methods, such as digital signatures or hashes, help to verify the integrity and authenticity of information content itself [378]. However, verification systems are currently mainly used in special fields of application, are even exploited for the legitimization of malicious content (e.g. via TLS certificates for phishing websites [311], or visual spoofing/misuse of verification stamps in OSNs¹⁷) and are sometimes difficult to use for end users (e.g. handling of signatures and keys). The use of verification also leads to a loss of privacy, as any information becomes unambiguously traceable. Even if approaches exist that enable this privacy preserving, e.g. anonymous credentials for user authentication [65, 66] with practical realizations¹⁸, they currently only play a minor role in practice.

¹⁷<https://www.forbes.com/sites/brucelee/2022/11/12/fake-eli-lilly-twitter-account-claims-insulin-is-free-stock-falls-43>

¹⁸<https://privacybydesign.foundation>, <https://abc4trust.eu>

The privacy vs. security problem becomes especially clear in the inverse approach - not to verify legitimate information but to be able to disprove malicious information in a provable way, by means of full-scale tracking of one's own being [78]. Thereby, 'immutable life logs' create '**credible alibis**' [78]. Thus, one could prove, e.g. in the case of being a deep fake victim, that this cannot have been produced and distributed by oneself. At the same time, such approaches have an enormous impact on privacy.

Next to verify legit information, a bunch of approaches try to **detect malicious information**. Such approaches could be based on either manual (crowd-based), automated, or hybrid detection. Thus, solutions exist for various information, for example to detect spam and phishing [239], click-bait [8], social bots [158, 12], fake news [321], hate speech [70], or synthetic data (audio [46], video [441], text [502]). The possibilities of automated analysis of information, however, furthers the security vs. privacy discrepancy. Detection methods are based on e.g. user behavior, information distribution patterns, the content, or source, which overall means that also legit behavior and content has to be analyzed [505, 397]. Moreover, various actors are always looking for ways to circumvent detection algorithms, which ultimately ends in a game of cat and mouse. For example, automated malicious e-mail detection methods are quite effective [496]. However, at the same time e-mail is still one of the biggest gateways for successful phishing attacks [187]. In addition automated detection algorithms could also directly be attacked [182, 285, 147]. Overall, a complete detection rate could never be reached. Nevertheless, after detection of malicious content there are several possibilities to react. Detected information could be **reduced** in its **exposure**. Either by simply deletion, or blocking, by reducing its visibility (e.g. down ranking in social media algorithms), or demonetization [418]. However, blocking and filtering information can also lead to discrepancies between freedom of speech/utility and safety. Thus, a high blocking rate may also prevent the delivery of legitimate information and vice versa.

This discrepancy between restricting communication freedoms and preventing harm (freedom vs. security) is also reflected in the **legal regulation** of malicious information, especially misinformation. Thus, general prohibitions on even deep fakes or targeted misinformation do not yet exist in Germany and Europe. At the systemic level, the current regulatory system so far relies essentially on the instrument of self-regulation [92], by the intermediaries involved (e.g. OSN providers) and by democratic discourse, and in cases of misinformation with individual relevance (personal rights, ...) also on the enforcement of existing individual claims. For the latter, regulation of misinformation takes place, e.g., through criminal law. For example, hate speech could be regulated based on §185-§187 StGB involving defamation, slander, and crimes against personal honor. Phishing could be regulated based on §269 StGB regarding the deception of the user, §263 StGB regarding financial loss, §§202 StGB regarding sensitive data spying or §12 BGB for faking legitimate identities in case of violation of name rights. However, law enforcement often appears difficult within digital infrastructures ('anonymity', mass of information, ...).

With the regulations of the *Netzwerkdurchsetzungsgesetz*¹⁹ and the *Medienstaatsvertrag* there are several provisions in German law which, i.a., address law enforcement on digital platforms and ensure the basic conditions of the communication system. In addition, there are two new EU regulations, namely the AI-act [96] and the Digital Services Act [94], which deal, among others, with the regulation and handling of misinformation and deep fakes. However, even though regulation is promoted on different levels, currently, self-regulation by platforms is still predominant in large parts. E.g., the reports of the different platforms under §2 NetzDG reveal that by far the most of removals of content are due to the community guidelines

¹⁹<https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html>

and not due to the legal provisions of criminal law and NetzDG²⁰.

In general, the regulation of information remains an area of tension. Overall, this is also due to the fact that other freedoms are violated in the regulation of information. Because one question remains: who and how will define what information is harmful and what is not²¹. This also applies for self-regulation²².

Due to the regulatory difficulties, measures to reduce credibility of misinformation, e.g. via **labeling of information** (based on detection approaches), as well as the empowerment of users are debated as additional measures to **increase** their **resilience** against misinformation [96, 95, 418]. Overall, this is the goal of transparency enhancing technologies with focus on the assessment of information.

2.3 Transparency Enhancing Technologies

As described above, technical and regulatory approaches cannot fully protect users in the digital world from the issues of (sensitive) information disclosure and the consumption of malicious information. Transparency and awareness-raising measures are therefore an important supplement to support users in dealing with these problems. In the field of privacy, such approaches are summarized under the term *Transparency Enhancing Tools/Technologies* (TETs) [225, 329, 507]. In this work, we adapt this term also for transparency approaches in the area of information consumption, as there currently exists no established umbrella term in this field. Hence, we differentiate between **Privacy TETs** and **Assessment TETs**. In the following, we want to give insights into various TET approaches and their properties.

2.3.1 Privacy TETs

Privacy TETs seek in particular to reduce existing *information asymmetries* between providers and users. They provide concepts and solutions to visualize meaningful insights into what, how, and why personal information is disclosed, collected, processed and stored by services and applications in an accurate and comprehensible way for the average internet user. The fundamental goal is to support users in the assessment of privacy risks [507, 330], i.e. to enable real informed trade-off decisions between privacy risks and benefits while using digital services or applications. This can already lead to users intending to take privacy-protecting actions [474, 24] and to increased privacy concerns [384]. But even if decisions are not directly changed by transparency solutions [41] (see also Sec. 2.1.1), they increase the user's basic understanding [474, 41]. To support and increase user's general privacy awareness, mental models and privacy literacy is the overarching aim of Privacy TETs. Thereby, critical privacy literacy could be described as 'knowing that' (declarative knowledge), 'knowing how' (procedural knowledge) as well as the ability of reflection 'to identify privacy risks and the actual level of privacy when using different online environments' [300, 444]. Increased basic understanding can in turn lead to improved trust in the digital world [225], and allow users to become part of the solution by participating and influencing social discourse and negotiation processes in the area of privacy, which can ultimately achieve more than individual protection [300].

²⁰e.g. <https://transparencyreport.google.com/netzdg/youtube?hl=en>

²¹<https://www.tagesschau.de/ausland/asien/tuerkei-desinformationsgesetz-101.html>

²²https://www.nrk.no/osloogviken/xl/tiktok-doesn_t-show-the-war-in-ukraine-to-russian-users-1.15921522

Types of Privacy TETs Janic et al. [225] classified Privacy TETs regarding to its content. They distinguish between TETs that provide insights into *possible, intended/reported* (e.g. via simplification of policies), and *actual* data disclosure. Zimmermann [507] proposed a categorization based on Hedbom [206] and Zimmermann [508]. Here, TETs are classified within six categories (Fig. 2.5), based on the following parameters:

The **Assurance Level (AS)** 'describes the extent to which data-subjects can determine the completeness and correctness of information'. The **Application Time (AT)** divides TETs on the basis of temporal display. Transparency information can be displayed either before (*ex-ante*), during (*realtime*), or after (*ex-post*) the actual data disclosure and processing. The **Interactive Level (IL)** distinguishes TETs according to their functionality. *Read-only* TETs only provide insights, whereas *interactive* TETs also enable intervention and thus control over information disclosure.

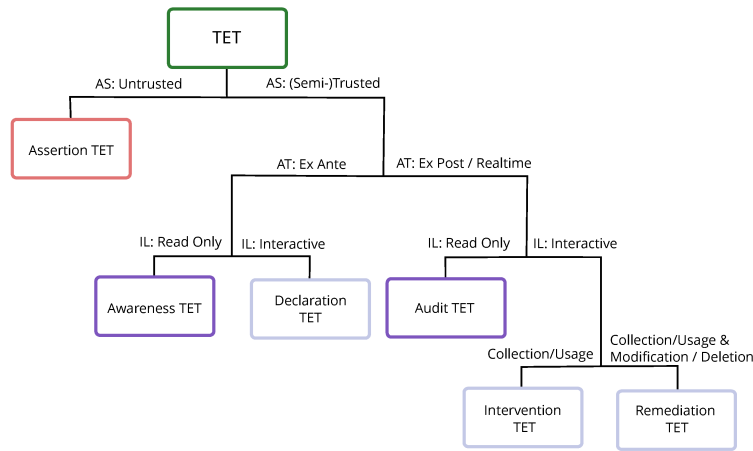


Figure 2.5: TetCat: Categorization of TETs by Zimmermann. Source: [507]

Additional categorization parameters are the **target audience**, which could be either data-subjects or external auditors; the **delivery mode**, which describes the need for activity of the data subject to get transparency information (push or pull); the **execution environment**, which describes if the TET is executed on the client itself, on server-side (offered by the data-controller or a third party) or hybrid; the **scope** describes the 'range of services or data controller' a TET considers; the **information source** describes where the transparency information is coming from; the **transparency dimension** means which part of information disclosure is considered (collection, analysis, usage, 2nd usage); and finally the **attacker model**, which means the threat focus on either other service users, service providers or third parties. **Data types** which are presented through the TET are distinguished between volunteered (explicit disclosure), observed (implicit disclosure), incidental (disclosure by others), derived (possible predicted through data analysis), and policy data (reported). Murrmann and Fischer-Hübner [329] proposed a similar categorization with different terminology. An additional and important category they consider is called **visualization** and describes the preparation and presentation of transparency information as well as the levels of detail. In the following, we give an brief insight into the general categories (Fig. 2.5).

Awareness TETs inform the user before the actual information disclosure. These are for example educational offers, which try to give general hints on how to behave and deal with digital systems, on

associated *possible* information disclosure and its risks. Based on e.g. text²³, videos²⁴, special tools²⁵, or games²⁶, it is shown which potential identifiable traces are left behind when using digital systems, why and how this is collected and processed, how tracking works, as well as hints on how to prevent this, which alternative services do exist²⁷, or how to handle the process of accessing or deleting disclosed data²⁸. Further Awareness TETs directly analyse services and applications regarding *possible* information disclosure and visualize and explain e.g. permission²⁹ [156], static code³⁰, or network³¹ analysis results. Other Awareness TETs simplify *intended/reported* data collection and handling of data controllers. These approaches try to either manually³², automatically [497, 440, 368], or hybrid [506] understand and aggregate privacy policies (or similar reported information), in order to make them easier accessible and comprehensible. For quick assessments, policy characteristics are ranked with privacy grades [497], simplified via labels [128, 127, 241] or iconification [133, 210, 123, 194], or they are made searchable more quickly³³.

Few ex-ante TETs also give intervention possibilities (**Declaration TETs**), for example cookie banner with opt-in/opt-out settings, their simplification[214], or approaches which support the handling of privacy settings within services [351].

In contrast, realtime/ex-post TETs mainly focus on the treatment of *actual* disclosed information. **Audit TETs** try to analyse and transparently prepare *actual* information flows between the user/application and connected services in realtime. This is done for specific applications, e.g. visiting websites³⁴ [296], using smartphone apps [369, 252, 427, 282, 458, 16], or for the whole home network and its different devices³⁵ [391]. Due to their 'firewall-like' functionality, most TET solutions in this area also offer the ability to intervene (**Intervention TET**) and thus control information flows, e.g. via blocking specific connections [252, 38]. Over a longer period, some of these approaches also offer the possibility of viewing the history of information disclosure (ex-post), i.e. who had access to which information and when.

In addition to the ex-post presentation of actual disclosed information, **Remediation TETs** enable the possibility to correct or delete them. This is usually realized via privacy dashboards, offered either by data controllers themselves based on legal requirements, or externally. Latter attempt to prepare disclosed data across different services and applications [19, 44, 365]. A broad overview of existing ex-post TETs is given by Murmann and Fischer-Hübner [329, 330].

Usable existing Privacy TETs are mainly available for 'traditional' digital communication, like websites or mobile apps. In this work we want to do a step forward into transparency regarding the disclosure through digitized analog information focusing on printed documents and audiovisual recordings of users through the surrounding IoT. Thereby, we focus on unconscious (implicit) information disclosure targeting the data-subject, likely to be a non-professional user.

²³<https://myshadow.org>, <https://datadetoxkit.org>

²⁴<https://blog.donotrack-doc.com>, https://pbs.org/wgbh/nova/labs/video_popup/5/34

²⁵<https://coveryourtracks.eff.org>, <https://amiunique.org>

²⁶<https://datadealer.com>

²⁷<https://privacytools.io>

²⁸e.g. <https://justgetmydata.com> or <https://justdeleteme.xyz>

²⁹<https://appcheck.mobilsicher.de>

³⁰<https://exodus-privacy.eu.org>

³¹<https://privacyscore.org>, <https://trackography.org>

³²<https://tosdr.org>, <https://privacyspy.org>, <https://foundation.mozilla.org/privacynotincluded>

³³<https://www.usableprivacy.org>

³⁴<https://privacybadger.org>, <https://lightbeam.chikl.de>, <https://duckduckgo.com/app>

³⁵<https://pi-hole.net>

2.3.2 Assessment TETs

Assessment TETs address the issue of misinformation by promoting media literacy and resilience via various means, such as teaching people to recognize misinformation [31, 263, 312, 403, 235], changing people's consumption behavior [151, 289], or labeling suspicious information with credibility indicators [243, 248, 323, 466]. Thereby, media literacy can be defined, as the 'ability to access, scrutinize, assess, and create information in a variety of forms [26].

Through increased awareness, such approaches can help to reduce the perceived credibility of misinformation [323, 248, 84, 353, 236, 312, 243], to reduce their distribution by reducing the sharing intention of such information [307], and to reduce the interaction with it [263, 466, 424, 67, 262, 259]. The overall goal is to improve the ability to assess information, thus to reduce the negative impacts of misinformation, and ultimately to enable users to become part of the solution. For example, through reporting malicious information instead of distributing and consuming it.

Types of Assessment TETs As there is currently no categorization approach for Assessment TETs, we adopted the TetCat approach [507]. We classified Assessment TETs in five categories (Fig. 2.6), based on the same parameters.

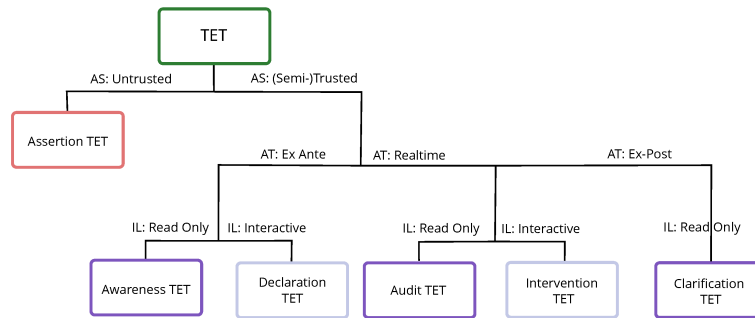


Figure 2.6: Categorization of Assessment TETs

The **Application Time (AT)** divides TETs on the basis of temporal display. Transparency information can be displayed either before (*ex-ante*), during (*realtime*), or after (*ex-post*) the consumption of information. The **Interactive Level (IL)** distinguishes TETs according to their functionality. *Read-only* TETs provide insights into consumed information, whereas *interactive* TETs also enable prevention of delivery and/or give interactive possibilities. The **scope** describes the type of malicious information (fake news, phishing, deep fakes, spam, ...), and **transparency dimension** describes which part of the information is considered (author, distributor, content, meta-data, ...). The remaining parameters of assurance level, target audience, delivery mode, execution environment, visualization, and information source could be directly adopted. The attacker model is not considered here, since the intention or the original producer of the misinformation is almost impossible to detect and is also not relevant for the reception and impact of it. In the following, we give an brief insight into the general categories.

Awareness TETs inform the user before consuming and falling for malicious information. These are mainly educational offers, which try to give general hints on how to deal with digital information. Based

on e.g. text³⁶, cartoons [415], quizzes³⁷, trainings [364, 259, 260], graphs³⁸, or games [22, 403, 477, 376, 183, 312] it is shown how easy it is to create, distribute and modify information, how misinformation is distributed (e.g. social bots) and which actors are involved, how users can spot deceptive information, as well as which impacts and risks malicious information can have. Overall, to increase the understanding of digital information flows and its risks. **Declaration TETs**, additionally enable users to block information before it is delivered³⁹.

In contrast, realtime TETs support the user in assessing and verifying information directly during its consumption. For example, approaches that provide the user with warning labels directly related to the suspicious information, e.g. spam and phishing content [124], misinformation [323, 248, 84, 353, 236, 307], or synthetic generated information⁴⁰.

Thereby, corresponding warnings can be either passive (**Audit TET**) or active (**Intervention TET**). With Intervention TETs the actual content is, e.g., interrupted by a barrier or overlaid with information about the questionable nature of the content and the user has to actively bypass the warning in order to consume the questionable content [124, 42, 466]. This reduces the speed of interaction with content and provides space for reflection.

Other realtime TETs do not visualize warnings but enhance the content with context information that support the user in the assessment process. This can be, e.g., related information about the specific topic to provide a broader picture⁴¹, information about the publisher [42, 243], the author⁴², the distributor⁴³ [106], or the content itself [202, 191, 466]. Corresponding context information can thereby either be abstracted, e.g. to a rating of the information [243, 191], or visualized directly⁴² [466].

The collection of indications if an information is questionable or the gathering of corresponding contextual information can be either manual (crowd-based, expert), automated, or hybrid. Thereby, crowd-based approaches [212] usually offer the interactive opportunity to participate in the solution process, e.g. by reporting suspicious information like spam and phishing⁴⁴ [283], fake and misleading news⁴⁵ [390, 243], hate speech⁴⁶, or deceptive designs⁴⁷.

Clarification TETs attempt to correct the impacts of misinformation after their consumption, by means of clarification and correction. In the area of news consumption, fact checkers in particular have established for this purpose⁴⁸. Such solutions systematically collect and sift news articles with high social interest, try to verify or debunk them using investigative approaches, and publish their results. Corresponding reports have either to be visited proactively, or they could be used directly as Audit TETs next to corresponding sources⁴⁹.

³⁶<https://www.dw.com/en/fact-check-how-do-i-spot-fake-news/a-59978706>, <https://datadetoxkit.org>

³⁷www.oswego.edu/cts/phishing-quizzes, <https://der-newstest.de>

³⁸<https://hoaxy.osome.iu.edu>

³⁹<https://privoxy.org/>, <https://jugendschutzprogramm.de>, <https://blog.mozilla.org/en/firefox/firefox-b-tch-to-boss-extension>

⁴⁰https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media

⁴¹<https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation>

⁴²<https://github.com/getcahoots>

⁴³<https://botsentinel.com>

⁴⁴<https://phishtank.com>

⁴⁵<https://fiskkit.com>

⁴⁶<https://www.coe.int/en/web/no-hate-campaign/reporting-hate-speech>

⁴⁷<https://deceptive.design>

⁴⁸<https://snopes.com>, <https://factcheck.org>, <https://correctiv.org>, <https://www.politifact.com/>

⁴⁹<https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works>

2.3.3 Preliminaries

As TETs focusing on end-users, it is important to create a usable interface that presents the transparency information in a simple and comprehensible way. Something being usable is defined as what 'a person with average or below-average abilities and experience understands how to use [...] to achieve something, without the effort being greater than the reward' [256]. As a general guideline, ISO 9241 [161] can be considered, an international standard for describing human-computer interaction (HCI) guidelines. In particular, three aspects are summarized here that are most significant in determining usability: effectiveness, efficiency and satisfaction. In addition, seven design principles of HCI are defined. These include conformity with user expectations, learnability, self-descriptiveness, fault tolerance, controllability, user engagement, and suitability for the user's task [162]. Other noteworthy design principles are, e.g., accessibility [361], or responsive design [172]. Latter provides a clear and readable interface independent of the used display. Overall, the cornerstone of usability is to 'not make people think' when using something with a user interface. To achieve this, e.g. unclear elements, technical terms, or noise should be avoided, determination of possible actions must be easy, and consistency is of essence when designing, e.g. by using common terms, action sequences, colors, or menu hierarchies [405, 2, 340]. In general, for TETs, an aesthetic and minimalist form of design is recommended, which serves the quick and easy understanding of transparency information [329].

Next to the interface design it is important that TET information does not overwhelm the user. This applies in particular to realtime TETS, i.e. solutions that are used constantly on a daily basis. Gomer et al. emphasize the phrase 'minimal distraction principle' [185], which expresses the desire of users for no distractions and notifications requiring interaction in everyday life. Friedman et al. [168] describes the need for the minimal distraction principle, as studies show that user's are overwhelmed by 'being informed' in everyday life, thus completely disengaging from the process. Thus, users should be able to receive and comprehend information, without 'unduly diverting the individual from the task at hand'. Overall, a transparency enhancing system should 'minimize information fatigue and maximize the likelihood that the user will not miss any useful information' [324].

It is therefore essential to address the following questions. What information is crucial and needs to be presented? How can this information be presented? In general, for quick and minimal access, corresponding transparency information has to be abstracted and simplified. It may be beneficial to go beyond text and find a graphical representation, to improve understanding, faster access, and reduce information overload [123]. For this, e.g. icons, colors, grades, graphs, maps, or charts can be used [329]. However, abstraction also has its drawbacks, like oversimplification, which could result in misunderstandings [133], or the difficulty of finding comprehensible global, generalized abstractions, e.g. due to cultural differences [211]. Moreover, to enrich trust and understanding in transparency information users should be assisted with explanations on how it is aggregated and why it is displayed [248].

To achieve both - simplification and explanation - transparency information could be presented, e.g., with hierarchical layering [329]. The first level is a highly abstracted representation of transparency information that only covers the most important aspects. Detailed information are then available by navigating down the layers. In this way, users are not overwhelmed, but are informed directly and introduced to more in-depth explanations, depending on their own pace and interests. In addition, an 'individualisation of TET's is considered desirable' and an adaption 'to the user's actual requirements and level of expertise' could improve efficiency and effectiveness of tasks the user want's to complete inside the TET [329].

3

Analog Information Exchange in the Digital Age

The further development of letterpress printing by Johannes Gutenberg in the 15th century, by means of movable letters made of lead, enabled the mass reproduction and dissemination of information, knowledge, and ideas and thus their accessibility to all levels of society. The accompanying change in the exchange of information enabled cultural, economic, and scientific upheavals and forms the basis for our present time [125]. For centuries, paper and printing techniques have been the foundation for the distribution of information.

Even though today's information exchange takes place primarily via digital devices, analog documents are used everywhere and thus still play an important role in our societies¹ [443]. Contracts, tickets, money, letters, invoices, ballots, certificates, or analog archives are just a small selection of examples. Moreover, technical developments allow an easy transfer of information from the analog to the digital world (scanners, cameras), and vice versa (printers). This information transfer, as well as easy-to-use processing software, allows anyone today to simply and inexpensively create, duplicate, and modify analog documents with high quality.

Despite the positive impacts of this development, however, also new issues arise. Due to the ease of creation and modification, printed documents are also an issue in crimes, such as fake IDs, forged certificates, counterfeit money, copyright infringements, blackmail letters, or as evidence in a criminal case. Hence, methods that allow statements about the process of creation of such documents are important to verify their credibility or for revealing the originator. Thereby, an important role is played by the attribution of a document, i.e. the assignment of a printed document to the printer used. The research field of printer forensics provides solutions for this. Tools and algorithms developed in this area can be distinguished between passive and active methods [79].

Passive methods take advantage of the fact that the quality of printouts is influenced by the corresponding printer mechanism and its components. This as well as several imperfections of such components produces artifacts within the printed document. Passive printer forensic methods try to find such artifacts or individual printing characteristics which are stable over several iterations, distinguishable among different

¹<https://www.idc.com/getdoc.jsp?containerId=prUS50453823>

printers, and robust against influences. These artifacts can be used as identification features, called *intrinsic signatures*, of specific printer technology, brand, model, or the device itself.

In contrast, active forensic methods focus on hidden information, called *extrinsic signatures*, that has been explicitly added to the document before or during the printing process. This information, e.g. the serial number of the printer device or a secure hash of the document, can then be used to identify the printer or to detect forgery [291, 73]. Examples in this field are the intentional adding of banding frequencies [314], halftone shifts [426] color-tile deterrents [174] or the EURion constellation [257]. These active methods, however, are only useful if the printing process can be controlled, for instance then printing money or IDs, unless one would integrate such a methodology directly into end devices.

— However, this is what currently is exactly done.

So-called machine identification codes (MIC), yellow dots, or tracking dots [165] are currently integrated into nearly all color laser printers and encode hidden information like the serial number of the printer or the date of print within each printout. One reason for this could be the instability of the printing process, concerning the extraction of intrinsic signatures (see Sec. 3.2). However the whole process is non-transparent, it is unclear who is responsible and what exactly is encoded within such embedded tracking dots.

Overall, this illustrates very well the dichotomy between security and privacy, also concerning analog documents. On the one hand, such methods are very important to protect the consumption of information, i.e. to assess the credibility and authenticity of printed information and to solve crimes. On the other hand, the availability of such methods allows inferring additional information not only from suspicious documents but from all and by arbitrary parties, resulting in privacy-invasive issues. Currently, users have presumably no awareness that when they exchange information with printed documents, more information is disclosed than the actual content of the documents.

Thus, in the following, we briefly describe the information disclosure through intrinsic properties of printed documents, exemplary review a concrete printer identification scheme, and evaluate its robustness. For this, in Appendix I-A, we present our developed benchmarking system *Twizzle*, which simplifies the comparison of such algorithms. Afterwards, we analyze the non-transparent tracking dots, their characteristics, and try to decode their content. On the one hand, for re-usability from the forensic point of view (transparency for verification) and, on the other hand, to create transparency for the user as to what additional information is disclosed when an analog document is shared. Additionally, we explore a first anonymization approach against this extrinsic signature to defeat arbitrary tracking. Finally, we present our declaration TET *DEDA* which implements the entire workflow of extracting, analyzing, and anonymization of such a tracking dot pattern. In conclusion, we briefly discuss the social reactions to the publication of the results.

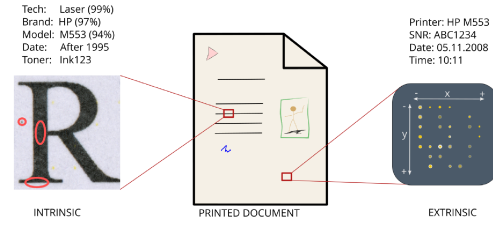


Figure 3.1: Intrinsic vs. Extrinsic Signatures

3.1 Customary Printer Technologies and their Functionality

Nowadays the content of analog documents is mainly created, processed, or modified with digital devices using editing tools and afterward analogized via printing technologies. In the printing industry, there are a large number of different technologies and processes, such as offset printing or dye-sublimation printing. In office or household environments, however, two types of devices are mainly widespread: laser and inkjet printers. In the following, we describe their structure and functionality, as this is crucial for the emergence of intrinsic signatures.

3.1.1 Laser Printer

The underlying technology for laser printers is called xerography or electrophotography (EP). The core of this technology is a constant revolving, cylindrical drum, whose surface is made of photoconductive material. This drum is called the Organic Photo Conductor (OPC) drum. The principle workflow of a laser printer consists of six steps [247, 79]. At first, a charge roller applies a uniform electrostatic charge to the surface of the OPC drum. Next, a controlled light source (laser or LED) is directed line-by-line onto the OPC drum through a rotating polygonal mirror and special optics. At certain points, which represent the negative image to be printed, the laser beam turns on and discharges these locations by exposure. The toner particles are then electrostatically attracted to the charged areas of the OPC drum through a developer roll. Next, the developed image is electrostatically transferred onto the paper. The image is then made permanent through a combination of heat and pressure. In the final step, the cleaning blade removes the remaining particles of toner from the OPC surface. Used inks are either powder or liquid toners, which contain the colorant in the form of pigments.

Regarding color prints, this process has to be iterated four times for cyan, magenta, yellow and black toners, which have to be printed overlapped (see Sec. 3.1.3). There exist mainly two solutions for a color laser printer architecture. The first solution, called multi-path or revolver type, has only one OPC drum with four development units. Most color models use an intermediate transfer belt on which first all colors will be transferred and from this developed onto the paper. The other architecture, called inline or tandem type, uses four separate printing units for each color. Here a very exact positioning of the paper is essential.

3.1.2 Inkjet Printer

Compared to EP, inkjet printers does not need an intermediate stage like the OPC drum. Instead, ink will directly be transferred onto the paper. In general, the technologies of inkjet printers can be distinguished between continuous and drop-on-demand (DOD) printing [247]. The general mechanism of DOD inkjet, which is mainly used in customary devices, consists of three basic components: the print head, the carriage and the paper feed mechanism. Here an inkjet printer has to perform two general motions. The vertical movement of the paper and the horizontal movement of the carriage with the print head. First, the paper feed mechanism advances the paper under the carriage. Next, the carriage moves the print head across the paper. The print head consists of many tiny nozzles which are arranged in several staggered columns. While the print head goes through the paper the ink will be ejected out of the nozzles onto the paper. In one pass of the print head several rows can be simultaneously printed through each column of nozzles. After the pass is finalized, the paper is advanced again and a new pass begins. This process is repeated until the image is completed.

There exist many constructions for this architecture which differ e.g. in speed and direction of the print head, number of printing passes or number of nozzles and columns. A newer architecture, sometimes called memjet printing technology, uses a stationary page-wide print head. Here the horizontal movement and the carriage is no longer required. The mechanisms which eject the ink can be basically classified into thermal, also called bubblejet (e.g. used by Canon), and piezoelectric methods (e.g. used by Epson). In majority the used ink is liquid and based on a mixture of water and pigments or dyes.

3.1.3 Halftoning

Continuous tone images, such as photographic originals, consists of a large range of tonal values. Most common printing technologies, including laser and inkjet, only have two operations, to transfer ink or not. Therefore tonal values have to be *simulated* through breaking up the image into very small dots. This process is called halftoning. The original digital image is divided into a structured grid with uniform 2D screen cells, where the screen frequency describes the number of cells per inch and is measured in lines per inch (lpi). Each screen cell simulates a tone value of the original image and consists of micro dots which are described by the resolution of the printer, measured in dots per inch (dpi).

By selectively filling the micro dots of a screen cell, different color shades can be simulated. The more cells in a screen cell are filled the darker this area, and vice versa. The number of possible tone values is determined by the size of the screen cells $S = \frac{dpi}{lpi}$ and is denoted by $T = (\frac{dpi}{lpi})^2 + 1$, whereby the +1 adds the white tone of the substrate (paper). From 150 lpi the human eye is usually no longer able to detect individual dots from a normal reading distance. For printers that can print real tone values, e.g. by varying the intensity of the laser beam of EP printer, the number of possible gray values per screen cell increases with $T = (\frac{dpi}{lpi})^2 * (g - 1) + 1$, where g is the possible number of real halftones.

For multicolored prints, the image is broken into several color channels. For this purpose the subtractive color model is used, which describes the mixture of the primary colors cyan, magenta and yellow to attain a wide range of colors. Typically, a separate black color is included, resulting in a CMYK color model. The halftoning process is applied to each color channel which are then successively printed, overlapping each other.

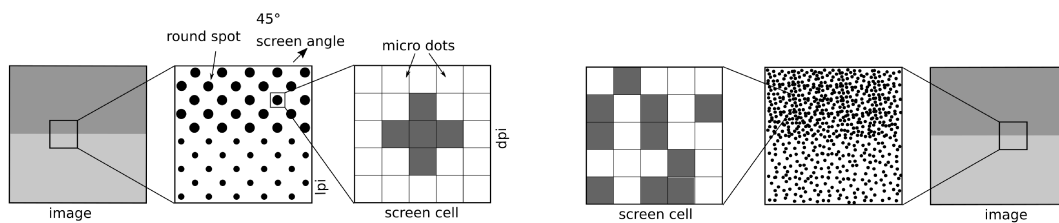


Figure 3.2: AM Halftoning (left) and FM Halftoning (right)

There exist several approaches describing how to fill the screen cells to simulate the tone values (see Fig. 3.2). Overall they can be categorized into Amplitude Modulation (AM), Frequency Modulation (FM) and Hybrid Screening [247]. AM halftoning, which is commonly used by laser printers, creates spots that are arranged in regular columns and rows and are equally spaced. Tone values are then achieved by varying the size of each spot in each screen cell. In comparison FM halftoning, commonly used by inkjet printers, create dots which have the same diameter but are arranged in different distances. In dark image locations the dots are in high density and conversely. The arrangement of the dots is mainly done

randomly. Common FM algorithms are e.g. error diffusion or dispersed dot screening. Hybrid screening combines FM and AM halftoning [205].

The screening process itself will be realized through a Raster Image Processor (RIP). The input for this is a page description in a high-level page description language (PDL) such as PostScript or PCL. Nowadays the RIP component is commonly a firmware program embedded into the printer.

3.2 Intrinsic Printer Signatures and their Robustness

The analysis of analog documents, such as their date assignment, detection of forgeries and alterations, or the attribution of the author have always been important information for their verification. Corresponding methods range from paleography (i.e. the dating of historical documents) [485], forensic handwriting analysis, and graphology [208, 334] which even tries to detect clues of the author's personality from the handwriting, to the analysis of typewriter documents [121, 190]. All these approaches extract and analyze properties and artifacts of a document, which generate identifiable intrinsic characteristics by the author (e.g. writing style², variations in typing keystroke force), or the corresponding writing tool (e.g. paper properties, ink composition, comparison of typewriter keys).

Today's inkjet and EP printing devices also do not produce uniformly perfect analog documents. As described in Section 3.1, these devices and their components are diverse, differently constructed and also contain mechanical imperfections. These differences are also reflected in the resulting documents and could be used as intrinsic signatures in order to be able to make statements about the printing process and thus to detect alterations/forgeries [456, 395] or to identify the originator.

Traditional forensic technologies [399, 111], like physical (e.g. Ramada-Spectroscopy) [170], chemical (e.g. Chromatography) [433] or microscopic [345, 338, 449] methods, which try to extract such identifying signatures, like ink characteristics, can give good results but are slow, require specialized equipment, educated employees and may destroy the document itself. Digital forensic science, on which we focus in the following, aims to improve the analysis in such a way that it can be carried out cost-effectively and automatically with standard commercial scanners [79, 291, 394]. Digitization thus simplifies not only the creation and forgery of analog documents, but also their verification process. On the other hand, such verification tools can be used by anyone, which poses a potential risk to the privacy of the document's originator.

However, this discrepancy and overall the effectiveness of passive forensic methods depends on the sensitivity and robustness of the intrinsic signature used. These properties, in turn, are insufficiently considered in the literature. Hence, in the following, we re-investigate an existing identification scheme which uses the halftone process as a signature to identify the source printer brand and model [380, 245, 246]. Its low complexity and overhead makes it especially interesting for broad use – the authors simplify the forensic process to capturing the image with devices as simple as smartphone cameras. We evaluate the robustness and resistance to forgery of this signature. In addition, we investigate whether the signature could also be used for prints that contain solely text. Another approach which uses banding artifacts [14, 79] is analysed in Appendix I-B. Finally, we discuss the impacts of intrinsic signatures on privacy and verification of analog documents.

²<https://sz.de/dpa.urn-newsml-dpa-com-20090101-220105-99-593400>

3.2.1 Existing Intrinsic Signatures for Identification

Depending on the content of the document to be examined, different signatures can be used for its source device attribution. Methods focusing on text documents mainly analyse the differences of texture and structure of printed characters. Distinguishing features are for example *micro textures* within printed characters [159, 108] (e.g. influenced by surface properties of the OPC-drum, properties of the toner or of fuser- or transmission system), *edge roughness*, overspray or satellite drops [345] (e.g. influenced by printhead movement).



Figure 3.3: (a) Differences of Micro Texture and Edge Roughness printed by HP M553 (left) and HP 4350 (right).
(b) Differences of color representations by HP M553 (left) and Ricoh MP305 (right)

These printed text characteristics are suitable for identifying the source printer technology used to create the analog document [176, 386, 110, 396], i.e., to differentiate between inkjet and EP printer, as well as for the identification of the specific printer brand and model [159, 126, 446, 492, 469, 315, 199, 230, 231, 450]. *Geometrical distortion*, caused by various mechanical imperfections (e.g. OPC-drum rotation or transmission system) and modes of operation, is another artifact which could be used as intrinsic signature for text [489, 221] as well as for image print outs [14, 59, 226] to identify the used printer model, brand and technology. Additional for image prints, the different implementations of the halftoning process and thus a different arrangement and distribution of the halftone dots can be used as an intrinsic signature of the printer model [396, 245, 244]. Since printers in private households are usually not periodically calibrated with standardized color profiles, the different color representations of halftone image outputs could also be used to identify the used printer brand and model [80, 448]. Furthermore, there are methods that analyse the structure of the used paper [83] or extract traces left on the paper by the paper feed construction, called spur marks [10, 11].

General Workflow of Signature Extraction In order to extract the corresponding intrinsic signatures, the documents to be examined are first digitized with customary scanners, with resolutions of 300-2400 dpi. After a subsequent pre-processing of the digitized document (e.g. noise reduction, distortion correction or color channel conversion) and extracting the important areas of the document (e.g. OCR [75] for char extraction), the feature extraction based on image processing algorithms follows. For example, for the extraction of texture properties from printed characters, Gray Level Co-occurrence Matrix (GLCM), DCT or Wavelet Analysis (DWT), histogram of gradients (HoG), Gabor Filtering or Local Binary Pattern (LBP) features are used [446, 315, 396, 159, 492, 447].

Extracted feature vectors are then used to either train a classifier [80, 448, 226, 489, 396, 159, 446, 126, 199, 230, 231, 450], like a Support Vector Machine (SVM), or to create a reference fingerprint for each printer model [108, 240, 492, 246, 59, 487, 315]. Hence, unknown printed documents could be compared

with known fingerprints via similarity measures (e.g. Euclidean distance) or by using the trained classifier to determine the probable source printer. The general workflow could also be seen in Figure 3.4.

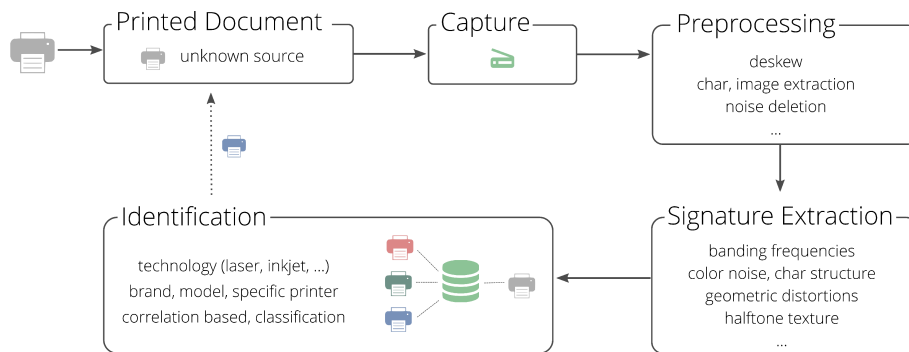


Figure 3.4: Workflow of passive printer forensic methods.

Overall, mentioned digital passive forensic methods achieve an identification accuracy over 90% for the assignment of the *printer model* used to create the document.

Weakness of Intrinsic Printer Signatures However, the complex printing process and environmental influences not only produce usable distinguishable signatures but even could change these signatures themselves. Compared to other device-specific signatures such as of cameras or scanners [186, 459], the printer space is very variable. Individual printer components can/must be replaced over time, e.g. toner cartridges or OPC drum, whereby original parts do not necessarily have to be used, but a variety of third-party suppliers exist. Additional changes in the printout due to the substrate used (e.g. plain vs. recycled paper), various drivers and their settings (e.g. toner save mode or resolution), age and fill level of the toner, general wear and tear over time, ambient temperature, different font types/sizes, or even the influence of the scanning device could potentially modify the extracted signature. Whether unconsciously or as a conscious attack. Only very few analyses take partial areas of these influences into account [315, 313, 159, 230]. Overall, existing signatures have not sufficiently been tested for their robustness against such influences, although this is important to judge the practical applicability for printer forensics and thus for verification of analog information, as well as to assess the privacy risks posed by intrinsic printer signatures. Moreover, in most studies, the accuracy rate is evaluated with a test set of only 5-20 different printers, which provides only a minimal indication of the sensitivity when used in practice.

3.2.2 Halftoning as Intrinsic Signature of EP Printer Models

As described in Section 3.1.3, laser printer commonly use AM Halftoning, which creates spots that are arranged in regular columns and rows and are equally spaced. A problem of such periodic dot placement is the vulnerability to interference effects like moiré pattern, which appears if periodic structures are superimposed.

To reduce such interference artifacts the overlapping color screens are arranged in different angles

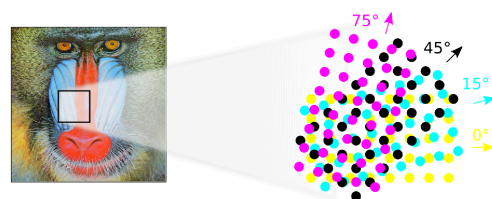


Figure 3.5: Arrangement of AM color screens.

(see Fig. 3.5). The best angle between two screens that is least likely to cause such artifacts is 45° . However, in four color process printing all four screens must be angled within a 90° limitation (above the screen would repeat itself). Typical angles are for example 15° , 75° , 0° , 45° for C,M,Y,K.

Ryu et al. [380] proposed that the arrangement of these screens is printer dependent and usable as intrinsic signature for identifying brand and model of a color laser printer. Therefore they developed a method to extract the angles of each color screen. They used nine color laser printers with 600dpi to print 40 color images each. The printed images were scanned with 2400 dpi and 25 randomly selected sub-regions (128x128px) were extracted and transformed into CMYK color domain. After binarization of each channel by applying Otsus Threshold [348] they used Hough tranformation [414] to estimate the angle values. These values were merged into a histogram. Half of the histograms of a printer were averaged and then used as a reference pattern. The remaining histograms were tested by correlation with the reference pattern. Kim and Lee [245, 246] improved this identification scheme by analyzing the entire halftoning texture of each channel instead of only measuring the angles. Hence the used frequency and spot function are also observed, which enlarges the range of differentiation. They also added the possibility to extract and evaluate the halftone texture with a common camera instead of 2400 dpi scanning. With this approach they reached an average accuracy of 86.14% to identify the printer model of a printed document, out of five models from two brands [246].

Overall, the low complexity and the fact that it is possible to extract forensic statements with simple smartphone cameras, makes the identification scheme interesting for widespread use. However, the forensic practicability as well as the question of whether this represents a problematic gap with regard to the privacy of creators of analog documents, i.e. whether transparency and awareness are necessary in this area, depends mainly on the effectiveness of the intrinsic signature. However, the authors did not evaluate the stability of the halftone signature with regard to potential influences. Thus we re-investigate this signature regarding its robustness and resistance to forgery of this signature and assess its sensitivity. In addition, we investigated whether the signature could also be used for prints that contain solely text. In the following we report about our experiment setup and results of the re-investigation.

3.2.2.1 Experiment Setup

To be able to re-investigate the halftone signature, we created two test patterns T1 and T2 (Fig. 3.6). T1 was created for testing the robustness. It consists of two images and 19 squares with different colors (e.g. CMYK colors), to extract the halftone texture as clearly as possible. T1 was then printed several times with different types of paper (recycling vs plain 80g), toner cartridges (third vs. original), operating systems (Windows 8.1 and Debian 8), drivers (e.g. Linux PPDs, Universal Drivers, Original Brand Driver), as well as different driver settings to evaluate the stability of the halftone signature regarding these parameters. T2 was prepared to explore the possibility of applying the scheme on colored text. It contains sentences in different fonts (Arial, Times New Roman) and colors and also some colored icons to compare the image and text halftoning.

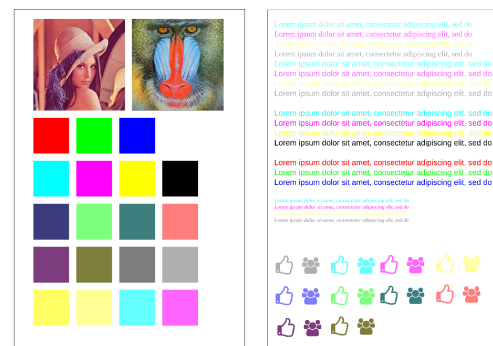


Figure 3.6: Test pattern T1 (left) and T2 (right).

Overall we examined 206 prints obtained by nine color laser printers from four brands (see Tab. 3.1). We used printers of the same model (P4, P7), and similar printers of the same brand (P6, P7) to assess also the sensitivity of the signature. All prints were scanned with an Epson Perfection V30 Scanner at 800 dpi resolution.

Table 3.1: List of printers used in our experiments

Label	Brand	Model	# devices
P1	Kyocera	P6021 cdn	1
P2	OKI	MC 361	1
P3	HP	CLJ M553	1
P4	HP	CLJ 5550	2
P5	HP	LJ Pro CM1415fnw	1
P6	Ricoh	MP C3003	1
P7	Ricoh	Aficio MP C305	2

The color squares of T1 were deskewed and extracted with morphological transformations and contour finding. Out of the squares a subregion ($128 \times 128 \text{px}$) was extracted and the halftone texture was analyzed. To estimate the angles of the different patterns we used the Probabilistic Hough Line Transformation [301]. All image processing operations were realized with the help of the OpenCV Framework³. Moreover, all texture patterns were evaluated and compared manually.

3.2.2.2 Robustness of the Halftone Signature

As expected, using different toner cartridges or types of paper did not influence the intrinsic signature. The halftone texture was also stable regarding different printers of the same model (tested with P4 and P7). However, our results show that the halftone screen is not fixed for a specific printer model. In many cases, the models realize different halftone screens for different driver settings. One of the most influential driver options, measured in our experiment, was the *resolution* setting. Many printer models provide different resolutions, e.g. for fast vs. high quality prints. The impact of the resolution onto the screen angles can be seen in Table 3.2.

Table 3.2: Different screen angles through different resolutions

Printer	Resolution	C	M	Y	K
P1	600 dpi	108	72	152	45
P2	600 dpi	75	18	45	45
	1200 dpi	51	128	27	0
	ProQ	104	161	45	45
P3	All	162	16	56	36
P4	All	72	18	0	45
P5	ImageRet3600	75	15	45	50
P6	600 dpi	27	63	0	45
	1200 dpi	18	72	27	45
P7	600 dpi	27	63	0	45
	1200 dpi	18	72	162	45
	2400x600 dpi	27	63	0	117

By varying the resolution, not only the angle but even the spot function and frequency could be changed (see e.g. Fig. 3.7). These changes were consistent regarding different drivers and operating systems.

³<https://opencv.org>

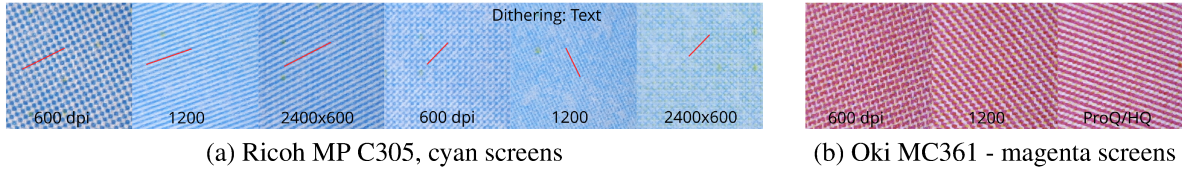


Figure 3.7: Different halftone screens, changed by driver settings.

Additionally, e.g. for the Ricoh Printers (P6, P7), there exist driver options named *Dithering* and *Gradiation*. Each of these settings produces different halftoning screens dependent on the resolution. Figure 3.7 (a) shows also the halftone screen with *text dithering* for P7 at different resolutions. For both models we get overall six different halftone screens through different driver settings. In comparison the HP models (P3, P4) produce the same halftone texture for all tested drivers and settings. P1 supports only 600 dpi resolution and P5 was only tested with a Linux driver that did not support resolution change. This means that a simple change of driver options is mainly sufficient to alter the texture and thereby the intrinsic signature.

Another weakness of the signature is the limited number of useful combinations of screen angles, resulting in possible overlapping of the intrinsic signature between different models. In our test set it can be seen especially within the same brand (P6, P7), but even among different brands (P2, P5), where the spot functions and frequencies were highly similar (Fig. 3.8).

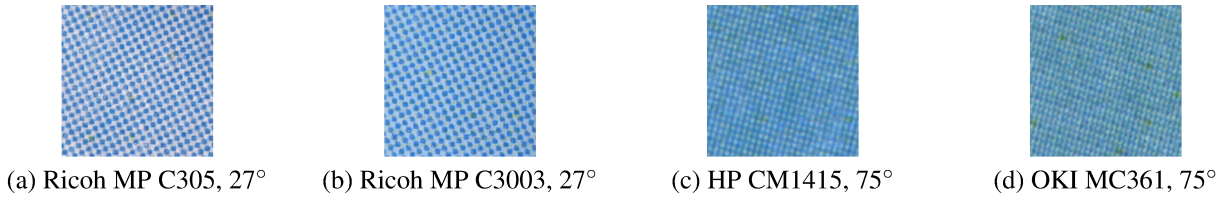


Figure 3.8: Overlapping of halftone cyan screens within the same brand (a,b) and among different brands (c,d).

Furthermore, the extraction of all four halftoning screens is not always possible, e.g. through dark image content. This increases the probability of overlapping (e.g. P2, P4, P5) and negatively affects the identification accuracy.

3.2.2.3 Spoofing the Signature

The halftone angles and textures measured in our experiments (Sec. 3.2.2.2) result from the default configurations of the printer model. But the frequency, angle or spot function are not hardware or device dependent. This means the halftone screen and thereby the intrinsic signature can be manipulated by the user.

For PostScript printers this can be done by sending a postscript file with the corresponding commands directly to the printer. In our experiment we tested this behavior with the *setcolorscreen* command [432]:

```
/sfreq 150 def %150 lpi
/sproc {dup mul exch dup mul add 1 exch sub
} def %round spot function
sfreq 75 /sproc load %75deg cyan screen
sfreq 15 /sproc load %15deg magenta screen
```

```
sfreq 0 /sproc load %0deg yellow screen
sfreq 45 /sproc load %45deg black screen
setcolorscreen
```

Figure 3.9 (a) and (b) show the original halftone screens of the cyan channel of Ricoh MP C305 and OKI MC 361 at 1200 dpi. Figure 3.9 (c) shows a manually set screen with a frequency of 170 lpi, an angle of 18° , and a line spot function, printed by the OKI Printer. As it can be seen this setting is a good simulation of the original halftone screen of the Ricoh printer. Furthermore, in Figure 3.9 (d) the screen is set with 50 lpi, 0° , and a rhomboid spot function printed by the OKI printer. This should visualize the wide range of possible screen settings. Additionally these settings can be changed in the specific PPD file. For non PostScript printers there is the possibility of using a Software RIP or digital halftoning to overlap the original screen and make the extraction more difficult.

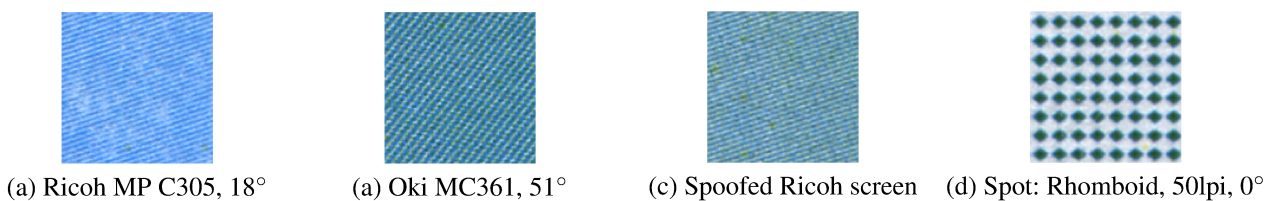


Figure 3.9: Halftone cyan screens of Ricoh Afficio MP C305 and OKI MC361 at 1200dpi and manipulated screens from OKI MC361

3.2.2.4 Document Linking and Application to Text

Despite the possibility of manipulation, the scheme may still be useful to identify forged documents: simple comparison of expected and detected textures indicates if this a claimed printer has been used, or if the printout is forged - either by altering the printer configuration, or using a different printer. It can also help link printouts to the same originating device, if, for instance, several ransom notes exhibit the same halftone texture. We extended our experiments to prints containing only colored text, to assess the applicability of the scheme for this purpose. Our results yield a match of the halftoning texture in colored text to the characteristic halftoning texture of the devices in 78% of the tested cases (cmp. Fig. 3.10). Only the Kyocera and Ricoh printers replace the texture with a different dithering in text. This dithering, however, is also characteristic and can again be used for identification.

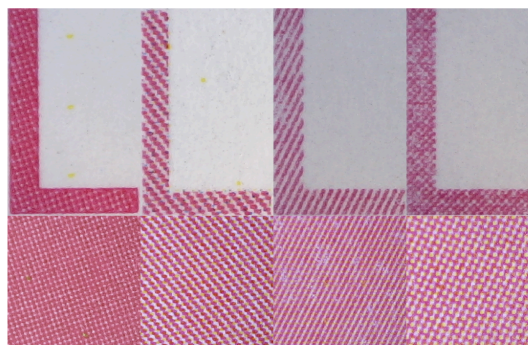


Figure 3.10: Halftone screen of a character and an image from (left to right): HP M553, Oki MC361, Ricoh MP C305 2400x600dpi windows driver and 600dpi cups ppd driver

3.2.2.5 Concluding Remarks regarding the Halftone Signature

Our results show that the halftone texture as an intrinsic signature is stable with respect to different types of paper, toner, operating systems and drivers. However, it is highly susceptible to different driver options, especially to the resolution setting. Furthermore, an attacker can spoof or modify the intrinsic signature. In addition, the sensitivity of the signature is affected by possible overlaps between different printer models. Based on those results we conclude that the halftone texture is not stable enough to identify the source printer of a document without any prior knowledge. For this purpose it would be necessary to determine every possible halftone texture of a printer model, especially considering different resolution settings. Even assuming that users tend to print with default settings, one should be aware of different halftone textures because of possibly different default settings among different operating systems, likewise mobile devices, and drivers.

Nevertheless, it is a promising feature for other forensic applications, like forgery detection or to evaluate the assignment of a set of suspicious documents to one source printer. For such cases the identification scheme can be adapted to colored text. To improve the identification scheme stability and make it more robust against attacks it could be combined with other identification approaches, such as the color noise signature [81].

Overall it can be seen that the analysis of possible influences on intrinsic signatures is important for an accurate evaluation of printer identification schemes and their practical applications.

3.2.3 Discussion: Impacts of Intrinsic Signatures on Privacy and Verification

An important information for verification of analog documents and the solving of crimes is the assignment of a document to its source printer. Intrinsic signatures, identifiable artifacts within printed documents, can help to detect the specific printer model used for the creation of the document with an accuracy rate of over 90%. However, this rate is achieved commonly with a test set of 5-20 different printers. Using it in the wild, with a much larger number of existing printing devices, misallocations may occur, since the signatures of different printers could overlap, as e.g. the analysis of the halftone as well as the banding signature (Appendix I-B) has shown. I.e. for the application and extraction of the used printer model from a document without prior knowledge, a data set of signatures of all existing printing devices would have to be created and maintained. Moreover, there exist many variable parameters, like driver settings, used substrate or toner, etc., which could potentially influence the intrinsic signatures. These parameters would need to be analysed for each printer model signature and incorporated into the database. But even after creation of such a database, the statement from any analog document would be limited to the printer model (including technology and brand) or toner used.

In terms of *privacy*, therefore, arbitrary tracking of analog documents and tracing back to the specific source device and thus to the originator by intrinsic signatures is feasible only at very high cost, if at all. Nevertheless, intrinsic signatures remain very important for the *verification* of analog information - especially when considered as circumstantial evidence in its entirety - remaining limited to use cases with prior knowledge. This includes, for example, the assignment of documents to a suspect or expected device: 'Did suspect X printed these documents?'; 'Are the documents (e.g. certificates, money, ...) printed from the expected device?', or the comparison of documents on the same origin, e.g. 'Are all blackmail letters produced by the same printer?'; 'Are all documents of the contract consistently from the same printing

device?'. This makes it very suitable for evidence collection, for example for law enforcement.

A supplement to the *privacy* problem: with prior knowledge, targeted tracking would be enabled with intrinsic signatures. Clarkson et al. [83] illustrate this with the example of paper-based voting. With their signature based on the surface of the paper it could be possible to de-anonymize paper ballots and thus the voting of individuals. They conclude with 'Anonymous surveys or reporting systems may not in fact be anonymous.'. This targeted tracking with intrinsic signatures is particularly difficult in the sense that it is not recognizable in any way and very difficult to remedy. With printer-specific signatures, however, it is more difficult because either a printer must be forced or the printer of the tracked person must be known. Overall the instability of the printing process and the limitation to the printing model could be a reason for embedding hidden extrinsic signatures within printouts.

3.3 Extrinsic Signatures of Printed Documents - Tracking Dots

In contrast to intrinsic signatures, which only allow statements up to the printer model used, intentionally embedded extrinsic signatures can provide exact statements about the printing process [73]. Such an approach are tracking dots, also named machine identification codes, or yellow dots.

Tracking dots first became known to the public in 2004 through an article in PC World⁴ and the subsequent analysis of the Electronic Frontier Foundation (EFF) [165]. They found that color laser printers print additional tiny and systematic yellow dots on each printout, invisible to the human eye. These are being generated at the firmware level [129] and represent encoded information such as the serial number of the printer or the date of the print [165]. This information can be read and decoded automatically. On the one hand, this is a helpful approach from a *verification* point of view, e.g. to assess the credibility of printed documents or to solve crimes. On the other hand, the embedding of hidden tracking data represents a severe limitation of *privacy*.

A response to a parliamentary question in the European Parliament in 2007 by Satu Hassi on tracking dots⁵ confirms their invasion of privacy: "[...] such processing may give rise to the violation of fundamental human rights, namely the right to privacy and private life. It also might violate the right to protection of personal data."

However, until today, nothing has changed in terms of existence, legal basis or transparency of embedded yellow dots and public attention has been lost. Since the origin and content of these yellow dots are largely unknown, we sought answers from some printer manufacturers. We received e.g. an official prefabricated statement from one manufacturer, in which they called the yellow dots "Document Colour Tracking Dots" (Fig. 3.11). Unfortunately, they were not able to give any answer and referred us to the Central Bank Counterfeit Deterrence Group

Axel.Holtzauer@KonicaMinolta.de Tel.: +49 664 Fax: +49 664 Langenhagen, Montag, 8. März 2010

— **Herstellereklärung Document Colour Tracking Dots**

Sehr geehrte Damen und Herren,
leider müssen wir Ihnen mitteilen, dass wir als Hersteller nicht in der Position
sind Auskünfte über das

"Document Colour Tracking Dots" zu geben.

— Ich möchte Sie darum bitten, sich an folgende Institutionen zu wenden.

CBCDG (Central Bank Counterfeit Deterrence Group)
Ms. Maureen Carroll
CBCDG Director
tel: +1-613-782-7947
e-mail: info@rulesforuse.org

oder

Deutsche Bundesbank.

Mit freundlichen Grüßen

Figure 3.11: Statement of Konica Minolta regarding Yellow Dots.

⁴https://www.pcworld.idg.com.au/article/8305/dutch_track_counterfeits_via_printer_serial_numbers/

⁵https://www.europarl.europa.eu/doceo/document/E-6-2007-5724_EN.html?redirect

(CBCDG), which also could not answer our request because tracking dots are "not a CBCDG product/technology". Further statements can be found in Appendix I-C. The most common answer was the mandatory labeling of printouts in order to track illegal activities such as counterfeiting. A regulation sounds plausible, as some printer manufacturers have probably signed an agreement with the U.S. Secret Service to fulfill "document identification requests" which might caused the tracking dots⁶.

However, since tracking dots represent a major gap between privacy and security and their current state remained mainly unclear we investigated them by ourselves. Overall, we found 6 distinct *tracking dot pattern* (TDP). We introduce three pattern to the public for the first time, analyse the code and structure for each TDP, completely decode one pattern, explain the information in one further code word from Pattern 4, and report about found correlations. Additionally, in Section 3.3.3 we explore an anonymization approach to prevent arbitrary tracking and in Section 3.3.4 we present our declaration TET DEDA which implements the entire workflow for extracting, analyzing, and anonymization of tracking dots. Finally, we discuss the social reactions to the publication of the results and report about an unknown type of tracking dots which we discovered in Appendix I-E.

3.3.1 Investigation of Tracking Dots - Background and Approaches

This section describes necessary definitions, related work, our data set used, and our approaches developed for extracting and analyzing yellow dots.

3.3.1.1 Definitions

The *tracking dot matrix* (TDM) is one prototype of tracking dots in a matrix of $n_i \times n_j$ cells which is printed repeatedly over the whole sheet of paper with a cell distance of Δ_i inches horizontally and Δ_j inches vertically. Each cell of the matrix stores one bit where a yellow dot represents "1" and an empty space represents "0".

A *tracking dot pattern* (TDP) is a format of storing tracking information. It uses a certain code, produces a TDM of a certain size and may include marking dots and a mask of empty cells. The TDP of a printer can be described as $(n_i, n_j, \Delta_i, \Delta_j)$.

Such a certain binary code could be algebraic or non-algebraic. In *algebraic* codes, A is the set of all code words. An information word a^* can be encoded in a code word $a \in A$ so that it is possible to detect or even correct a certain amount of erroneous bits. This is helpful when transferring data via a distorted channel such as yellow dots on a sheet of paper. A code described by the parameters (n, l, d_{min}) encodes information words of length l and adds $k = n - l$ redundant bits to the code words of length n . The amounts of ones in a code word a is called weight and noted as $w(a)$. The minimal weight among $2^l - 1$ nonzero code words is the minimal Hamming distance d_{min} . Binary codes can detect $f_e = d_{min} - 1$ errors in a distorted word $b = a \oplus e$ where e is the error word. A distorted word can only be reconstructed if the error word e has a weight of $f_k = \lfloor (d_{min} - 1)/2 \rfloor$ or less.

An even *parity code* $(n, l = n - 1, d_{min} = 2)$ is a systematic code where the parity bit k is calculated by $k = \bigoplus_{i=1}^l u_i, u_i \in \{0, 1\}$. The weight of such code words is always even. An odd parity code is a parity code where the parity bit is $k = k \oplus 1$.

⁶<https://www.scribd.com/doc/81897582/microdots-pdf>

A *product code* (n, l, d_{min}) with an interconnected block interleaver is a code chain that consists of an outer code $(n_1, l_1, d_{min,1})$, an interleaver and an inner code $(n_2, l_2, d_{min,2})$ where $n = n_1 \cdot n_2$, $l = l_1 \cdot l_2$ and $d_{min} \geq d_{min,1} \cdot d_{min,2}$. An outer code writes code words of length n_1 row by row into a matrix before an inner code reads the words column by column of length l_2 and encodes them [249].

"*One hot encoding*" $(n, n, 2)$ is a type of *non-algebraic* constant-weight code. It consists of one "1" and $n - 1$ zeros so that the weight of a code word a is $w(a) = 1$. Any error word e can certainly be detected with $w(e) \in [1..n] \setminus \{2\}$.

3.3.1.2 Related Work

The first investigation of tracking dots was done in 2005 by the EFF [165]. Using manual analysis, they were able to decode the structure and the content of one specific tracking dot pattern (Pattern 4). They found that it contains the printer's serial number as well as date and time of the print. In addition, they developed a first TET regarding this pattern. However, this required the user to identify and extract the pattern from a printout and then input the analog dots into a digital matrix to finally obtain the decoded content.

Based on this work van Beusekom et al. [228, 455] described a method to automatically extract the TDP's width and height and concluded the TDP class from it. They detected three different TDP classes and their rotations. This method succeeds with an accuracy up to 93% for detecting the TDP class. However, they did not attempt to decode the content or code structure of the patterns. Instead, they developed a method for comparing two prints to analyse if they are printed from the same printer or not. To achieve this, they extract the TDM as an image, using a pixel threshold to match a TDM's repetitions on a sheet. They reached an accuracy of 91% for comparing different printouts with regard to their source device. However, for later decoding and useful transparency, numerical TDMs are needed.

When scanning a document, it is typically skewed by few degrees. Van Beusekom et al. [454] introduced a method for deskewing using geometric text-line modelling. This method was also used for their automated TDP analysis. However, this is not applicable for documents with very few text.

Embar et al. [129] developed a method to obfuscate TDM's by filling the page with a grid of yellow dots. In a practical example the points were embedded as the background of a Word document. When printing the whole page in yellow, some printers were found to produce white dots instead.

3.3.1.3 Dataset for Analysis

For the investigation of tracking dot patterns and development of extraction methods, a sufficiently large data set is necessary. However, due to the necessity of accessing a large number of printing devices, the creation of this proves to be difficult (time-consuming, costly). Data sets in the print area are therefore mainly based on crowd-based collections. By means of open calls, users are asked to scan test printouts and submit them with additional meta data.

An existing data set, which was created crowd-based, is the MIC data set⁷. It was compiled by the EFF in collaboration with the DFKI. It consists of scanned printouts of 132 different printers, from the period between 2005 and 2010. It provides the printer's manufacturer, model and serial number.

⁷<https://madm.dfki.de/downloads-ds-mic>

In order to be able to include more current models in the data set we have launched our own open call in 2020 in cooperation with the artist Wolfgang Plöger⁸. This enabled us to collect 40 new printer sets⁹. The meta data for this were again labeled by normal users. It includes information about the manufacturer, model, serial number, date and origin.

As the crowd-based approach can be prone to errors, e.g. incorrect labeling, we have also generated our own printouts of overall 22 printers. Thus, in addition we have clearly labeled samples whose source devices can also be further examined, if necessary. These samples were obtained from local organizations (esp. printers in the university, library or copy stores). At these locations, larger quantities of printers of the same type were used, which often corresponded to the same order and thus had similar serial numbers. Each printout has been digitized with Epson Perfection V30. The content of the documents consists of either images or text (Fig. 3.6 in Sec. 3.2.2.1). Next to brand, model, and serial number, this data additionally contains the date, information about the used driver, resolution and toner.

Overall, the combination of these three data sets enabled us to investigate 1515 prints by 135 printer models from 19 different manufacturers, a total of 194 printers (Tab. 3.3). This covers the majority of well-known printer manufacturers.

Table 3.3: Data set by manufacturer

Manufacturer	# Printers	# Models	# Prints	Has Dots	Found Pattern
Brother	4	4	16	✓	2
Canon	26	18	153	✓	2/5
Dell	5	3	40	✓	4
Epson	9	8	73	✓	3/4
Hewlett-Packard	47	20	421	✓	2
IBM	1	1	9	✓	?
Konica Minolta	29	20	216	✓	3
Kyocera	6	5	37	✓	2
Lanier	1	1	9	✓	1
Lexmark	7	6	57	✓	2/6
NRG	1	1	9	✓	1
Okidata	10	7	92	✓	2
Panasonic	1	1	1	✓	2
Ricoh	9	7	89	✓	1/2
Samsung	8	7	58	×	
Savin	1	1	12	✓	1
Tektronix	4	4	35	×	
Unknown	1	1	9	✓	3
Xerox	24	20	179	✓	4

3.3.1.4 Tracking Dot Extraction

This section describes our first method on reading arbitrary TDP and transforming a sheet into a list of TDM for further analysis of previously unknown TDP. The goal was to map the analog tracking dots into a digital matrix.

First, the empty areas of the document must be detected, as the yellow dots in these areas are visible. Therefore we mask the printed areas using Gaussian Blur and a global threshold. After a color space conversion to HSV and exposure of the yellow color range, the set D of all recognised yellow dots is extracted by a contour detection algorithm [430].

⁸http://www.wolfgang-ploeger.com/open_call.htm

⁹www.dataset-tracking-dots.com/dataset_overview.html

Next, the page needs to be aligned so that the tracking dots can be separated by straight lines into a grid. Because of the manual scanning process, the sheet might have been skewed by α degrees and must be corrected by a rotation of $-\alpha^\circ$. It is possible to correct a skew up to 45° by taking advantage of the fact that on the sheet a TDM is being repeated many times in a straight line. Remember that the set D of yellow dots might be distorted. To approximate α , we calculate the angles between each two dots from D , quantise them and find the most occurring value.

When mapping the dots into a matrix, the cell separating grid might be shifted due to inaccuracies caused by a limited scan resolution and therefore skip a column or row each few centimetres. To prevent this, cropping the sheet to a square of 7.5 cm (3 inch) per side is recommended. Tracking dots extraction can be applied to all possible cropped squares from a page.

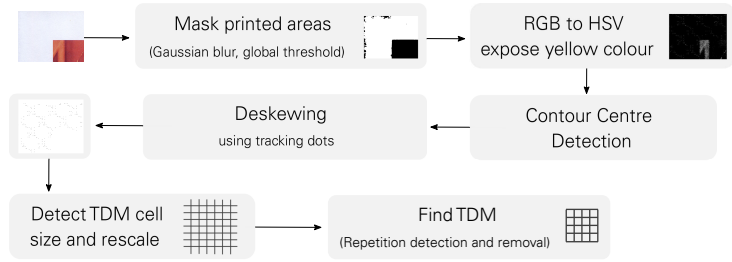


Figure 3.12: Pre-processing steps of the extraction workflow.

Afterwards the tracking dots are mapped from the page into a grid. In a matrix, all cells of the same column are exactly one below the other. Due to imperfections in the scanning and/or printing process, the x coordinates of dots from the same column vary slightly. We call this bias. Let's assume Δ_1 and Δ_2 are the two smallest local maxima of the neighbouring dots' horizontal distances' frequency with $\Delta_1 < \Delta_2$. Δ_1 typically is the biased distance between dots of the same column and Δ_2 is the distance between dots of neighbouring columns: $\Delta_i = \Delta_2$. The vertical dot distance Δ_j can be calculated analogously. The grid shall be placed in a way so that the most occurring x coordinate of D is in the center of a cell and the most occurring y coordinate is in the center of a row. The cells have the size $\Delta_i \times \Delta_j$ inches. The tracking data yields '1' where there is a yellow dot in the grid cell and '0' everywhere else.

As a TDM is printed repeatedly over the whole sheet, its dimensions could be detected given the matrix of all yellow dots. Let's assume a function that calculates the likelihood of two columns being identical. Then for each column c we calculate the median distance to each column \hat{c} where the likelihood that the content of c and \hat{c} is above a given threshold. The most occurring distance is assumed to be horizontal separation distance n_i cells. The vertical separation distance n_j can be calculated analogously. If the sheet is being cut into pieces where each contains $n_i \times n_j$ cells, a list of (possibly distorted) TDMs result. For TDPs using a redundancy code, all TDMs shall be removed from that list where the redundancy check fails. Otherwise a TDM prototype can be estimated by overlapping all found TDMs and setting the value '0' or '1' by a majority decision.

3.3.1.5 Analysis Methods on TDPs

Due to the repetitions of a pattern over the entire printout, each pattern probably contains marker of its beginning. To find possible markers, we overlapped all TDMs of the same pattern from different printers such that the resulting matrix shows only a dot where all matrices show a dot (intersection). Dots that appear in all TDM samples do not contain information and can be used as orientation markers therefore (red in figures).

Furthermore a TDP may contain empty cells, rows or columns that need to be skipped when reading the data. To determine them, all TDMs of one pattern were overlapped so that the resulting matrix shows a dot where at least one matrix shows a dot (union, see Fig. 3.13). Thereby, the cells become visible that are empty on all TDMs. They may mark spaces / separators between data blocks. If there is only one dot in each of these blocks of size n , then its data may be stored in a “one hot encoding” of length n – written row by row or column by column.

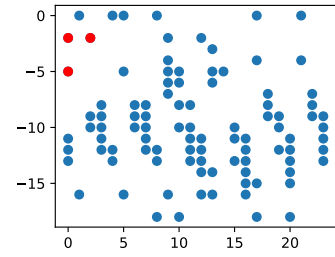


Figure 3.13: All matrices of pattern 3 united.

The information stored in a matrix was attempted to reveal by analyzing the inference of metadata (e.g. the printer’s serial number) to the matrix.

For the printers, for which both the serial number (or other possible device ids) and the TDP were known, a known-plaintext attack could be achieved. Same properties of printers with the same pattern (same characters in the serial number, same model number, ...) were used to find correlations within the pattern in order to possibly decode it. In order to recognize dynamic content such as date or time within the pattern, different printouts from the same source device were compared. If the pattern remains constant for each print, we conclude that it only contains fixed information.

3.3.1.6 Improving the Extraction

After in-depth analysis of the existing patterns, we can use the findings about structure, content and redundancy checks, to improve the extraction algorithm.

To find a valid TDM, the tracking dots must be extracted. Here, pre-processing steps are similar to the first extraction algorithm (Fig. 3.12). After masking printed areas, exposing yellow color and contour detection, the scan is align based on detected dots. If the amount of dots per inch is realistic (min/max), the sheet is cropped into squares of 3x3 inch, to minimize scan artifact influences.

Afterwards, for each known pattern, tracking dots are mapped from the page into a grid, which is setup-ed based on the patterns known dot distances (cell size of $\Delta_i \times \Delta_j$; see also Tab. 3.4). Afterwards each square is split into matrix subsets based on the known TDM-sizes ($n_i \times n_j$). I.e. the biggest necessary size to extract the TDM. These subsets could overlap.

Then any repetition of the TDM must be selected from the sheet. Initially it must be shifted so that the marking dots are on the desired place (top left corner). If an offset TDP does not provide marking dots, the matrix can be shifted according to its empty space. An offset repetition has to be removed from the extracted matrix if it contains any. Next, markers and spaces have to be removed from the matrix so that it only contains cells that belong to the code word. The result’s redundant bits can be checked according to the code’s description. If the check fails, the algorithm has to be repeated using another prototype from the TDM repetitions on the sheet until a valid TDM has been found.

Depending on the pattern, several valid identical TDMs must be found until it reaches a pattern score. This is important especially for patterns without parity bits (like 3 and 5), where a slipped dot produces a valid TDM with a valid “one hot encoding”, but the extraction would be wrong.

The whole process is done in an iterative round robin mode, for each known pattern characteristic. If there is no pattern match which full fills all redundancy checks and reaches the score, the whole process is repeated with another yellow color range. If there are valid dot distances found, the extraction is improved by firstly analyse the whole sheet regarding the pattern which fits best on these distances.

3.3.2 Results of Tracking Dot Pattern Analyses

Six different Tracking Dot Pattern (see Tab. 3.4) were detected in our data set using the proposed extraction algorithm (Sec. 3.3.1.4) and manual examination.

Table 3.4: Dimensions of detected TDMs

Pattern	n_i	n_j	Δ_i	Δ_j
1	32	32	0.02 in	0.02 in
2	18	23	0.03 in	0.03 in
3	24	48	0.02 in	0.02 in
4	16	32	0.04 in	0.04 in
5	16	16	0.02 in	0.01 in
5s	16	32	0.02 in	0.01 in
6	8	21	0.04 in	0.03 in

Table 3.5: Patterns by Manufacturer

Pattern	Manufacturers
1	Lanier, NRG, Ricoh, Savin
2	Brother, Canon, Hewlett Packard, Lexmark, Okidata, Ricoh, Kyocera, Panasonic
3	Epson, Konica Minolta
4	Dell, Epson, Xerox
5	Canon
6	Lexmark

The patterns may appear rotated (90° steps) and/or flipped. The companies Lanier and Savin belong to the company of Ricoh [373] which use Pattern 1 together with the company NRG. Pattern 2 is being used by five different independent manufacturers (see Tab. 3.5).

All detected TDPs were analyzed and some matrices decoded. The patterns were evaluated according to their information density, capacity, error detection rate and conspicuousness. We also analyzed the number of yellow dots generated by each pattern in the best, worst and average case, depending on the content of the TDM.

For a TDM, let *col* be the column and *row* the row index number. All patterns use a kind of repetition code because their matrices are spread over the whole sheet of paper. Therefore, forward error correction can be achieved using the repetitions of a matrix. The amount of repetitions depends on the size of the printed area and the size of the matrix. Each section relates to only one prototype of the matrix. The same statements always apply to its repetitions.

3.3.2.1 Pattern 1

The first pattern is printed offset (Fig. 3.14). Therefore its dimension is detected as 32×32 cells although the unique matrix with the spacing uses $32 \times 16 = 512$ cells. This section deals with the prototype of the matrix in rows 0-15 and columns 0-15. The pattern marks its beginning with two neighbouring dots (red in figure) and stores information in every second column in every second row. All even rows do not contain any dot except the marking ones. Each row is one code word. Let s be the index of the first column that contains the first code word bit in the row. s is either 2 or 3 depending on the printer. In our figure s is 3. This pattern has been discovered on nine different devices.

Redundancy check The pattern uses a $(7,6,2)$ even parity code. It stores eight code words row by row which contain $8 \cdot 6 = 48$ information bits in total. A TDM is considered as valid if the amount of dots is even in all rows. Error words with an even weight produce code words. To detect them, all valid TDMs have to be compared and chosen by a majority decision. Row 31, for example, contains the code word (1100000) and passes the parity check. The correct word might have been (0000000) as well united with the error word $e = (1100000)$. This error with $w(e) > 2$ cannot be detected.

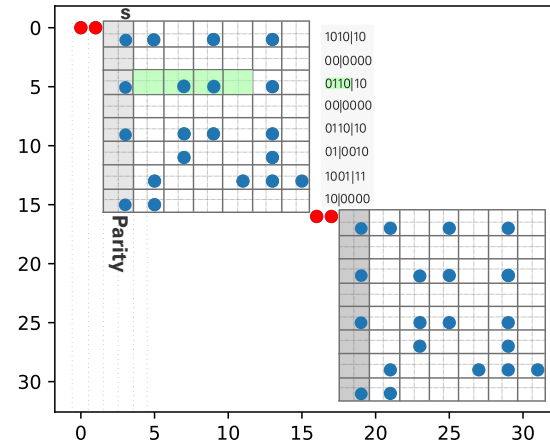


Figure 3.14: Pattern 1: Marking (red) and other tracking dots (blue).

Decoding The pattern contains the printer's serial number as 4 binary bit blocks in the (7,6,2) even parity code. Being a systematic code makes it easily readable. For the rows 1, 3, 5, ..., 15, column $s = 3$ contains the parity bit and the information bits can be found in every second column from $s + 2$ to 15. A binary chain has to be read from left to right starting with the bottom row. Each 4 bits of this chain represent a binary number. The resulting 11 binary numbers before the last one represent the printer's serial number. The first and fifth number may represent letters, where "9" stands for "P" and "0" stands for "W" or "Q".

Example: Figure 3.14 contains the words (1101010), (0000000), (1011010), (0000000), (1011010), (0010010), (0100111) and (1100000). The information bits without the leading parity bit are (101010), (000000), (011010), ..., (100000). Splitting this chain into 4 bit chunks results in (1010), (1000), (0000), (0110), ..., (0000). Reading the chain as well as the chunks backwards and transforming them into decimal numbers gives us the string "079496016015". The serial number of this printer is W794P601601.

Conspicuousness The amount of dots per code word a is determined by its even weight $w(a) \in \{0, 2, 4, 6\}$. From 64 code words with an equal probability of occurrence, 35 have a weight of 4 (54,7%)¹⁰. This makes eight code words \cdot 4 dots + 2 marking dots = 34 dots per matrix at average (0.67 dots per bit).

3.3.2.2 Pattern 2

Pattern 2 uses $18 \times 23 = 414$ cells (Fig. 3.15) and was found on 59 devices. Depending on the printer, each dot in the figure is represented by one or two printed dots. Three dots in the first two rows mark the beginning of the pattern (red in figure). The TDM consists of eight blocks named A to H (from left to right) situated in rows 2-6, 7-11, 12-16, 17-21 and columns 1-8 and 10-17. Row 22 as well as columns 0 and 9 are separators. The pattern considers every second column in rows with an even index and every first column otherwise, so all cells are being considered where $(col \bmod 9 + row) \bmod 2 = 0$. This pattern stores four signs from a (4,4,2) "one hot encoding" and interleaves it in a (5,4,2) odd parity code - for each block. This results in a (20,16,4) product code where each code word can differentiate between 4^4 different states. The pattern consists of eight code words, so it stores $4^{4 \cdot 8}$ different states in total. This is equivalent to storing 64 bits.

¹⁰Calculated with binomial distribution

Example: Block G contains

0010
0010
0001
0010
1100

the last row is the parity.

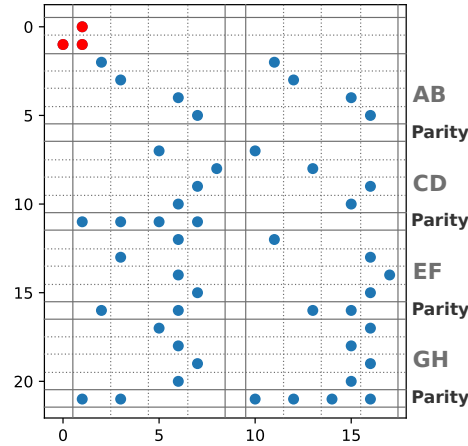


Figure 3.15: Pattern 2: Marking (red) and other tracking dots (blue) aligned into blocks A-H with parity. The estimated grid has been added to this figure for readability.

Redundancy Check Each of the first four rows of each block contains information bits as a “one hot encoding”. The fifth row contains the parity bits of the outer encoding which make an odd amount of dots in each column of each block. This helps to detect errors e with $w(e) = 2$ which are not detected by the “one hot encoding”. Each inner code word contains exactly one “1”, so 1, 3 or 4 faulty bits can be detected. The product code can detect any number of faulty bits that is 2 or odd. Moreover, error correction is possible.

Example: If $a = (1000\ 0100\ 0010\ 0001\ 0000)$ from block A is being distorted with an error word $e = (1100\ 1100\ 0000\ 0000\ 0000)$ then e is one of the few error words with weight 4 that produces a code word which matches the parity bits with $b = a \oplus e = (0100\ 1000\ 0010\ 0001\ 0000) \in A$.

Decoding To obtain the information part from the matrix, the first four rows of each block can be interpreted as a number in \mathbb{Z}_4 . For example Block A from Figure 3.15 contains the information bits $(1000\ 0100\ 0010\ 0001)$, which then represent the numbers 3, 2, 1, 0.

Block A equals block B in all samples. We found that they correlate to the printer’s manufacturer. Hence, from the information in block A, the printer manufacturer can be concluded (see Tab. 3.6). Hence, the example with 3,2,1,0 represents a printer of the manufacturer Okidata.

Table 3.6: TDM’s block A and B by manufacturer

Manufacturer	Brother	Canon	HP	Kyocera	Lexmark	Okidata	Ricoh	Panasonic
Block A	2301	3021	3021	0123 / 2013	0213	3210	2310 / 0132	2103

Obviously there has been a preference for chains of distinct digits to identify the manufacturer. The advantage of these numbers is that they produce less dots. Only if a block consists of all distinct numbers,

the amount of added parity dots is minimal. For Kyocera and Ricoh printers there were found two distinct identifiers which match the manufacturer. There was a small set of Canon printers which used Pattern 2 instead of Pattern 5. All of these patterns have the same identifier as HP. At this point, we were undecided whether the correlation still held true, however we found that these manufacturers are in cooperation¹¹, which could explain the duplication. Further, all other printers were always clearly assignable to the manufacturer by block A and B.

Regarding block C, we found a clear correlation to the printer model across the entire data set (see Tab. 3.7). Some models share the same information, however these models are very similar. Perhaps this information is more coarse-granular and shows the specific model series.

Table 3.7: Known models and their corresponding block C information

Manufacturer	Block C	Models (Number of Printers)
Canon	3230	i-SENSYS MF8080Cw (1)
HP	3120	LaserJet 5500 (6)
HP	3031	LaserJet M479fdn (1)
HP	3021	LaserJet 500 M551 (1)
Ricoh	3001	Aficio CL3000 (1)
HP	2321	LaserJet M477fdn (1)
HP	2220	LaserJet 4650 (2)
Canon	2130	LBP7660C (2)
HP	2120	LaserJet 4600 (10)
HP, Brother	2031	HP LaserJet M281fdw (1), Brother MFC914C-CDN (1)
HP	2021	LaserJet M1415fnw (1)
Ricoh	2000	Color Laser AP206 (1)
HP	1320	LaserJet 5550 (4)
HP	1220	LaserJet 3500 (1), LaserJet 3700 (6)
Brother	1131	MFC-L3750CDW (1)
Kyocera	1123	FS-C5020N (1)
HP	1120	LaserJet 9500 (1)
Kyocera	1111	P6021 (1)
Kyocera	1031	Ecosys M6635 cidn (1)
Kyocera	1001	MC361 (1), C841 DN (1)
HP	0321	M553 (1)
HP	0320	LaserJet 2550 (4)
Kyocera	0031	Ecosys M6630 cidn (1)
Lexmark	0231	910 (1), C912 (1)
HP	0220	HP LaserJet 2500 (3)
Panasonic	0220	DP-C264 (1)
Kyocera	0200	C2630D (1)

Printers using this pattern have serial numbers like CNBB002529, CNBC55MOPR, or JPGMC52527. The information of blocks D-H is uncertain. It may contain encrypted digits from the serial number, random hashed values or some other data. Overall, we could not find any correlation of these blocks to serial information. Date and Time are not included.

Conspicuousness The amount of dots per code word ranges from four to eight. If one word produces four dots, it consists of four different inner code words and all parity bits are 0. In the worst case one word produces eight dots: This occurs when all parity bits are being set to 1. The average amount of dots from all code words is six. There are $4^4 = 256$ different code words per block. 192 of these either contain three identical numbers and one different one or two identical numbers and two other numbers of which both differ. Both cases produce exactly two parity bits which are 1. All code words contain exactly four dots from the information bits. This sums up to 4 information dots + 2 parity dots = 6 dots. For the whole pattern this means $8 \cdot 6 + 3$ marking dots = 51 dots at average (0.80 dots per bit).

¹¹<https://www.hp.com/de-de/hp-news/press-release.html?id=33932>

3.3.2.3 Pattern 3

Pattern 3 (Fig. 3.16) consists of 27 blocks of six bits where each block uses two columns and three rows. The pattern's beginning is marked by three dots (red in figure) which do not fit into one block. The pattern is being repeated over the whole sheet. Its detected shape is 24×48 cells because each vertical repetition of the pattern is being shifted by +8 columns. The unique pattern consists of $24 \times 16 = 384$ cells. 28 devices with this pattern were found in our data set and is used by Epson and Konica Minolta.

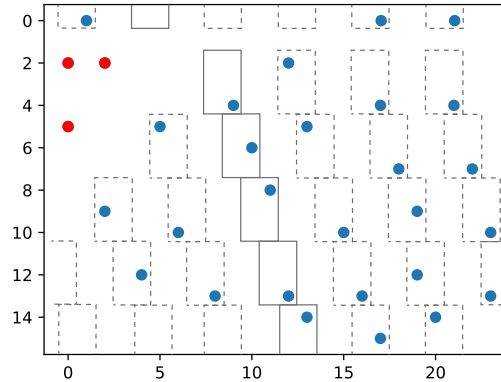


Figure 3.16: Pattern 3: Marking (red) and other tracking dots (blue). Rectangles have been added to this figure to mark our detected code word blocks. The solid boxes indicate the pattern's offset.

Redundancy Check The source alphabet contains six elements which are being encoded with a (6,6,2) “one hot encoding”. The pattern consists of 27 blocks where each block stores one code word. A code word weights “1” and therefore produces exactly one tracking dot.

Decoding In our first investigations, we could not find correlations with any of the printer's known properties. However, based on the improved data set (see Appendix I-E), we compiled a set of printers with Pattern 3 sharing similar serial numbers. Overall, serial numbers of printers with Pattern 3 could have different structures. The first set contains of four different Konica Minolta Magicolor 2300DL printers. Their serial numbers consist of ten digits, where the first four are identical and the following six are different (5311299678, 5311071099, 5311041279, 53111038848). The second set contains of eight Konica Minolta bizhub C250i printers. Here serial numbers consist of 13 digits, where the first nine are identical and the following four are different (AA2M021002826, AA2M021000969, AA2M021002388, AA2M021003041, AA2M021001337, AA2M021001024, AA2M021003604, AA2M021003921). Figure 3.17 shows the dot assignments of the TDMs as heat maps for general Pattern 3 (34 printers), as well as for the two sets mentioned above. It turns out that the more similar the serial number, the more similar the TDMs. This means that in comparison to Pattern 2, serial number information is stored plain. Nevertheless, we have not yet found a method to read the serial number from the TDMs. Analysed with different prints on different dates, we conclude that date and time information are not included in Pattern 3.

Conspicuousness Considering the markers, there are $27 + 3 = 30$ tracking dots in total (0.43 dots per bit). The pattern can differentiate between 6^{27} states in total. This is equivalent to storing

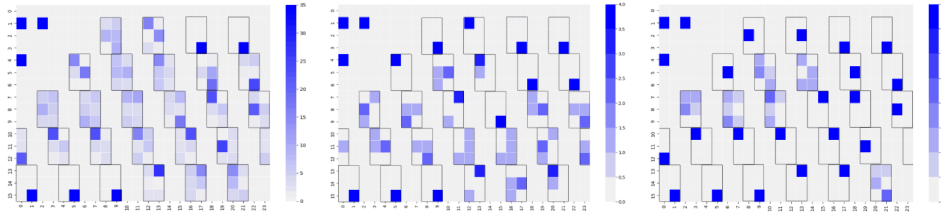


Figure 3.17: TDMs as heatmap of: 34 different printers with Pattern 3 (left); 4 Konica Magicolor 2300DL with similar SNRs (middle); and 8 Konica bizhub C250i with highly similar SNRs (right).

$$\ln(6^{27})/\ln(2) \approx 69.8 \text{ bits.}$$

3.3.2.4 Pattern 4

Pattern 4 (Fig. 3.18) is being used by Dell, Epson and Xerox printers. According to a research fellow at Xerox, the U.S. government and his company have a “good relationship” [451] which might be the origin of this pattern. Pattern 4 uses $16 \times 16 = 256$ cells and is being repeated offset (16×32 cells in total). This section relates to the matrix in rows 0-15 and columns 0-7. There are three or seven marking dots (sometimes called “separators”) in row 6 although they are missing on some printers. The pattern encodes words with a (8,7,2) odd parity code and interleaves 14 code words in a (15,14,2) odd parity code. This results in a (120,98,4) product code. The parity bits are in row 15 as well as in column 0. For some printers the outer parity does not cover the inner parity bits (e.g. see Fig. 3.18 col 0, row 15). The pattern stores 98 information bits in total. Overall 19 devices in our data set use this pattern.

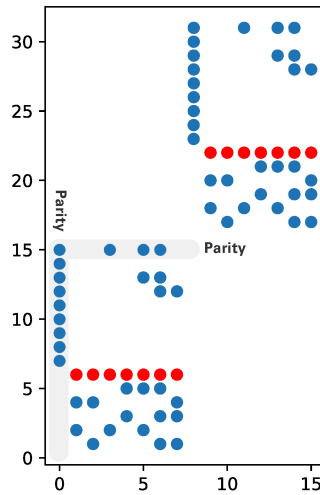


Figure 3.18: Pattern 4: Marking (red) and other tracking dots (blue).

Redundancy Check Code words in rows 1 to 14 as well as each of columns 1 to 7 must show an odd amount of dots. The product code allows error correction.

Decoding Each row has to be transcribed into a binary number excluding the leading parity column. The resulting number can be transformed into the decimal system. The TDM contains the date and time of the print. The manufacturers Epson and Xerox add six digits of the serial number as well. Dell's TDMs within our data set do not include them. The minutes can be found in row 14, the hour in row 11, the day in row 10, the month in row 9 and the year in row 8. The middle of the serial number is a concatenation of the numbers from rows 3, 4, 5. Row 12 correlates with the manufacturer (Tab. 3.8) and row 7 has been constantly empty (except parity bit). The meaning of the information in rows 1, 2 and 13 does not correlate with any of the printer's known features. Row 15 contains parity bits and row 0 is always empty.

Table 3.8: Number in TDM's row 12 by manufacturer

Manufacturer	Dell	Epson	Xerox	Xerox
Row 12	20	3	0	4

Conspicuousness If we assume that the pattern stores an arbitrary serial number and a date where the hour ranges from 0-23, the minutes from 0-59, the day from 1-31, the month from 1-12 and the year from 0-127 then it produces between 12 and 76 tracking dots. At average it generates 46 tracking dots (0.47 dots per bit).

3.3.2.5 Pattern 5

Pattern 5 (Fig. 3.19) is used exclusively by Canon. There are two variants of the pattern, which have the same overall structure. Those patterns have a matrix of $16 \times 16 = 256$ or $16 \times 32 = 512$ cells. We named the second variant Pattern 5s. Both variants are repeated without white spaces / offset in between. Vertically, the pattern is repeated evenly, with a horizontal displacement. This displacement is variable with different printers and could therefore contain information value.

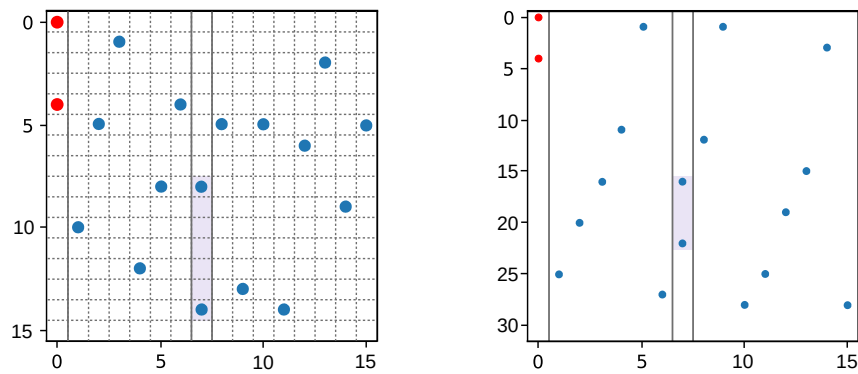


Figure 3.19: Pattern 5 (left) and Pattern 5s (right): Marking (red) and other tracking dots (blue)

Both variants always contain exactly 18 dots. Two dots in the first column mark the beginning, set in row 0 and 4. Column 7 consists of 2 dots as well. However, those dots do not have a fixed position compared to the marking dots, but they are always at the same distance from each other, namely 5 empty cells. The functionality of this column is currently not clearly apparent. One functionality could be an orientation marker, as the pattern is sometimes printed flipped horizontally, which can be detected since column 7 thus appear in column 9. All other columns of the pattern (1-6 and 8-15) consist only of one dot as “one hot encoding”.

Redundancy Check The source alphabet contains 16 elements which are being encoded with a (16,16,2) “one hot encoding” respectively 32 elements encoded with (32,32,2) for 5s. The pattern consists of 14 columns where each column stores one code word. A code word weights “1” and therefore produces exactly one tracking dot. In addition there is column 7 with weight 2 storing exactly two dots, but not with more information as the distance between both dots is always identical.

Decoding The pattern observed on the printouts does not show any difference when printed timely apart. Thus, it does not appear to contain the time of the printout. The first part of the pattern (columns 1-6), including the special column 7, correlates with the specific printer model within our data set. However, we had too few samples to verify this clearly. The second part of the pattern after column 7 is similar, if TDMs with similar serial numbers are compared. However, it differentiated too much to contain the information in a simple encoding such as in Pattern 1 or 4. Simultaneously, they are too similar to contain a hash-like encoding. Currently our decoding attempts have not been successful so far.

Overall, serial numbers of Canon printers can have formats of different length (8-9 signs). Since the columns of the pattern are always filled with 18 dots, the encoding of the serial numbers of different length would have to map to an encoding of the same length.

Conspicuousness Considering the column 0 and 7, there are $14 + 4 = 18$ tracking dots in total (0.3 dots per bit and 0.24 dots per bit for 5s). Assuming all columns but column 0 contain information, four bits could be encoded per column, given the 16 different positions a dot could have in the first variant. This would equal 60 bits of information for the whole pattern in the first variant and 75 bits in the second variant 5s.

3.3.2.6 Pattern 6

In further investigations we found a new pattern within Lexmark printers in our data set (Fig. 3.20). Those patterns have a matrix of $8 \times 21 = 168$ cells.

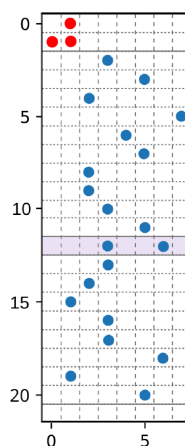


Figure 3.20: Pattern 6: Marking (red) and other tracking dots (blue)

Three dots in the first two rows mark the beginning of the pattern, similar to Pattern 2. The pattern is repeated evenly both vertically and horizontally without offset in between. Information bits are found as (8,8,2) “one hot encoding” within row 2-11 and 13-20 (18 rows). Row 12 might be a separator or

orientation marker. It contains two dots with same distance and on the same position within our samples. Considering the marker and row 12, there are $18 + 3 + 2 = 23$ tracking dots in total (0.39 dots per bit). Taking row 12 as marker, the pattern can differentiate between 8^{18} states in total. This is equivalent to storing $\ln(8^{18})/\ln(2) \approx 54$ bits. Involving row 12 as information this could increase to 57 bits. Overall, Pattern 6 is currently not yet deeply investigated, especially because we only have a total of four samples of this pattern in our data set.

3.3.2.7 Summary and Comparison of the Patterns

Table 3.9: Pattern Comparison

Pattern	Δ_i	Δ_j	Size		Capacity	Density		Dots/in ² avg / max
			Cells	in ²		Bits/cell	Bits/in ²	
1	0.02 in	0.02 in	512	0.21	48 bits	0.09	234.38	166 / 244
2	0.03 in	0.03 in	414	0.37	64 bits	0.15	171.77	137 / 180
3	0.02 in	0.02 in	384	0.15	69 bits	0.18	454.43	195 / 195
4	0.04 in	0.04 in	256	0.41	98 bits	0.38	239.02	112 / 186
5	0.02 in	0.01 in	256	0.05	60 bits	0.23	1200	360 / 360
5s	0.02 in	0.01 in	512	0.10	75 bits	0.15	750	180 / 180
6	0.04 in	0.03 in	168	0.18	54 bits	0.32	300	127 / 127

Table 3.9 gives an overview over the detected patterns. It notes the amount of cells of one unique matrix including the spacing to its closest repetition. The capacity for storing information bits is given as well as the amount of bits that one table cell and one square inch can store. The density is the quotient of the capacity and the size. The amount in bits per cell shows the efficiency of the patterns regardless of the cell distance whereas the number in bits per square inch does consider the cell distance. The more dots per square inch are printed the more visually conspicuous the matrix is on the paper. The amount of dots per matrix depends on the encoded information. Its minimum, average and maximum are given in dots/in², divided by the pattern's size. The (n, l, d_{min}) code description (Tab. 3.10) follows with the amounts f_e and f_k of faulty bits that can be detected and/or restored correctly (forward error correction).

Table 3.10: Code parameters

Pattern	Dots per word avg / max	(n, l, d_{min})	f_e	f_k	f_e
1	34 / 50	(7,6,2)	1	0	14%
2	51 / 67	(20,16,4)	3	1	15%/100% ^{1),2)}
3	30 / 30	(6,6,2)	6 ¹⁾	0	100% ¹⁾
4	46 / 76	(120,98,4)	3	1	14% ²⁾
5	18 / 18	(16,16,2)	16 ¹⁾	0	100% ¹⁾
5s	18 / 18	(32,32,2)	32 ¹⁾	0	100% ¹⁾
6	23 / 23	(8,8,2)	8 ¹⁾	0	100% ¹⁾

1) An error of two bits per code word might not be detected in any case. Presence of parity bits has not been revealed.

2) For the inner code of the code chain

Comparing the patterns leads to the following observations: Pattern 1 encodes information per in² very densely because it uses the binary system and produces few redundancy so it only detects single faulty bits and cannot correct errors at all. Pattern 4 is similar to pattern 1 but adds error correction and therefore displays information less densely. The density per cell of Pattern 4 is higher than of Pattern 1, because Pattern 1 leaves nearly every second column and row empty, assuming a cell distance of $\Delta_i = 0.02$ in.

Pattern 2 uses the “one hot encoding” and a parity code. The “one hot encoding” with length 4 still can display information quite densely and can detect distributed erroneous bits quite well. On the other hand it only detects at maximum 4 faulty bits whereas a “one hot encoding” with a higher length has a lower information density but can detect errors of a higher weight. The parity bits lead to a high redundancy. However, Pattern 2 minimizes its dots because it sets the parity to an odd amount of ones. Blocks where all four inner code words are different occur a lot more often than a block where all code words are the same. Therefore the outer parity code deals with odd amounts of “1” more often and sets “0” as the parity bit in this case.

Pattern 3 uses a “one hot encoding” with length 6. The presence of parity bits has not been found so we assume that all bits are information bits. A “one hot encoding” with length 6 stores little information per sign but the absence of additional redundancy helps this pattern to create a higher density than Pattern 2. The “one hot encoding” allows six faulty bits to be detected. Pattern 5 and 5s represent a similar structure with a ‘one hot encoding’ of length 16 or 32 in each column, and without additional parity bits.

Pattern 4 can be the least conspicuous one because it produces the lowest amount of dots per in² in the best case. Though the worst case is likely to occur if the pattern is being used to store a big variety of information.

Pattern 2 and 5s produces at maximum 180 yellow dots per in² which is the lowest maximum for all patterns. Pattern 3 produces just 15 dots more but stores 2.6 times of the information of Pattern 2 per in². Pattern 4 does not use an explicit marker. It can be aligned by the definition of the free space of its offset pattern but this is a lot more computationally expensive than finding the three marking dots of Pattern 2. Code words of Pattern 1 can be created by the random distortion on the paper. This makes it difficult to find unambiguous information and also to determine whether a sheet does or does not contain this pattern. The most efficient patterns are 3 and 5s. Storing between 454-750 bits per in², they represent the highest density. Because error correction can be achieved using the repetitions anyway, it is reasonable to focus on a high error detection capability rather than on forward error correction. This pattern have the highest error detection capability. A small cell distance of 0.02 in is useful to allow many cells per area. This is possible due to the few dots produced by the “one hot encoding”. In the worst case 180-195 tracking dots are being produced which is similar to the other patterns. In addition, Pattern 3 and 5s have enough capacity to store the same information as in all other patterns without time and date information.

Concerning the information content only pattern 4 includes date and time information regarding the print. All other patterns are constant for each printer and do not vary by each print. Hence we conclude that they do not contain dynamical information like date or time.

3.3.2.8 Analysis of the Improved Extraction Approach

Table 3.11 shows the amount of printers where the prints’ tracking information passed the redundancy check successfully at least out of one sheet of its documents. It shows the amount of printers belonging to a pattern as well as the overall success to extract a valid tracking dot pattern out of a printout within our data set. Overall, valid TDPs were successfully extracted from 960 out of 1505 printouts, and 129 out of 193 printers.

From Pattern 1 the extraction of a TDM often is ambiguous. A randomly distorted matrix could be valid according to the redundancy check by Pattern 1. Especially the markers from Pattern 2 are valid markers for Pattern 1 and can therefore be interpreted as a Pattern 1 matrix more easily. The spots around valid

Table 3.11: Successful extracted patterns with accepted redundancy check for known patterns within our data set

Pattern	Print Outs	Printer
1	133	9
2	462	59
3	170	28
4	135	19
5	47	6
5s	13	8
Out of	960/1515	129/194

dots have to be checked carefully before deciding on Pattern 1 for a sheet. The prints of which the tracking information could not be decoded showed sparse matrices or were scanned unluckily. Especially Pattern 3 prints were hardly distorted on the scan.

Furthermore, we evaluated the extraction of the printers' manufacturer and serial number from a valid TDM. 100% of the extracted serial numbers are part of the printer's actual serial number. The manufacturers for all but one printer could be decoded correctly. The erroneous printer¹² has possibly been labelled wrong in the data set. Overall, many of the scans across the data sets were collected by individuals, thus the quality of the data can differ strongly. Some scans were presumably labeled wrongly, and some scanner and their settings produced unusable data. Therefore, we have taken first steps towards a unified and cleaned data set (see Appendix I-E).

In comparison to previous work by van Beusekom et al. [455], our method maps the tracking dots into a grid and therefore transforms them into a matrix. For each two prints, van Beusekom et al. aimed at deciding whether they come from the same printer or not. To achieve this, they did not extract the TDM as a matrix but as an image, using the TDM's repetitions and a pixel threshold to separate true dots from distortion. This may allow false positives: The prints from two different printers might be detected as from the same origin if significant dots were removed. In contrast, our method uses the code's redundancy check. Comparing only valid TDMs, a classification of two different printers as identical is very unlikely. Moreover, it enables the digital visualization and decoding of known pattern content of extracted TDMs.

3.3.3 An Anonymisation Approach against Surveillance through Tracking Dots

Tracking dots on a sheet reveal information about the printer, which is often owned by the originator, and are therefore a threat to privacy. Tracking dots information have no controlled access and can be read and decoded by anyone. Moreover, today's printers are often connected to the internet for e.g. automatic updates, or simplification of cartridge reordering. In these connections, the serial number could be part of the information exchange and thus a link to the owner can be easily established¹³. This would be precarious e.g. in case the sheet is a critical leaflet about the government in a dictatorship. For this reason we introduce methods for removing tracking dots from scans and for masking tracking dots on printouts. Each anonymization method was successfully tested.

3.3.3.1 Removing Tracking Dots on Scans

When scanned documents are being sent via the internet, they might still contain tracking information. Tracking dots may have a strong effect: In 2017, scanned NSA documents has been leaked and published

¹²Sample 99 in the DFKI data set

¹³see e.g. <https://www.hp.com/de-de/privacy/privacy.html>

in the online medium The Intercept¹⁴. The most probable reason for having identified the Whistleblower Reality Winner was embedded tracking dots¹⁵. Tracking dots can mostly be removed from scans (Fig. 3.21) by clearing the original document's empty areas as detected in Section 3.3.1.4.

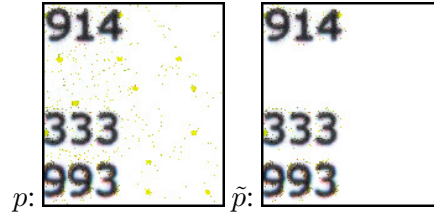


Figure 3.21: A scan before (p) and after (\tilde{p}) automatic tracking dot removal. Yellow colours are darkened equally on p and \tilde{p} for better visibility

3.3.3.2 Masking Tracking Dots on Prints

Our idea for anonymization of printed documents was to add a custom TDM as a mask on top of the printer's TDM to prevent restoring a word b correctly. The ambiguity of correcting b must be high enough to allow many possibilities on restoring a masked TDM: Some code words can be detected as wrong by their content, e.g. if a decoded word contains a month greater than 12 or an invalid serial number. Embar et al. [129] have mentioned to print a full unit matrix on the sheet to prevent an unambiguous decoding of a Pattern 4 matrix. For pattern 2, 3, 5, and 6, filling all blocks completely with dots is a safe option to anonymize the TDM (see also Fig. 3.24 right TDM), as the information in the code words is not completely known and might use a further error correcting code. However, printing a yellow dot in all cells makes the TDM very conspicuous and uses a lot more toner. This might not be necessary for pattern 1 and 4. Because we know the codes, for a given TDM we want to find a mask that has as few dots as possible but makes the decoding ambiguous when it is being united with the original TDM. The mask has the same size as the printer's TDM and must cover all of the TDM's repetitions on the sheet constantly. For each pattern there is a different algorithm to create a mask.

Pattern 1 The mask for Pattern 1 (Fig. 3.22) is being created as follows. Let $s = 3$ if the information dots are placed in the odd columns and $s = 2$ otherwise. On each of the rows 1, 3, 5, 7, 9, 11, 13 and 15, one empty cell has to be chosen at random where the cell number must be one of $s, s + 2, s + 4, s + 6, s + 8, s + 10, s + 12$. These cells must carry a dot on the mask. The parity will be broken and the adversary does not know which dot has been added. From rows containing exactly one dot, it is unambiguous which dot we added. If we added two dots to each row that was empty on the original TDM, the parity would reveal that we have added an even amount of dots which must be smaller or equal than two. Therefore empty rows must be added three dots. Exactly the same mask starting at column 0, row 0 has to be repeated from column 16, row 16. This adds at least eight dots to each TDM.

If our algorithm is known to the adversary and he wants to restore the masked TDM, there are $\prod_{r=1,3,\dots,15} d(r) \geq 3^8 \approx 2^{12.7}$ possible code words where $d(r)$ is the amount of dots in row r with $d(r) \geq 3$. If the parity bit in row r is not being set, the adversary will remove one of each of the dots. On rows where the parity bit

¹⁴<https://theintercept.com/2017/06/05/top-secret-nsa-report-details-russian-hacking-effort-days-before-2016-election/>

¹⁵<https://qz.com/1002927/computer-printers-have-been-quietly-embedding-tracking-codes-in-documents-for-decades>

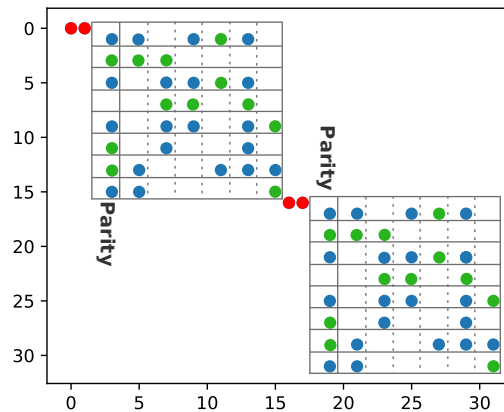


Figure 3.22: Pattern 1 mask example (green)

is being set, itself might as well be considered as the flipped bit from our mask. Using this interpretation, an additional possible code word can be concluded from the unchanged information bits. The ambiguity can be low if each of the the original TDM's rows contain very few dots. To increase it, to each row a dot can be added at a randomly chosen cell. The parity in these rows will be correct.

Pattern 4 The mask (Fig. 3.23) for columns 0-7, rows 1-5 shall carry a dot on a randomly chosen empty cell for each row and fill the outer parity row 15 completely with dots between columns 0 and 7.

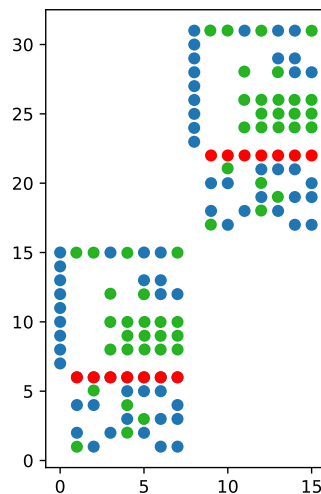


Figure 3.23: Pattern 4 mask example (green).

To obscure the manufacturer, row 12 has to be modified. All known manufacturers use one of the numbers 3, 4 or 20 resp. binary numbers 11, 100 and 10100. From their disjunction 10111, none of the original number can be concluded. So it is sufficient to fill columns 3, 5, 6 and 7 of row 12. To hide the date of the printing process, rows 8, 9 and 10 have to be treated in a specific way: Filling columns 3-7 in row 8 hides the domain of the year, filling columns 4-7 in row 9 hides the domain of the month and filling columns 3-7

in row 10 hides the domain of the day. The hour and minute without the date are irrelevant and not being considered any further here. If rows 3, 4 and 5 are filled so that each row contains at least s dots, then at least s^3 different serial numbers can be concluded from the TDM. A copy of the mask shall be placed on columns 8-15, rows 17-30. Overall, this adds between 6 and $8+5+4+5+2+12=36$ dots to each TDM.

Mask Calibration In practice, the prepared mask needs to be printed in such a way that it hits the printer's native tracking dots although their position on the sheet is unknown. For this, we propose to use a calibration sheet in combination with our extraction algorithm to calculate the mask orientation for a specific printer. This calibration sheet contains markers in each edge on the Cartesian points A, B, C, D . After scanning it can be aligned so that its coordinate system matches the original calibration image's coordinates. An additional marker O is located in one edge of the page and used to maintain the page orientation. A contour detection algorithm [430] is used to find the markers in the scan, which are printed in the toner's colours cyan and magenta.

First, the scan needs to be rotated so that point O is in the same edge as in the original image. Then all other points A', B', C', D' need to be matched against the markers at A, B, C, D in the original image by filtering markers according to coordinate ranges and focusing on the smallest x and y coordinate for each marker. After applying a perspective transform mapping A' to A, B' to B etc., any point $Z \in \{A, \dots, D\}$ on the scan would represent Z' on the test image so that $Z = Z'$. However, the process of scanning a document causes geometrical distortions. These distortions are minimal near the alignment markers A, B, C, D . Let's assume the printer's TDP has the parameters $n_i, n_j, \Delta_i, \Delta_j$. We need to find the offset coordinates x_o, y_o near the top left corner where the first TDM begins. If four valid TDMs have been found at points $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ where $\hat{A} = (x_{\hat{A}}, y_{\hat{A}})$ is close to $A, \hat{B} = (x_{\hat{B}}, y_{\hat{B}})$ is close to B , etc. then the best estimation of x_o is the average of the TDM's offsets at $\hat{A}, \hat{B}, \hat{C}, \hat{D}$. We use modulo to calculate the offset given the coordinate. If the pattern's offset is close to 0, a constant $c \in \mathbb{R}$ must be chosen so that all x values can be mapped into a range $(x+c) \bmod (n_i \cdot \Delta_i)$ that allows us to calculate a meaningful average:

$$x_o = \text{average}((x_{\hat{A}} + c) \bmod (n_i \cdot \Delta_i), (x_{\hat{B}} + c) \bmod (n_i \cdot \Delta_i), (x_{\hat{C}} + c) \bmod (n_i \cdot \Delta_i), (x_{\hat{D}} + c) \bmod (n_i \cdot \Delta_i)).$$

y_o can be calculated analogously.

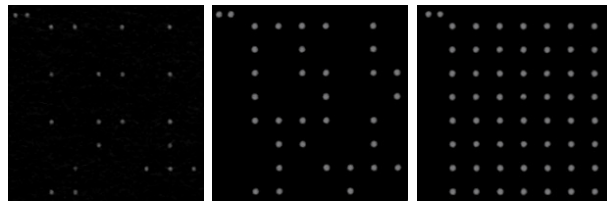


Figure 3.24: Pattern 1 practical mask example. From left to right: original TDM, masked TDM, fully dotted TDM.

With this extracted information the mask can be generated. If a page shall be printed and anonymized, the calculated mask shall be transformed into an image having cell distances of Δ_i, Δ_j . It must be printed at $(x_o|y_o)$ and at all points where a repetition of the TDM begins: $\{(x_o + x \cdot \Delta_i \cdot n_i | y_o + y \cdot \Delta_j \cdot n_j) \mid x, y \in \mathbb{Z}\}$. This anonymization workflow was successfully tested with a Ricoh MP C305 (Pattern 1; see Fig. 3.24). In practice we do not print only the mask but overprint the entire matrix (original and mask) with slightly larger dots, as there was always a tiny offset between both which would allow the conclusion of the original pattern.

3.3.4 Towards a Declaration TET - The DEDA Toolkit

To simplify end-user access to hidden, unknown, embedded information in color laser printouts, we incorporated our findings on yellow dots into a Declaration TET. The categorization parameters of this TET according to Zimmermann [507], could be found in Table 3.12. Thereby, the whole workflow of *Dot Extraction*, *Decoding* and *Anonymization* could be obtained with our *Deda* toolkit¹⁶.

Parameter	Value
Target Audience	data-subject
Data Types Presented	implicit / observed
Execution Environment	client-side
Application Time	ex-ante / ex-post
Interactivity Level	interactive
Delivery Mode	pull

Table 3.12: Classifying the envisioned TET according to parameters described by Zimmerman [507]

The enabled transparency could be used on one hand to raise privacy literacy and knowledge regarding which additional information is disclosed when sharing analog documents. On the other hand it could be used to reuse embedded tracking dots for forensic purposes, e.g. for the verification of analog documents. In addition, the tool enables intervention regarding data disclosure through tracking dots before sharing them. Both for digitized analog documents and the print itself. The user interface of the TET was implemented with eel¹⁷. In the following we describe the structure and functionalities of Deda.

3.3.4.1 Transparency - Analysis and Verification of Printouts

This part of the TET enables transparent insights regarding embedded yellow dots within analog printouts. The first functionality Deda provides is the **extraction and decoding** of embedded tracking dots. For this, the printouts to be analyzed have to be color scanned with at least 300 dpi, preferably saved with lossless compression. Inputs could be single files up to folders with a bunch of scans in common image formats (png, tiff, jpg, bmp). By using the improved extraction algorithm, each input image is analyzed. If a pattern is detected, the found pattern is displayed, as well as the corresponding information which are decodable and inferable. This includes for example information about the manufacturer, the model, serial number, or the printing date (see e.g. Fig. 3.25 (a)). Results can be exported as an image or PDF.

In addition, Deda simplifies the **comparison of documents** regarding their source devices, based on embedded yellow dots by automation. The input is equally to the analysis part, and the following output is a comparison of the documents based on embedded yellow dots and listing of the different patterns and affiliations of the printouts (see Fig. 3.25 (b)).

As an additional feature, the third part enables the **generation of an own tracking dot pattern**. This could be used, e.g., as a hidden signature for documents which should be protected. The input for this is a PDF document in which the pattern is to be embedded, as well as the information for the pattern. Currently, only Pattern 4 is provided. Here, the date, time, a number with six digits and manufacturer information could be embedded. Additionally the dot size can be chosen by the user. Using a printer without tracking dots, the document can be printed with the embedded, self-created pattern.

An interesting request of a user revealed, that this is also possible with the copy mechanism of a color laser printer. If there is an analog document without a signature, this document could be signed with

¹⁶<https://github.com/dfd-tud/deda> or <https://dfd.inf.tu-dresden.de>

¹⁷<https://github.com/ChrisKnott/Eel>



Figure 3.25: Analysis Part of the Deda Toolkit

yellow dots by placing it in the feed tray of the color photocopier device, and make a color copy without putting anything on the scanner. Afterwards yellow dots of the copier device are embedded within the specific document. This procedure was successfully tested with an Ricoh Aficio MP C307 device.

3.3.4.2 Intervention - Destroying Yellow Dot Patterns

The interactive part of the Deda TET enables the obfuscation of embedded yellow dots for anonymous sharing of analog documents.

The first feature enables the automated **deletion of tracking dots within scanned analog documents** (see Sec. 3.3.3.1). Currently only image files are supported. Afterwards, the user has the possibility to share digitized analog documents without embedded source device information. Note, that our method for removing yellow dots on scans is rather simple. Of course the recovery of tracking dot information, if deleted all from empty areas, becomes very difficult. However, depending on the content of the document it could be possible to recover the TDM out of printed areas. Especially when large areas are printed in the cyan colour range. The user should therefore take a closer look at the result to rule out a TDM recovery. The second **anonymization** feature is **for print outs** and consists of two steps. The first step is to generate a anonymization mask for the specific printer with which anonymized prints should be enabled. Therefore, the user has to print and scan a specific calibration sheet with the printer to be anonymized, to find the printer's TDM as well as its native positions (see Fig. 3.26 (a) for the DEDA UI and Sec. 3.3.3.2 for the Calibration functionality). With this extracted information the mask can be generated.

In the second step the user has to merge the mask with the document to be printed anonymized (Fig 3.26 (b)). Afterwards the masked document could be printed with ambiguous tracking dots. In our data set the position of the tracking dots was statically always at the same place. Thus, it can be reused for all following documents. Since the recognition does not always hit the exact position of the original dots, the user also has the option to vary the positioning ($(x_o|y_o)$ shift), as well as the size of the dots.

Overall, the page margin must be identical for printing the calibration sheet and the anonymized document. If the margin has not been zero (borderless print), there might remain unmasked tracking dots on it. If this

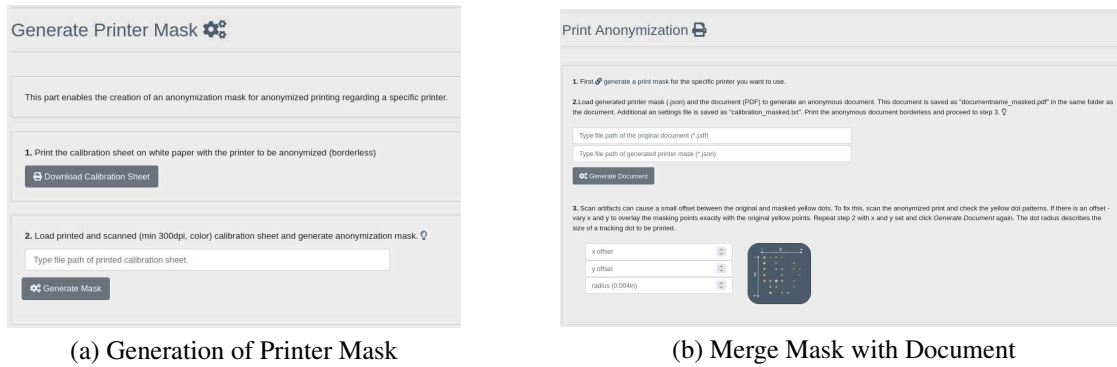


Figure 3.26: Anonymization of Printouts within Deda

is not possible, the margins have to be cut off, otherwise TDMs might be reconstructed.

3.4 Concluding Remarks & Summary regarding Analog Information Exchange

Due to digital advances, nowadays anyone can easily create, duplicate, and modify analog documents in high quality. Despite the positive impacts, printed documents are also used for illegitimate purposes. This requires methods which allow to verify the credibility of analog documents or to reveal their originator. An important task here is to identify the source device of a printed document. Thereby, digitization not only simplifies the creation and forgery of analog documents, but also their forensic examination. Digital printer forensic approaches improve traditional forensics, by determining the source printer automatically using commercially available recording devices such as scanners or cameras. This makes forensic examination cheap, simple, and therefore available to everyone. However, such digital printer forensic methods not only allow the investigation of crimes, but potentially also the arbitrary tracking of authors of analog documents. In this chapter we analyzed the resulting discrepancy between privacy and security regarding printed documents.

In the first part we investigated *passive printer forensics*, which attempt to extract specific artifacts from printed documents. These artifacts are used as identification features, called *intrinsic signatures*, of specific printer technology, brand, or model. With an exemplary analysis we showed that a lot of influencing parameters could change these signatures in the variable printer space. Currently the robustness of intrinsic printer signatures against possible influences, as well as their sensitivity, is insufficiently investigated (see Sec. 3.2.3). Overall, digital passive methods are a useful complement to traditional forensics and simplify the examination of documents, but remain limited to use cases with prior knowledge. However, due to the limited informative value of intrinsic signatures, which extend at most to the printer model, and the potential susceptibility to varying print parameters, we argue that tracing unknown analog documents back to the specific source device and thus to the originator using passive printer forensics is only feasible at very high cost, if at all. Therefore, they do not leave a significant gap between privacy and security. In contrast, *active printer forensic* methods focus on *extrinsic signatures* which are explicitly embedded within a printout. Such an extrinsic signature are tracking dots which are currently integrated into nearly all color laser printers. Thereby, additional tiny hidden yellow dots are printed over the entire printout, which encode information about the printing process, like the serial number of the printer. Since this information

is integrated into every color laser printout, each of them can be clearly traced back to its source device and therefore potentially also to its creator. On the one hand, an effective tool for forensic investigations and solving crimes, on the other hand, a strong intrusion into privacy. This is especially critical with regard to whistle blowers, journalistic work, or the possibility to express criticism anonymously; all important components of democratic societies. However, the properties, encoding and information content of these tracking dots were mainly unclear.

Thus, in the second part of this chapter we investigated tracking dots. In total, we discovered six distinct patterns and succeeded in automatically extracting and interpreting the structure of the patterns as well as decoding Pattern 1, 4 and partially Pattern 2 and 5. Based on our investigations we developed a transparency tool called DEDA. Transparency on one hand for verification of information (re-usability of tracking dots), and on the other hand to visualize which information a printer reveals in terms of privacy. In addition, we developed an anonymization approach to defeat arbitrary tracking and integrate it within the DEDA toolkit. Finally, we found a new type of tracking dots, which we named inverse tracking dots and which we report in Appendix I-E with additional open challenges regarding the analysis of tracking dots.

After publishing our results, we briefly attracted media attention¹⁸. Here, privacy was in particular the focus and presumably also the conspiracy character. However, after a few days this has subsided and the subsequent attention focused on the potential of verifying information. Law Enforcement Agencies as well private people and companies were interested in information and analysis of embedded yellow dots regarding verification. The Feedback also indicated that German law enforcement authorities most likely do not have official access to the tracking dot content. Probably, not even the intelligence service, which apparently showed interest in the anonymization.

The discrepancy between privacy and security in relation to tracking dots is also reflected in a user study we conducted in 2022 (see Appendix I-D). While comparing advantages and disadvantages of tracking dots, about a quarter of the participants perceive privacy impacts to be less important than benefits (verification), almost half weigh privacy constraints more highly, and the remaining perceive pros and cons to be balanced. However, nearly all participants would like to see more transparency. An interesting finding was that, despite the fact that tracking dots have been known since 2004 and have been mentioned several times in the media, three quarters of the participants were completely unaware of their existence.

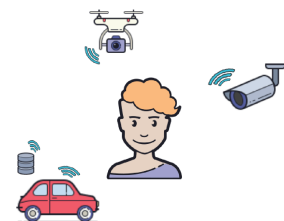
¹⁸<https://www.deutschlandfunk.de/farblaserdrucker-tracking-dots-unlesbar-machen-100.html>, <https://heise.de/-4090876>, https://www.theregister.com/2018/06/27/german_researchers_defeat_printer_tracking_dots/, ...

4

Information Disclosure while using or being surrounded by Digital Devices

Next to analog documents, nowadays information is mainly exchanged via digital devices. Wireless communication and the establishment of mobile devices enable digital information to be received and sent anywhere and at any time. Further, visions of ubiquitous computing and the Internet of Things (IoT) [476] have become an integral part of everyday life. Tiny, unobtrusive, and networked sensors that measure and monitor their environment are continuously integrated into all types of physical objects [481]. Merged with artificial intelligence [141], these objects are enhanced by more elaborated abilities of perception and agency. Overall, this enables a host of new possibilities, improvements, and efficiency gains. For example, in energy supply (smart meter), production (smart factory), medical care (mHealth), city administration (smart city), or home automation (smart home).

Despite the wealth of benefits, ubiquitous use in both public and private spaces brings challenges and implications, especially with regard to privacy, i.e. the unconscious disclosure of private information. The tiny sizes of sensors, their seamless integration into the physical world, and the more natural tactile or audiovisual interaction, are increasingly blurring the boundary between the digital and analog worlds. Additionally, these smart devices usually provide only the interface to the environment, with the sensor data being aggregated and processed on more powerful, external servers. Overall, the invisibility of sensors simultaneously makes data acquisition, transmission, processing, and storage invisible to the user. Thus, in addition to the more 'conscious' use of digital information exchange, like via messenger, e-mails, or websites, information is now also exchanged digitally simply when using (smart) physical objects, at the same time revealing sensitive data about their users. This range from location data through mobile device sensors, health, or motion data through wearable sensors like fitness trackers, presence and energy use through smart meters, to biometric data through the use of audiovisual smart devices, e.g. smart speakers.



One vivid example of this development is the *digitization of mobility*. Future smart cars will be able to communicate with other vehicles and the surrounding transport infrastructure, thus enabling connected driving. This will be realized with integrated sensors which can send and receive continuous wireless broadcast signals. This communication will increase road safety, reduce CO₂ emissions, and improve traffic efficiency. However, to avoid errors in the communication and to protect the integrity of the traffic data, messages have to be authenticated. Overall, this enables movement profiling and thus limiting location-based privacy. For this reason, approaches have been developed that attempt to minimize the possibility of location tracking. However, the trade-off between privacy, security, and utility becomes very clear in the various solutions. As the vehicle communication takes place unconsciously in the background, in the *first part* of this chapter, we take a closer look at this trade-off and analyze to what extent the European ambitions will protect location privacy while connected driving through the city.

In addition to wireless communication, smart cars are also equipped with a range of additional sensors. By integrating radar and camera sensors, vehicles are empowered to understand their environment and react to it automatically. One primary concern with such sensors is, that not only the owner of these smart devices but everyone who enters their recording radius is affected¹. The people who are recorded, however, have no way to recognize this data collection and its processing. Consequently, there is a *foreign control* over the personal data of such *bystanders*. Therefore, in the *second part* of this chapter, we want to highlight the issue of bystanders in the world of IoT, especially with a focus on audiovisual recording devices. We introduce our first concept and prototypical implementation of a transparency solution for bystanders, which inform users about nearby recording devices. Finally, we analyze how the bystander issue is perceived by participants from the German culture and whether our transparency approach is seen as a sensible solution.

4.1 Privacy-Utility Trade-Off during Connected Driving

The ongoing digitization of mobility enables a range of new opportunities and efficiency benefits. For example, it simplifies the sharing of vehicles such as cars or bicycles. However, when using established sharing systems, sensitive location data such as start and destination points and times, or even the entire route traveled, are disclosed [464]. Thereby, regular use enables the creation of motion profiles (trajectories), and thus determines the user's points of interest (PoIs) (e.g. residence, workplace, or school). Out of this, further information could be predicted, e.g., whether the user has children, any illness, or religious views. Even with the removal of directly identifiable data, it is possible through locality behavior, again assign individual persons or user groups² [203]. Moreover, such location data is often shared with third parties, e.g. advertisers or city planners [464].

With the increasing digitization of the entire transport infrastructure, such location information is not only disclosed by shared vehicles, but by all vehicles. Cooperative Intelligent Transportation Systems (C-ITS) will enable wireless communication between vehicles (V2V) as well as the surrounding transport infrastructure (V2I), summarized as vehicle to everything communication (V2X). This is achieved by

¹<https://github.com/tevor-threat/scout>, <https://www.wired.com/story/tesla-surveillance-detection-scout>

²<https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy>

vehicles using their on-board units (OBU) to regularly send data such as position, speed or direction by broadcasting to all receivers in the vicinity and to receive broadcast messages from other surrounding vehicles or the infrastructure via road side units (RSUs). This enables so-called vehicular ad hoc networks (VANETs), whose participants are always informed about vehicles in the vicinity. Shared information can be used, e.g. to avoid collisions by coordinated vehicle movements, to warn about accidents and traffic jams, or for intelligent traffic light control, thus ensuring safer and more efficient road traffic overall.

To achieve this goal it is important that vehicles can trust the data they receive, i.e. to *verify* received information. In order to ensure that the information is not faulty or manipulated (whether un- or intentionally), common systems rely on the authentication of sent messages by means of digital signatures and certificates. Each recipient can determine the authorization of the sender using these signatures and verify the integrity of the message.

However, this approach poses major problems regarding location-based privacy. Every vehicle regularly exposes its identity and position to its environment. Hence, the introduction of V2X communication can facilitate the tracking of vehicles and thus the creation of motion profiles, not only by service providers but from everyone with sufficient antennas. While the data collection, and processing through the use of, e.g. shared vehicles, is still explained to the user via privacy policies, V2X communication takes place completely in the background.

One common solution is to pseudonymize the communication. Instead of using one signature, that unambiguously identifies a specific vehicle, a vehicle would use a number of different signatures which are changed while driving according to a specific change strategy. Overall it is important to use an effective and robust pseudonym changing scheme as the effectiveness of the protection of vehicles data privacy relies strongly on it. A major issue with pseudonym changing schemes is that pseudonyms might be linked and thus a trajectory can be reconstructed or vehicles even can be deanonymized.

Therefore, in this section we analyse a pseudonym change strategy, which is recommended in the security guidelines of the European C-ITS platform and has good chances to be included in a future European standard [143]. By simulating a realistic traffic scenario and attacking the applied pseudonym scheme, we try to determine to what extent it effectively prevents vehicle tracking. Next to the experimental analysis, in Appendix II-A we describe theoretical considerations regarding this scheme and in Appendix II-B we investigate additional wireless identifier which could potentially be used to bridge the pseudonym change.

4.1.1 Pseudonymization of V2X Communication - Related Work

The topic of pseudonymization of vehicles in VANETs is being addressed in a large amount of work. A detailed overview is given by Petit et al. [357]. They give a broad insight into the functioning of pseudonyms as well as various implementations in concrete pseudonym schemes.

In order to ensure that pseudonyms protect privacy effectively and at the same time do not hinder V2X applications, a number of requirements must be met. A pseudonym has to be unique, available, only valid for a limited period of time, and it must be impossible for an attacker to link multiple pseudonyms [357]. Overall, on one hand, pseudonyms should not contain any information that could be used by an attacker to deduce the actual identity of the vehicle. On the other hand, authorities, such as the law enforcement, must be able to resolve pseudonyms, e.g. due to malicious behavior of a vehicle.

Petit et al. [357] illustrates the functional principle of pseudonyms in V2X communication using an abstract life cycle. Thereby, they identified five phases: pseudonym issuance, usage, change, resolution,

and revocation (Fig. 4.1).

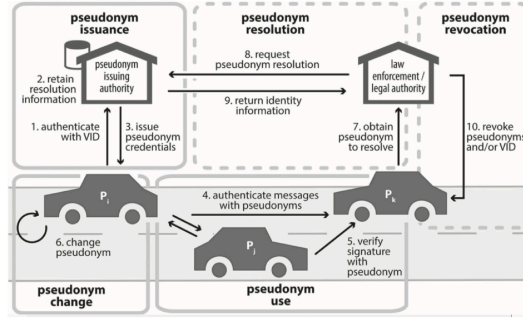


Figure 4.1: Schematic representation of the five pseudonym life cycle phases [357].

In this chapter, we focus on the pseudonym change part. Therefore, in the following we describe related work regarding pseudonym change strategies and evaluations regarding their unlinkability characteristics.

4.1.1.1 Pseudonym Change Strategies

For pseudonyms based on asymmetric cryptography the changing strategy is crucial. Unsuitable strategies can allow an attacker to link different pseudonyms. Many different strategies have been proposed in recent years [139, 357].

The easiest way would be to use a new pseudonym for each new message. This offers a high level of anonymity protection, as the risk of pseudonym concatenation is minimal [445]. However, it also leads to an extremely high pseudonym consumption, which in turn requires that each vehicle has a large number of pseudonyms in its list. The decisive disadvantage, however, is that safety mechanisms and the benefit of connected communication are made impossible, since they require at least a temporary concatenation of messages. For example, it would hardly be possible for a vehicle to reliably determine how many other vehicles are in its vicinity. This strategy is almost unanimously regarded as unsuitable and rejected [357]. In order to enable V2X functionality, pseudonyms must be used for a certain number of consecutive messages. A strategy for this would be a change based on *fixed parameters*. For example, the change can be triggered after a certain time, number of sent messages or distance [139, 357]. However, fixed parameters offer only limited protection against attackers. As soon as the attacker knows the parameters, vehicle tracking becomes trivial, as he can reliably predict future changes [357]. In order to prevent predictability, a *random* value can be added to the fixed parameter that determines the change. Another possibility is that a vehicle dynamically initiates a pseudonym change based on its *mobility parameters*, e.g. as soon as it changes its direction of travel or speed. To avoid ineffective pseudonym changes, e.g. only one vehicle changes its pseudonym or only one car is on the road, other strategies use environmental parameters. For example *traffic density* or *cooperative synchronous* changes with surrounding vehicles. Since attackers could also use the content of sent messages to predict the movement of a vehicle, and thus to link pseudonyms, there is the idea of a short period of *radio silence*, i.e. the vehicle does not send messages, after changing the pseudonym. The longer the radio silence, the more difficult the linkage of pseudonyms. However, the radio silence of the vehicle undermines the functionality of V2X-based safety applications, e.g. for collision avoidance. Because of these aspects, this strategy is considered by some to be unfeasible [357].

Overall, the different strategies show clearly the discrepancy between utility and privacy for this area.

4.1.1.2 Evaluating Unlinkability of Pseudonymization

Different literature analyzed pseudonym changing strategies such as Boualouache et al. [51]. They give a comprehensive overview of pseudonym change strategies and classify them according to various characteristics. Wiedersheim et al. [482] showed in their work that simple pseudonym changes do not provide satisfactory pseudonymization, since it is easy for an attacker to link pseudonyms which are changed in the observation areas.

An important basis for the evaluation of pseudonym changing strategies are traffic simulations to imitate the complexity of the issues and achieve results that are close to reality. Various studies have dealt in detail with the simulation of traffic scenarios to investigate certain pseudonym change strategies.

Troncoso et al. [445] carried out a successful attack on the asymmetric pseudonym change strategy in the US model IntelliDrive using a simple urban traffic simulation. They considered a squared, 1km² big city. The streets are all one-way streets, straight, regularly arranged and divide the city into 100 identical square blocks with a side length of 100m. Each vehicle is randomly assigned a home and a work address. Each morning, all vehicles drive to work at approximately the same time and spend a set time before returning home, hence there are two traffic peaks per day. Vehicles always use the shortest route to their respective destinations. Each time a junction is crossed a message is sent. The pseudonyms are taken from a pool and can be reused. The scenario used is rather unrealistic and the attack carried out cannot be transferred to the European model.

Buttyán et al. [63] introduced the concept of mix zones for vehicular pseudonym schemes, which is helpful for the analysis of the effectiveness of pseudonym changing strategies. They simulated a traffic scenario on a simplified map of Budapest with the traffic simulator SUMO. The simulation scenario is more realistic than the one discussed by Troncoso et al. as it is based on real streets in a real city. Nevertheless, the street map of Budapest's city centre is highly simplified and limited to a few major roads. In addition, the simulation artificially determines how many vehicles appear in traffic.

Förster et al. [163] presented a framework for the evaluation of pseudonym change strategies, with the help of which different steps in an evaluation can be conceptually well analysed. They based their work on the findings of Buttyán et al. [63]. Additionally they presented a new type of attack strategy and applied it with good success to the Budapest scenario. They analysed two different pseudonym change strategies. To the best of our knowledge, there is no study which uses realistic traffic scenarios to address the issue of traffic complexity.

4.1.2 The European Pseudonym Scheme

For the European context, security and privacy of V2X communication are primarily defined by the EU's C-ITS Platform, whose policies build upon specifications of the European Telecommunications Standards Institute (ETSI). The C-ITS platform's security policy [143] and certificate policy [142] are envisaged to govern security, privacy and trust aspects of ITS deployment within the EU. The certificate policy defines the European C-ITS trust model and builds upon the PKI architecture presented in [140] (see Sec. 4.1.2.1). The security policy [143] proposes legal entities and bodies to take over the roles defined in the certificate policy, defines security levels for several ITS message types and prescribes mandatory minimum controls for V2X communication. Further, the security policy defines a strategy for regular change of pseudonyms, based on a proposal by the Car-2-Car Communication Consortium (C2C-CC) [139]. The C-ITS security

policy recommends the pseudonym change strategy, described in Section 4.1.2.2.

4.1.2.1 European ITS Trust Model

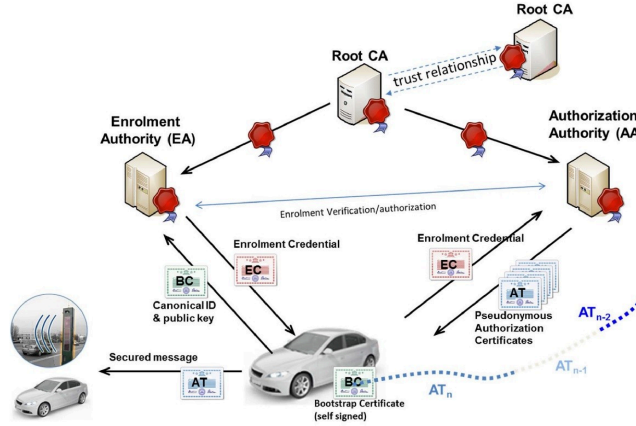


Figure 4.2: ETS ITS PKI architecture [140].

In the ETSI documents, classic asymmetric cryptography is used to implement pseudonymized V2X communication. Accordingly, a hierarchical public key infrastructure (PKI) exists for issuing the necessary certificates. Vehicles and RSUs participating in the communication are collectively referred to as *ITS stations* in ETSI jargon. As can be seen in Figure 4.2, the ETSI ITS PKI architecture comprises three types of authorities [140]. The Root Certificate Authorities (CA) act as trust anchors for the whole PKI and issues certificates to Enrolment Authorities (EA) and Authorization Authorities (AA). Enrolment Authorities are responsible for providing authorized ITS stations (via global unique id - Canonical ID) with credentials that they can use to demonstrate that they are authorized to send specific types of V2X messages. To that end, ITS stations receive Enrolment Credentials (ECs). Using their EC, ITS stations can obtain Authorization Tickets (ATs) from Authorization Authorities.

The separation of EA and AA aims at increasing ITS stations' privacy by enabling them to trustworthily sign messages without the need to reveal their identity. Authorization tickets do not contain identifying information of the ITS station and are the certificates an ITS station uses to sign the V2X messages it sends³. Hence, an ITS station can sign messages while at the same time not revealing its long-term identity but only a short-term pseudonym. Receiving ITS stations can still trust in the authenticity and legitimacy of received messages, as ATs are only issued to ITS stations that have been authorized to send the respective messages by an EC, which in turn has been authorized to enrol ITS stations by a commonly trusted root CA. In order to further increase ITS stations' privacy by reducing the risk of long-term traceability, i.e., linkability of messages sent over time, ITS stations should regularly change their ATs.

4.1.2.2 European Pseudonym Change Strategy

The European C-ITS pseudonym change strategy considers a combination of changes according to fixed parameters and added random values. As a compromise between privacy and utility (technical or economic

³In order to make the text more readable, we use the short phrase "to sign a message with a certificate" instead of the more correct one: "to sign a message with the private key which belongs to the public key certified with the certificate".

feasibility), this strategy defines the objective, that at least 95% of all journeys can be divided into at least three segments. The basis for this numbers is the ‘exemplary estimate’ that 95% of all trips are longer than 10min or longer than 3km [139].

According to this strategy, the first pseudonym change occurs at the start of a new journey. The start of a new journey is considered when the vehicle’s ignition was switched off for at least 10 minutes, the ignition is then switched on again and the vehicle is moving. The purpose of these conditions is to avoid that frequent shorter stops e.g. at traffic lights are counted and lead to a pseudonym change. The second pseudonym change takes place randomly within a distance of 800–1500m from the starting point of the journey. The third pseudonym change takes place minimal 800m after the last change and an additional driving time of 2–6min. The fourth pseudonym change takes place randomly between the next 10km–20km. Every further pseudonym change takes place randomly between every 25km–35km.

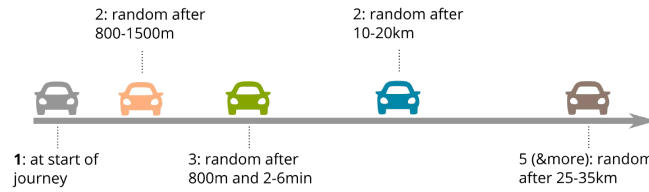


Figure 4.3: European C-ITS Pseudonym Change Strategy.

4.1.2.3 V2X Message Content

Since C-ITS messages are not encrypted, its also important to consider their content, as it can potentially be used to track vehicles. According to ETSI, there are different types of messages for V2X communication, e.g. Cooperative Awareness Messages (CAMs) or Decentralized Environmental Notification Messages (DENMs) [138, 137]. Here we focus on CAM messages, which are the basis of V2X communication. CAMs are broadcasted periodically (e.g. 10Hz). They are unencrypted but signed with a certificate. The message contains mandatory and optional vehicle information [137]. For location privacy, certain information can be considered critical, e.g. vehicle position, direction, speed, or size. Besides position and time of a message, here we will also focus on vehicle width and length, since these parameters do not change over time and are therefore well suited for linking pseudonyms. The vehicle length and width are given in decimeter [136].

4.1.3 Experimental Setup for Evaluating the Scheme

Our analysis of the European C-ITS pseudonym scheme is based on the slightly customized framework described by Förster et al. [163], which allows the evaluation of different pseudonym change strategies and attackers. Overall the framework consists of five phases. These are (i) modeling vehicle mobility traces, (ii) applying the pseudonym change strategy onto mobility traces, (iii) observing parts of the pseudonymized traces, (iv) learning of traffic statistics and attacking the observed pseudonymized traces and finally (v) evaluate the success rate of the attacker. In the following the different steps are described in detail.

4.1.3.1 Traffic Simulation

In the first phase, mobility traces were generated using the SUMO traffic simulator [286]. For the traffic scenario we used the Luxembourg SUMO Traffic (LuST) [85], a realistic urban traffic scenario. It includes a very detailed road network of the medium-sized city of Luxembourg with a total of 931km of roads on a total area of 156km². The road network also includes traffic lights at crossroads, inductive loops, and polygons of buildings and car parks. Figure 4.4 (left) gives an impression of the extent and level of detail of the road network.

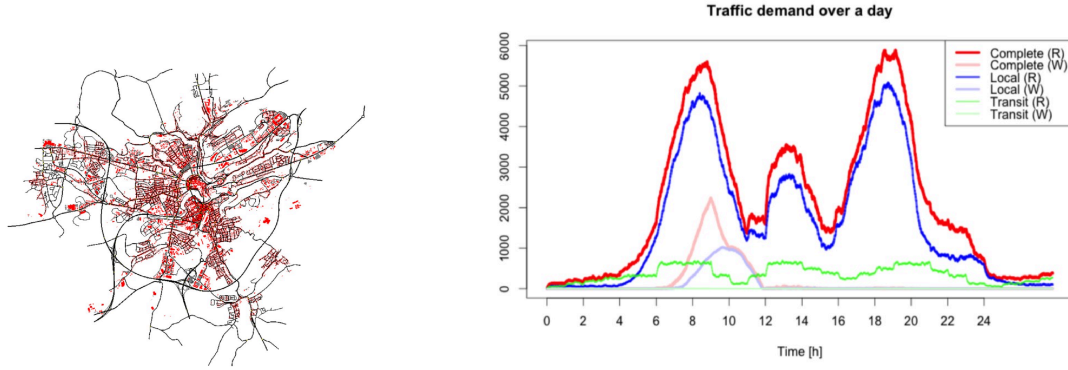


Figure 4.4: Left: Overview of the road network of the LuST scenario; Right: Number of vehicles in the LuST scenario over 24 hours [85].

In addition, the LuST-scenario provides traffic data generated for 24 hours. The generation of these data included extensive real information on the size and age structure of the population, schools, jobs and housing estates and known traffic characteristics of the city. The vehicle routes have been optimized so that the vehicles adapt their route to the traffic situation instead of choosing the shortest path. The LuST-scenario thus represents one of the best elaborated and most realistic scenarios available for SUMO. Figure 4.4 (right) shows the distribution of the vehicles in the LuST-scenario over 24 hours. The values marked with (R) show active vehicles and (W) show vehicles waiting for their journey to begin. In this work, we modified the LuST scenario slightly. As the main goal is to analyse personal vehicles, we reduced the data set by excluding bus traffic. Additionally, we removed transit traffic as well as the traffic of all vehicles whose journey begins or ends outside the city. Both adjustments only slightly reduced the total traffic. We used the Traffic Control Interface API (TraCI API) from SUMO to execute the simulation and to extract the traces of each vehicle in a one second granularity (similar to CAM broadcast). Each trace contains the following information: Vehicle ID (VID), time (in sec), position (in SUMO coordinates) and traveled distance. The result is a set of all mobility traces T .

4.1.3.2 Adding Metadata

Since CAM messages also contain the vehicle length and width, we added both properties to the mobility traces in our simulation. To get a realistic assignment of vehicle sizes we used a snapshot of all registered vehicles running in Luxembourg from November 2019 (provided by the Luxembourg authorities and their public data platform) [290]. This dataset contains 566,133 registered vehicles including trailers, busses, motorcycle etc. Since the attacking scenario is designed to trace individual traffic, we focused only on registered vehicles in the main category M1 (Regulation (EU) 2018/858; motor vehicles designed and

constructed primarily for the carriage of persons with max. nine seats [146]). This results in 322,359 vehicle entries with their type, brand and model. Because the official database did not contain any vehicle sizes we matched the given vehicle specifications with a database on vehicle properties⁴ (including a total number of 8639 types of passenger vehicles) to get the length and width (in a granularity of decimetres) for each model. Finally, we randomly assigned a model with corresponding length and width to each simulated vehicle, according to the distribution of [290].

4.1.3.3 Applying the Pseudonym Change Strategy

In the next step the mobility traces T are pseudonymized according to the C-ITS pseudonym change strategy (Sec. 4.1.2.2). The result of this phase is a set of pseudonymized traces $P = \{\text{pseudonym, time, position, length, width}\}$.

4.1.3.4 Modelling the Attacker

We consider an attacker who has control of multiple observation points (wireless transceivers) in the city, each point with a restricted observation radius of 100m, due to the limited range of the used V2X wireless technology (IEEE 802.11p or 5G sidelink). Observation points are stationed at traffic lights on junctions and therefore being able to observe all the traffic in and out of the junction in the range of his receivers. Similar attackers are also described by [163] and [63]. Further, we consider an observing attacker, who only listens but does not send messages nor tries to manipulate the communication.

The goal of the attacker is to track vehicles that change their pseudonym between two observation zones, i.e. the goal is to link different pseudonyms of the same vehicle. Everything outside the observation zone is called *mix zone*. We considered three attackers of different strength with observation points at 50, 100 and 200 intersections (fig. 4.5).

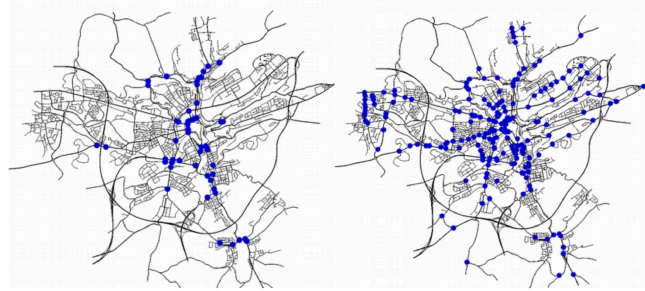


Figure 4.5: Attacker with 50 (left) and 200 (right) observed junctions in the Luxembourg scenario.

4.1.3.5 Observing Vehicles

To simulate the modelled attacker the set of pseudonymized traces P is reduced by the traces whose position is not within the observation range, which was done with the context subscription function of the TraCI API. The result is a set of observed pseudonymized traces O . From O it can be determined whether pseudonym changes have occurred within the observation zones. These observed changes can easily be linked by the attacker [482]. After linking all pseudonym changes which have taken place in the

⁴<https://www.cars-data.com>

observation area, the observed pseudonymized traces are reduced to *exit* and *enter* events, resulting in event traces E . An exit event describes an observed pseudonymized trace where the vehicle leaves the mix zone and enters the observation zone. The enter event describes the trace where the vehicle leaves the observation zone and enters the mix zone (see Fig. 4.6).

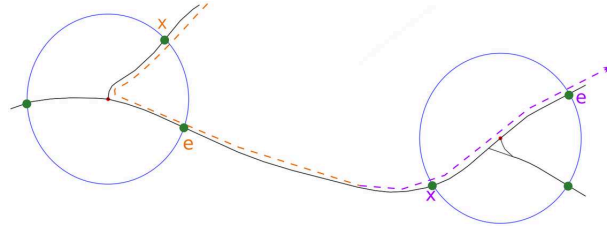


Figure 4.6: Observation zones (blue) with 100m radius and fine granular observation points (road entries) in green. Mix zone traversal with pseudonym change (dashed lines) with exit and enter events.

4.1.3.6 Learning Traffic Statistics

During the fourth phase, the attacker executes his attack on the set of observed traces. His aim is to track vehicles driving through the mix zone, i.e. to find out which enter events belong to which exit events, even if the pseudonym has changed. To make this possible, the attacker first has to generate knowledge about the traffic flow in the city. Therefore he uses the data of observed vehicles that have not changed their pseudonym between entering and leaving the mix zone to build up statistics. In this way, for each pair of observation zones, the attacker can record the number of vehicles that travel back and forth between the points, as well as the average time it took them.

To improve the original approach [163], we added fine granular observation points (road entries) to the observation zones and map each event to the nearest fine granular point (see Fig. 4.6). Thus, the statistics are not generated for pairs of observation zones but for fine granular pairs of road entries of observation zones. The learned statistics about how many vehicles drove through such a pair of observation points and how long it took them on average are finally used to attack the pseudonym changes between these points, i.e. to track vehicles traveling through the mix zones.

4.1.3.7 Attacking the Pseudonymization Scheme

The attack consists of the attacker using all events that have not yet been matched with any other event during learning to construct a weighted bipartite graph. Each node in this graph represents a particular event. Each edge connects an enter event with an exit event. The weight of these edges represents the probability that the corresponding events belong to the same vehicle (Fig. 4.7).

For the calculation of the weight the statistical data obtained during learning is used. The weight of the edge (see algorithm 1) between an enter event e , observed at the fine granular observation point $e.position$, and an exit event x , observed at $x.position$, increases the more vehicles ($nr_vehicles$) previously traveled between $e.position$ and $x.position$. It decreases the more the time interval between the two events ($travel_time$) deviates from the average time (avg_time) measured between $e.position$ and $x.position$. To improve the attack by [163], we added a validity check using the minimal and maximal learned time and added a safety margin to make sure the matched results could be realistically possible and to reduce the

Algorithm 1 $\text{weight}(e, x)$

```

if  $e.time \geq x.time$  then
    return 0
else if  $e.length \neq x.length$  or  $e.width \neq x.width$  then
    return 0
else if no trip from  $e.position$  to  $x.position$  in statistics then
    return 0
else if  $travel\_time > max\_time$  or
 $travel\_time < min\_time$  then
    return 0
else if  $(travel\_time = avg\_time)$  then
    return  $nr\_vehicles$ 
end if
return  $nr\_vehicles / \text{abs}(travel\_time - avg\_time)$ 

```

size of the graph of matched sets. Additionally, we excluded all exit-enter event pairs which are not of the same vehicle length and width. After the graph has been constructed, that matching is calculated, for which the graph reaches a maximum cardinality and a maximum weight. The attack is not only applied once to all events in the database, but the registered events are processed at intervals of a certain length t . The length of the time intervals between the individual executions of the attack determines the size of the resulting graph and thus the size of the graph matching problem. Here, an attack interval of 300s was selected. Within the interval all enter events between t_0 and t_{300} and all exit events between t_0 and t_{600} are collected (Fig. 4.7). Then the weight between all pairs of exit and enter events is calculated. Unmatched exit events between t_{300} and t_{600} will be used again in the following attack interval. The output of the attack is a set of matches M , which are the decisions of the attacker of which pseudonyms belong to the same vehicle.

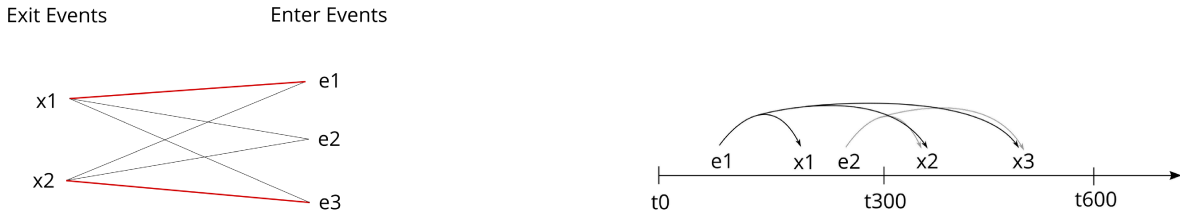


Figure 4.7: Bipartite graph of exit and enter events and attack interval.

4.1.3.8 Evaluating Attack Success

In the last phase, the success of the attacker is evaluated by comparing reconstructed traces, represented by the set of matches M , with the original mobility traces T from the first phase.

4.1.4 Results of the Analysis

In the following section we describe the results of the simulation.

4.1.4.1 Mobility Simulation

In total, the simulated traffic comprised 214,315 vehicles with overall 209,654,111 mobility traces T . The average travel time is 16.33min and the average travel distance is 8.62km. In total, the vehicles performed 602,742 pseudonym changes. The number of pseudonym changes made during the journey per vehicle can be seen in Table 4.1.

Table 4.1: Number of pseudonym changes per vehicle.

nr. of changes	1	2	3	4
nr. of vehicles	2,758	38,900	168,444	4,213
percent	1.3%	18.2%	78.6%	2.0%

1.3% of the vehicles performed only one change during their journey (at the start); 18.2% completed two and 78.6% three changes. Thus, in this realistic traffic scenario, the specified goal of the pseudonym change to divide the journey of at least 95% of the vehicles into 3 segments, is not achieved. However, this assumption can also be questioned in general when considering various traffic studies (see also Appendix II-A).

4.1.4.2 Width and Length of Vehicles

Considering the distribution of registered vehicles and their properties in Luxembourg [290], we can identify an issue of V2X pseudonymization in general. A number of 99 (0.03%) vehicles had such a unique combination of length and width that they were only once registered in Luxembourg and thus could always be tracked. Additionally, 18,903 (5.86%) vehicles have a size that they share with less than 100 other vehicles (72 sizes for which only two vehicles exist). These vehicles, although not unique, can simply be deanonymized, as it is very unlikely that a vehicle of the same size will be at the same time in the vicinity. On the other hand, there are also 8,703 (2.67%) vehicles, which are of the same size. Hence, drivers driving vehicles of this class are more likely not to be tracked, due to a higher probability of other vehicles of the same size in the area. The whole distribution of size properties can be seen in Fig. 4.8.

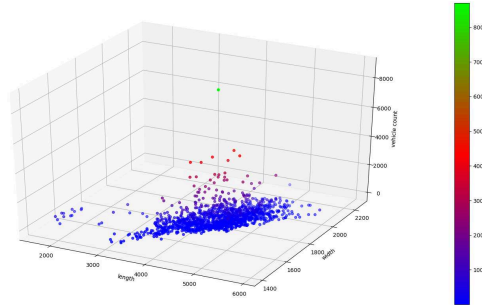


Figure 4.8: Distribution of registered vehicles in Luxembourg regarding size properties (length and width).

4.1.4.3 The Attacker View — Observing Vehicles

The attacker's observations and attack capabilities naturally vary with the number of observation points. Table 4.2 gives a general overview about observation possibilities of the attacker.

Table 4.2: Observation Possibilities of the Attacker

Observation Zones	200	100	50
Nr. of vehicles	205,497	195,813	169,028
% of all vehicles	95.89%	91.36%	78.87%
Observed changes	86,605	64,736	41,778
% of all changes	14.37%	10.74%	6.93%
Mix Traversals	152,037	115,965	67,693

With 200 observation points within the city, the attacker can see overall 95.89% of all vehicles driving through Luxembourg. Furthermore, he can directly observe 86,605 pseudonym changes, because they were performed within an observation zone. Overall, with 200 observed junctions, there are 152,037 mix zone traversals, where a vehicle leaves an observation zone and enters another one with a changed pseudonym (independent on how often it changes between the two points). With only 50 observation points, 21% of the cars were never seen.

4.1.4.4 Linking Pseudonyms of Mix Zone Traversals

Next to directly observed pseudonym changes, the attacker tries to link mix zone traversals, i.e. link exit and enter events to a specific vehicle even if it changed the pseudonym. With the slightly modified original attack (only position and time), the attack success, compared to [163], is rather poor (see Tab. 4.3-1). A possible reason could be the realistic traffic scenario: The road network is extensive, so that vehicles can travel from one observation point to the next in several different ways; traffic light waiting times and possible traffic jams cause delays in the traffic flow. These reasons mean that the average travel times measured from point to point are less meaningful. Additional false positive rates are not mentioned by [163].

Table 4.3: Linking Pseudonyms

Observation Zones	200	100	50
Mix Traversals	152,037	115,965	67,693
1) Linking with Position and Time			
Precision	11.87%	7.50%	5.25%
Recall	20.44%	15.46%	13.31%
F1-Score	15.02%	10.10%	7.53%
2) Linking with Position, Time and Size			
Precision	71.12%	59.46%	50.87%
Recall	76.14%	71.62%	73.91%
F1-Score	73.55%	64.98%	60.26%

However, adding the static vehicle characteristics of length and width to the attack, the success rate increases enormously (see Tab. 4.3-2). With 200 observed intersections within Luxembourg, the attacker was able to correctly link an enter with an exit event, and thus the pseudonyms of the corresponding vehicle, with $> 70\%$ precision. It can also be seen that the number of observation points has a high influence on the success rate of the attacker. The smaller the observation possibilities the worse the precision of linking. One reason for this is an increased number of events which have no counterpart in the observation area, but which could erroneously be matched with another event.

The influence of traffic volume on the success of the attacker can be seen in Figure 4.9. An increased traffic flow decreases the precision of the attacker. This was expected, as more vehicles with the same size characteristics appear side by side in the observation zones and therefore more event combinations occur.

4.1.4.5 Linking the Entire Journey of Vehicles

Table 4.4 shows the distribution of mix zone traversals per vehicle as well as the attacker's success in linking the entire journey of them. Entire journey means the tracking of a vehicle from the moment the attacker observed it for the first time until the last time. The number of mix traversals describes how often the vehicle appears with different pseudonyms, i.e. mix zone traversals the attacker has to link correctly to follow the entire route of the vehicle.

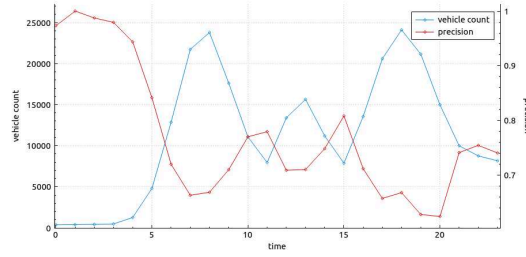


Figure 4.9: Correlation between traffic volume and precision of the attack.

Table 4.4: Mix traversal possibly linked by the attacker.

Observation Zones	200	100	50
Observed Vehicles	205,497	195,814	169,029
Linked Journeys	82.99%	83.64%	89.75%
0 Mix Traversals	83,822	97,012	107,616
Linked	100%	100%	100%
1 Mix Traversals	91,624	81,793	55,144
Linked	73.88%	69.03%	73.07%
2 Mix Traversals	29,740	16,855	6,258
Linked	63.28%	60.75%	60.64%
3 Mix Traversals	311	154	11
Linked	63.02%	51.30%	72.72%

Vehicles with zero mix zone traversals can be directly followed over the entire route, cause they either have been observed only briefly or they performed their pseudonym changes directly within observation zones. Vehicles with three mix zone traversals had four pseudonym changes (first at start, see Tab. 4.1), which have all taken place unobserved by the attacker (within the mix zone). After all, about 60% of these cases could be linked correctly. Trips that were divided into three segments (two mix zone traversals), which is the goal of the C-ITS pseudonym change, could be linked with a success rate of $> 60\%$.

Overall $> 80\%$ of observed vehicles could be linked over the entire route, regardless of the number of observation stations. Even though the success rate of the attacker decreases with an increasing number of mix zone traversals, tracking cannot be prevented effectively. The creation of comprehensive motion profiles is especially feasible for stronger attackers (e.g. RSUs such as traffic light systems), as success depends on the distribution and count of the observation points. However, targeted attacks are also conceivable, for example on special points of interest such as workplaces (e.g. police station) in order to find out who works there, where a person lives or when she comes to work. Thus, observation stations can be set up in a targeted and gradual manner.

4.1.5 Transparency for Connected Driving

If V2X communication will be realized in the EU with proposed pseudonym change, thus maintaining the gap between utility and location privacy, this should at least be communicated transparently to the user. We could think, for example, about ex-ante transparency. Similar to the privacy labels of Emami-Naeini et al. for IoT devices [128, 127], vehicles could be equipped with transparency information, to inform the user about location privacy risks before purchasing a vehicle. Vehicle characteristics that influence the location privacy (e.g. very specific length and width) as well as information on anonymity conditions (e.g. minimum distance, or driving time) could be integrated. Thus, the user is already informed about possible privacy risks at the time of purchase and can react accordingly in required situations. Another

approach could be a transparent visualization of the state of location privacy in the vehicles infotainment system (Audit TET). The state could be predicted depending on journey and pseudonym characteristics. Parameters could for example be the number of pseudonym changes, travel distance, travel time, number of vehicles in the vicinity (e.g. with the same length and width), or also other switched-on identifiers like Wi-Fi or Bluetooth. Depending on the sensitivity of the trip (e.g. journalistic work, doctor's visits) or the preferences of the user, one could also consider making the changes interactive and give the user the option to increase the pseudonym change intervals (Intervention TET).

4.1.6 Concluding remarks regarding Connected Driving

In this chapter we used the slightly improved framework from Förster et al. [163] to analyse the effectiveness of a pseudonym change strategy, recommended by the European C-ITS platform. For this purpose, we simulated a realistic urban traffic scenario within Luxembourg and added realistic vehicle characteristics of width and length, which are transmitted via V2X communication. Further, we modelled attackers of different strength (number of observation points), which try to link pseudonym changes with the help of learned traffic statistics and vehicle properties.

Overall, linking pseudonyms with simple traffic statistics within a realistic city scenario is more challenging than related work suggests. However, the consideration of additional information from the V2X communication, such as length and width of the vehicle, enormously improves the linking of pseudonyms, and thus enables tracking of vehicles and generation of motion profiles. Around 80% of the vehicles observed could be tracked from the first to the last observation point, regardless of the strength of the attacker. However, the stronger the attacker, the more vehicles and longer distances can, of course, be observed. Furthermore, the precision of the attack increases with the number of observation points. Besides the observation possibilities, the success of the attacker is also particularly influenced by the number of vehicles in the vicinity and especially their properties. For example, the length and width of a vehicle can be unique in its vicinity, so that the attacker can clearly track this vehicle, regardless of the pseudonym scheme used. In addition to length and width, other information such as Wi-Fi or Bluetooth identifiers could potentially be used to bridge the pseudonym change. We describe a detailed investigation of such further identifiers in Appendix II-B. Furthermore, improved attacks, which are not only based on simple traffic statistics, e.g. with more information about the road network and with improved learning algorithms could increase the success of linking pseudonyms.

Overall, our results suggest that the introduction of VANETs, in addition to enabling more effective and secure road traffic, also enables vehicle tracking, thus reducing location privacy in the future, even with the C-ITS pseudonym scheme. Although it prevents simple attacks, the achieved protection is strongly dependent on the journey (e.g. surrounding vehicles, distance, travel time), and vehicle characteristics (e.g. size), which leads to the fact that some trips can be clearly traceable. Thus, in addition to the issue of location tracking of mobile devices [20], the simple use of a physical vehicle will also restrict location privacy.

4.2 Transparency for Bystanders in IoT regarding audiovisual Recordings

In addition to wireless communication, smart cars are or will also be equipped with a range of additional sensors. By integrating radar and camera sensors, vehicles are empowered to understand their environment and react to it automatically. From the location privacy perspective, based on camera systems (also used e.g. by RSUs), vehicles could thus also be identified and tracked automatically, e.g. via Automatic Number Plate Recognition (ANPR) [328].

Overall, with the development of IoT, not only vehicles but a variety of physical objects are equipped with unobtrusive, audiovisual sensors. Thereby, the capturing and processing of *acoustic and visual signals*, e.g., through surrounding voice assistants, augmented reality (AR)⁵ glasses, or smart cars, could represent a significant intrusion into personal privacy by allowing a variety of profound inferences regarding further sensitive information, like personality [234], health status [265], sexuality [470], or political orientation [253]. Moreover, instead of used-when-needed such smart devices are always-on, always-recording, and always-connected.

One primary concern in this regard is that not only the owner of such smart devices, but everyone who enters the recording radius of their sensors is affected. These *bystanders* currently have no way to recognize these devices, to know what is being recorded, how that data is being processed, and who is involved, let alone object to the recording. Consequently, there is a *foreign control* over the personal data of bystanders, interference is hard, giving consent difficult, and being informed about its ramifications impossible, in consequence.

Therefore, in this chapter we introduce our transparency concept and its prototypical implementation, to enable insights into surrounding, audiovisual IoT devices in everyday life. I.e. a transparency solution for bystanders, regarding the possible recording of biometric data such as voice or face, as well as their processing and storage. In addition, we conducted a semi-structured interview to analyse whether and how the bystander issue is perceived by participants of the German culture as well as whether our transparency approach is considered as a sensible solution.

4.2.1 Bystander Perceptions regarding IoT

In addition to the difficulty of detecting tiny recording smart devices, bystanders are usually unaware of how they work, such as whether and which data a device can collect at all [299, 9, 495]. Most are dissatisfied with this situation, want to better identify such devices and want to be informed about data collections that make them uncomfortable [333]. However, when and how much one feels uncomfortable varies between individuals and depends on several technological and contextual factors [333, 504]. The type of recorded data matters considerably; While the capturing of biometric data like face, voice or fingerprints raises greater concern among bystanders, other sensor data is scarcely relevant [495, 298] (even other sensor data can result in privacy issues, e.g.⁶).

Contextual factors which impact perceived intrusiveness pertain the trust and relationship to the device owner [299], the device's primary functionality [9], the specific location of the recording [298], the recording's purpose, if data is shared with third parties, and if it is used to infer additional information

⁵<https://www.theverge.com/c/22746078/ar-privacy-crisis-rethink-computing>

⁶<https://www.technologyreview.com/2019/06/27/238884>

[333, 504]. Many concerns are situational and are much higher in the private context than in a public one [333]. As exemplified by Bernd et al.'s [40] interviews with nannies, the professional context is also less of a concern for some than the private one. Few participants in their study, however, also indicated that they changed their behavior during work (less singing, joking, ...) if they were aware of domestic surveillance.

Besides behavioral change, bystanders use protection mechanisms like blocking/obstructing the sensor or negotiating with the device owner. However, they would like to learn better strategies and gain better control over their data [299, 9]. However, this control should neither lead to an overload of notifications nor take place completely automatically in the background, but remain with the user [89].

4.2.2 Towards Transparency: A Background

In order to design an appropriate transparency solution for bystanders, several challenges have to be considered: How can devices be detected? How can transparency information be transmitted to the bystander? Which information should be transmitted?, and: How can this information be presented so that it is easily accessible in everyday life? Therefore, in the following, we take a closer look at related work regarding these questions.

4.2.2.1 Detecting Recording Devices

To be able to give the bystander a transparent overview of their environment, first of all, surrounding audiovisual recording smart devices have to be detected. On the one hand, this could be done using technical approaches. Possibilities are for example automated visual recognition [360] (e.g. object detection via AR glasses), or by analyzing wireless network streams [77, 408, 398]. However, such technical approaches only work for certain recording devices, hence, cannot give an entire overview and are therefore not suitable for daily usage.

Another approach is to adopt regulation of IoT devices, meaning that recording of audiovisual data must be made identifiable. Several proposals have been made out of mostly ethical reasons, to give people the possibility of being informed, and thus can allow or deny access to their personal information [269, 472, 185]. In 2018, the EU settled on the GDPR [145] (Sec. 2.1.2). Here, article 13 states, that if personal data is collected, the data subject (DS) shall be provided with information about the data controller (DC) as well as with insights into data processing and storage. Thus, recent studies base their work on the assumption of regulation due to the GDPR [324, 71, 102]. In contrast, other work questions whether the GDPR alone may not be concise enough to deal with the complexity of IoT yet, due to the imbalance between DC, the person owning the recording device, and data processor (DP), the company, which handles the processing of the data [280].

4.2.2.2 Information Transmission

If we assume that the identification of such devices is regulated, we need a channel to transmit information about data collection and processing to the user. Devices could, for example, be equipped with visual (e.g. LEDs) or acoustic indicators to highlight their recordings and facilitate their general recognition [412]. However, such approaches cant provide further information about data processing and possible privacy implications. One approach to alerting bystanders to the presence of such devices while also

communicating transparency information is to display analog signs on site. Like current warning signs for video surveillance in public areas (e.g. regulated in Germany through BDSG⁷), data controller could offer a sign for possible data acquisition, e.g. when entering the smart home. The DTPR project⁸, e.g., creates signs by using a simple visual language and unified taxonomy that allows users to understand the complex information flows of the IoT more easily and adds QR codes for further information (Fig. 4.10). However, we argue that such analog information transmission is only suitable for specific scenarios, like



Figure 4.10: Left: Analog sign example by DTRP Project. Middle: Practical analog sign within Dresden. Right: Practical surrounding recording devices without any transparency within Dresden.

at the entrance of a department store, but not for the entire dynamic IoT world. Such concepts would lack information, are not suitable for mobile IoT devices, flexible and adaptable.

To remedy the weaknesses of analog approaches, other TET approaches communicate the presence and characteristics of IoT devices digitally. This enables the current state of the bystander's vicinity to be visualized directly, e.g. on a mobile device. Thereby, this information can be transmitted either directly or indirectly [324].

Research using *direct communication* describes a setup in which IoT devices announce transparency information via a short wireless signal, which in turn is received, interpreted, and visualized by a bystander's personal device. Langheinrich [268, 269] proposed one of the first TET concepts of this type called *privacy awareness system* (pawS). For audiovisual data recordings, which he calls *active policy announcement*, he proposes a *privacy beacon*, which communicates with the user device over a short-ranged wireless link. Thereby an identifier is transmitted to the user device, which will in turn communicate with a service proxy on the network to request the privacy policy. In addition, it provides an infrastructure for comparing the user's privacy preferences with service policies of smart devices, thus enabling interaction. Several technologies for the communication between IoT and user device are proposed, such as Infrared, Bluetooth, WiFi, and ZigBee. Infrared was used as the communication technology, due to the readily availability in then-common PDAs. Morel et al. describe another design of a direct transparency and consent system for IoT [324, 71]. As the communication channel, Bluetooth Low Energy (BLE) is proposed over alternative short-ranged wireless technologies, due to high availability within smart phones, and privacy benefits due to no involvement of other systems. IoT devices are enhanced with a *Privacy Beacon*, which is responsible for the declaration of this device. Beacons broadcast in set intervals all information required by the GDPR within BLE *advertisement packets*. Additionally,

⁷https://www.gesetze-im-internet.de/bdsg_2018/_4.html

⁸<https://dtp.helphelpplaces.com>

they suggest the usage of the BLE Attribute Protocol to enable a lightweight communication of consent between the user's device and IoT device. The aspect of visualization of transparent information was perceived as important in both contributions, but omitted. The advantage of direct communication is that transmitted information can be received passively, i.e. the use of additional services is not required; thus no further information is disclosed. However, equipping smart devices with beacons generates additional costs, and bystanders could only be notified within beacon range, making preventive notification (Ex-Ante) difficult.

With *indirect communication*, on the other hand, the user device connects to an external database that contains information about nearby smart devices. Das et al. offer a solution for transparency and consent tracking in IoT, using such a communication setup [102]. For this, they introduce IoT Resource Registries (IRR), which “act as a location-aware lookup service that supports the discovery of nearby IoT resources”. IoT device owners are expected to register their device in the local IRR, including corresponding privacy information. Additionally, they can provide APIs of IoT devices for consent purposes, such as opt-out [103]. A *Personalized Privacy Assistant (PPA)* is proposed as a smartphone app, which queries the local IRR and notifies the user about surrounding IoT devices and their characteristics. In their implementation, the user devices detect specific WiFi access points and Bluetooth beacons to determine the location of the user⁹. Shaw et al. [400] propose the use of IoT maps. Similar to the previous approach, a central database is offered by device owners, administrators or crowd based, storing IoT device location as well as information about owner and characteristics. Existing IoT devices are visualized on a map in relation to the bystander's location. With indirect communication, smart devices can also be detected outside their proximity, enabling preventive insights and notifications (Ex-Ante TETs). Further, smart devices do not need to be equipped with additional hardware, thus avoiding manufacturing costs. On the other hand, it yields costs for maintaining the needed infrastructure; it produces effort for device owners who have to register their devices in registries or set up their own; the correctness of transmitted transparency information is highly dependent on the timeliness of the corresponding registry; and mobile smart devices cannot be detected. In addition, the passive nature of notification is lost, leading to potentially new privacy issues, such as the disclosure of location information.

4.2.2.3 Information Content

Next to an appropriate transmission channel, the information content itself has to be considered. Existing approaches are mainly based on regulatory requirements for privacy policies and sensor characteristics [324, 71, 102]. The GDPR for example requires, that the data subject must be informed about the processing of their personal data, which includes for example identity and contact details of the data controller, purposes of the processing, recipients, data retention time, as well as information regarding rectification and erasure of personal data (see Sec. 2.1.2). Morel et. al. [324] represent the point of view, that position and range of the IoT device should be included as well, to make the given information more contextual.

⁹<https://www.iotprivacy.io>

4.2.2.4 Presentation of Information

Finally, the information presentation has to be considered. Since IoT devices are ubiquitous, the transparency solution must also be omnipresent. Therefore, a solution needs to be unobtrusive and must provide a usable interface in an everyday scenario without overwhelming the user [272, 472, 269] (see also Sec. 2.3.3). Users probably don't want to constantly read privacy policies of nearby IoT devices in their daily routine (see Sec. 2.1.2).

Hence, various projects aim to aggregate and simplify privacy policies [497, 440, 368, 506, 127]. Initial approaches in this area were based on the idea of standardizing policies in machine-readable form. A promising standardization approach was the World Wide Web Consortium's (W3C) "platform for privacy preferences" or P3P [98]. Based on this simplified representations were developed¹⁰. Kelly et al. [241], for example, proposed a visualization as "privacy nutrition label" based on P3P, which improved accuracy and time in understanding and finding policy information. Unfortunately, neither P3P, or any other standardization has been established in practice.

Current approaches, therefore, try to either manually¹¹, automatically [497, 440, 368], or hybrid [506] understand and aggregate privacy policies. The drawback of manual, mainly crowd based, approaches is the difficulty with processing large amounts of policies, while automated ones could fail in full accuracy. Afterwards, however, policies are e.g. ranked with privacy grades [497], or they are made searchable more quickly¹². *ToS;DR* or *PrivacySpy*¹¹, for example, define evaluation mechanisms that take different aspects of privacy policy characteristics into account. Based on sub-scores of the policy properties, a final privacy rating is derived for websites or companies. In this way, the user receives a quick, comprehensible, and discreet indication of the current privacy state when visiting a website.

Other approaches attempt to simplify the individual aspects of a privacy policy based on iconification [467, 133, 210, 123, 194], like the Mozilla privacy icons project¹³, which tried to iconify data handling properties for faster information acquisition and comprehension. Unfortunately, no standardized iconification has prevailed and a lot of developments have been discontinued. Moreover related work in this area was mainly focused on web browsing or mobile app usage [329]. For IoT transparency their has to be done additional work, cause additional information could be important, like the usage of a wake-up word before data processing, or local processing of biometric data (e.g. voice transcription directly on the device). Overall, simplification also leads to a loss of information, which must be taken into account when developing a transparency solution, i.e. weighing up between functionality and usability.

4.2.3 Designing an Audit TET for Bystander in IoT

In the following we present our concept and prototypical implementation of an Audit TET for surrounding audiovisual IoT devices. The categorization parameters of this TET according to Zimmermann [507], could be find in Table 4.5.

¹⁰e.g.: <http://www.privacybird.org>

¹¹e.g. <https://tosdr.org>, or <https://privacyspy.org>

¹²<https://www.usableprivacy.org>

¹³https://wiki.mozilla.org/Privacy_Icons

Parameter	Value
Target Audience	data-subject
Data Types	observed / reported
Execution Environment	client-side
Application Time	realtime / ex-post
Interactivity Level	read-only
Delivery Mode	push

Table 4.5: Classifying the envisioned TET according to parameters described by Zimmerman [507]

4.2.3.1 Information Transmission

Similar to related work, we assume a regulation of audiovisual IoT devices. In particular we expect a direct communication channel which requires that IoT devices identify themselves through some kind of broadcast signal. A regulatory implementation could be done, e.g., during manufacturing by integrating a wireless beacon which broadcasts its data from switch on. Following Morel et al. [324], we used BLE as a communication technology due to its low power consumption, low cost, generally suitable range, and high availability in current mobile devices. The practicality of BLE for such a use case may also be demonstrated by its use for proximity tracing in various Covid-19 apps.

The IoT device continuously broadcasts information using BLE GAP (Generic Access Profile), which defines the *central* and *peripheral* role. GAP offers 5 link layer states: Advertising, Scanning, Standby, Initiating and Connection. The advertising mode features two possible methods for inter-device communication: Advertising packets and Scan Response packets. The former can be sent in an interval from 20ms to 10.24s, and can contain up to 31 bytes of payload. IoT-devices use the peripheral role and broadcast information using these advertising packets. Instead of explicitly embedding transparency information within these packages, it will be transmitted by reference. Thus, it is more flexible to changes in privacy policies and bypasses the limited package size of beacon specifications in terms of transmitting entire policies.

User devices use the scanning mode to receive such broadcasts. Due to the short advertising interval, devices are detected in real-time. If data is received, the payload of the advertisement package is extracted, included reference IDs are queried in a database, which returns the privacy information of the IoT device. To maintain the passivity of the approach, the database is stored locally and could periodically be updated. After retrieving the information, the user device can directly inform the user about the current situation regarding nearby recording devices and their characteristics.

For the daily use of this transparency solution, our approach for notification and visualization is to support user devices that are directly accessible, i.e. wearables like smart watches or smart glasses. Thus, the user can be informed quickly and efficiently without being torn from his activities. The user's smart phone would take over the scanning mode, query correlating information when devices are detected and then display a notification and/or forward important information to a wearable of the user. The whole communication concept could be seen in Figure 4.11.

4.2.3.2 Information Content

Related work is based on regulatory requirements for privacy policies and sensor characteristics [324, 71, 102]. Here, e.g. the GDPR expects that the data controller (DC; the IoT device owner) must make all transparency information public to the data subject. However, currently, IoT recordings are usually

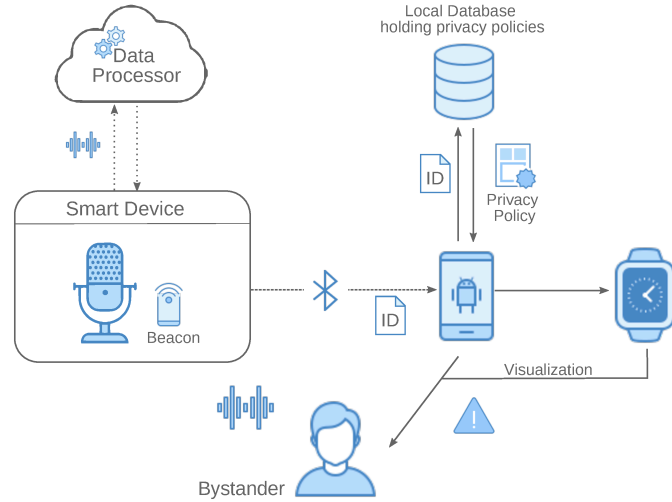


Figure 4.11: Communication setup for transparency solution

handled by the data processor (DP), e.g. the manufacturer, who also decides contract terms due to the power imbalance between DC and DP [281]. Thus, we argue, that for IoT transparency information about the DP are almost more important than about the DC, cause in most cases the DC does not have complete control over the data process owed to the current design of IoT devices. Hence, our solution provides information about the data handling by the DP as well as device characteristics. In addition to the associated privacy policy of the DP, the following information are considered: recording channel, type and name of the device, retention time, involved third parties, usage of a trigger/wake word before recording, and whether the connection to DP is encrypted. Our scenario requires for the DP to embed a BLE beacon with a corresponding transmission ID into the IoT device during production, and then to enter information about this ID into a corresponding database. Information about the DC are excluded for now, but has to be considered in future, because they have also (in part) access to the data, even though most of them probably have no idea how to handle them.

4.2.3.3 Information Visualization

In order to provide the user a quick and clear insight into the current situation regarding nearby recording devices, we prepare transparency information in a hierarchical structure of three layers.

The first layer represents the corresponding *recording channel*, i.e. audio or video recording, and the number of devices in the vicinity. Thus, the user is provided with a quick and comprehensible overview of which biometric data is possibly captured and from how many devices. This layer corresponds to well-known traditional warning signs (e.g. for video surveillance) and should always be accessible, for example as a notification of the smart phone or as a icon on the smart watch. For further information, the second layer shows the IoT device types (e.g. smart speaker or smart tv) per recording category. The third layer displays the device models and manufacturers of the specific IoT devices (e.g. Amazon Echo) and its privacy policies.

Rather than simply presenting a link or the policy itself, we decided to mock-up a simplification by color coded icons. As there is no widespread standardized icon set, we adopted the Mozilla Privacy Icons¹⁴ merged with own designs, to take IoT device characteristics into account, like the activation of data

¹⁴https://wiki.mozilla.org/Privacy_Icons

capturing through a wake-up word. The mocked-up iconification can be seen in Figure 4.12.










Retention period	Third-party use	Wake word	Connection to servers
 No storage beyond processing	 Intended Use Only	 Data processed only after wake word	 Connection encrypted
 30 days	 Limited re-use	 Data is always processed	 Connection not encrypted
 Indefinitely			

Figure 4.12: Mocked-up Privacy Policy Simplification via Iconification.

4.2.3.4 A First Prototype

As our focus lies on wearables as the user device, we decided to use a "Fitbit Ionic" smart watch for the first prototypical implementation. These watches mainly function as fitness trackers, but offer high customizability of the user interface. The UI is shown in Fig. 4.13. The left and middle image represents the default watch face, showing time and two *complications*. Complications are small badges, holding a specific part of information, which is usually represented with text and an icon.



Figure 4.13: First smart watch prototype

The left complication shows the first layer of the Audit TET. A situation with no recording devices nearby is conveyed through a crossed-out eye icon, and enriched with a corresponding text denoting "Private" (Fig. 4.13 left). An ear icon, with the text "Audio", reveals that there might be devices listening to the user (Fig. middle). An eye icon and the text "Video", visualizes that video recording devices are nearby. In addition, with the used colors the privacy state is preliminary classified. Thus, the first layer provides a quick insight into whether and what information could be recorded at the moment, i.e. face or voice. By tapping the TET complication, a popup shows a short text describing the found smart devices (Fig. 4.13 right). Additionally, a button allows to request additional information about the device, on the connected smartphone.

4.2.3.5 Pre-Evaluation

To gain a first impression of the usefulness, we presented our solution to a group of 18 users (Tab. 4.6) at the Dresden Science Night¹⁵ and had them evaluate it. For this purpose, a booth was set up with an

Table 4.6: Overview of the participants

Age	10-20	20-30	30-40	40-50	50-60
Count	5	5	2	4	2

IoT sample device (Amazon Echo) and the implemented smart watch TET solution. The project was explained to the visitors and they could try out both the Amazon Echo and the smart watch. When the Amazon Echo was activated, the communication channel was also triggered and the transparency display of the smart watch visualized an audio recording device nearby. In addition, the user was prompted to visit different rooms, making the change in state of the transparency solution practically visible when leaving and entering the exhibition space. After the practical evaluation, the users were asked to fill out a survey.

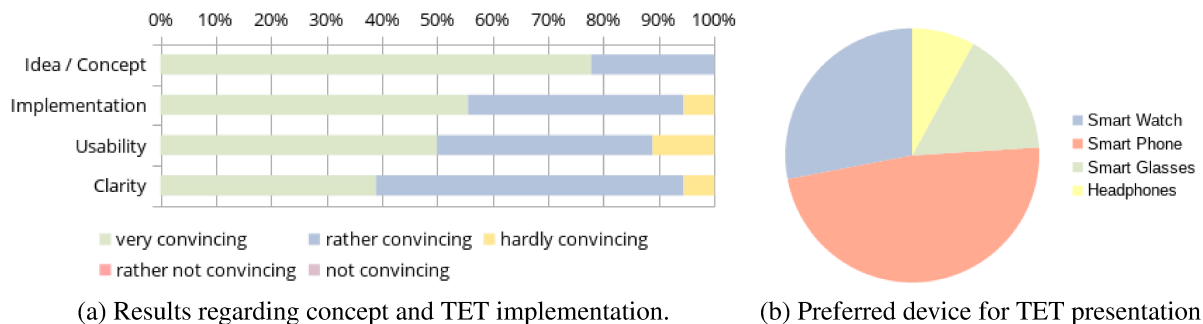


Figure 4.14: Pre-Evaluation Results with 18 participants.

The overall feedback regarding the transparency solution developed was very positive (Fig. 4.14). Especially the overall concept of the solution was felt to be very useful. In addition, 95% of the participants stated that they wanted to use this solution in daily life. More than half (59%) of them said that they had already found themselves in a situation where the app would have been useful. The deficiencies in implementation, usability and comprehensibility were mainly due to the fact that several participants wore a smart watch for the first time. In this regard, a transparency solution via smart phones was preferred by most of the participants (Fig. 4.14).

Suggestions for improvements and remarks were in particular a desired control of the own data (e.g. by integrated switch-off function), as well as a listing of the range of the device located nearby. The icons of the first layer were perceived partly ambiguous. Furthermore different design suggestions were introduced.

4.2.3.6 Extending the Prototype

Based on the pre-evaluation we extended our transparency solution. First we started an implementation for smart phones based on Android. Figure 4.15 shows the implementation of the information hierarchy (layer 1 to 3). The first layer represents the main view, where the 2nd layer can be accessed by clicking “Audio” or “Video”. Layer 3 can be accessed by layer 2 respectively.

¹⁵www.wissenschaftsnacht-dresden.de

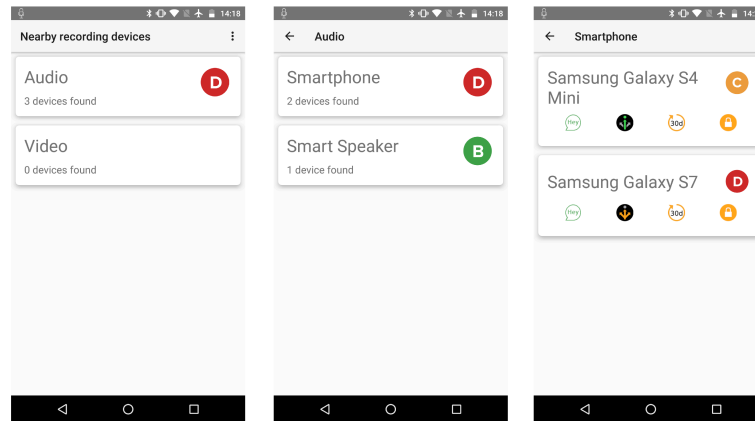


Figure 4.15: Smart phone implementation (From left to right: layer 1, 2, and 3)

For better usability in everyday life we implemented the first layer also for WearOS smartwatches (Fig. 4.16). Similarly to the first prototype, a complication for watch faces provides a quick overview of the current privacy state. The left image shows a private state, without any recording devices nearby, whereas the middle image shows three recording devices nearby, with all being audio processing devices. On click, the view on the right image is shown, where the first information layer is visualized, as well as the worst processing device from a privacy perspective and whether the device is always recording or is waiting for a wake word. Through the synchronization by the DataLayer API, changes on the smartphone are immediately reflected on the smartwatch.



Figure 4.16: Wear OS implementation

Another feature to support usability in everyday life and the principle of minimal distraction (see Sec. 2.3.3) is the usage of a persistent notification with a text and corresponding icon (see Fig. 4.17, left). This enables a quick overview of the users' privacy state without overwhelming.

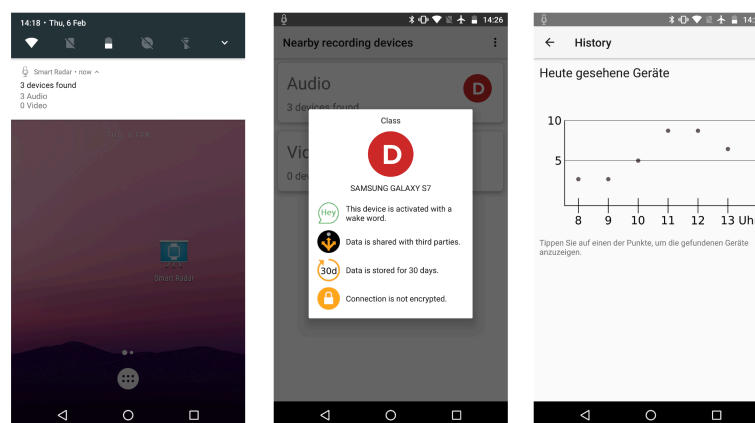


Figure 4.17: Smart phone implementation (ltr: notification, popup view, history view)

Hence, the complication on the smart watch and the smart phone notification visualize the number of nearby devices and their recording channels with the icons seen in Figure 4.18. To enhance the comprehension of the first layer we customized the icon set based on the pre-evaluation.

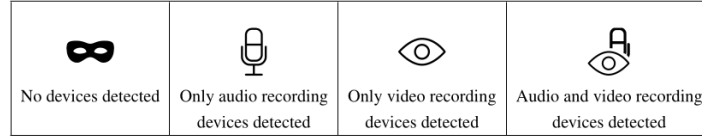


Figure 4.18: Changed Icons for the First Layer.

In addition to the simplification of privacy policy characteristics with color coded icons and textual description, we mocked-up an overall privacy scoring system (grades from A to F) for each IoT device type based on their privacy policies, inspired by TosDR¹⁶. The worst grade is then also displayed in the parent layers of the presentation (see Fig. 4.15 grade D). Thus, the first layer provides a quick insight into what information could currently be recorded, i.e. face or voice, as well as an abstract view of how this data is handled from a privacy perspective. At each layer, the user also has the option of tapping on the category grade to display the specific privacy properties of the worst device via a pop-up (Fig. 4.17, center).

To enable ex-post insights regarding information disclosure by surrounding IoT devices, the amount of found devices is recorded and presented in a history view (Fig. 4.17, right). Thus, users can reflect on their privacy state over the past days, which also may help to raise awareness about IoT-devices in everyday situations. Individual points can be tapped, to reveal the devices which were found at this specific time. The overall communication structure was implemented as shown in Figure 4.11. Hence, IoT devices are equipped with a BLE Beacon, which continuously broadcast advertising packets in an interval of 1s. For the implementation the AltBeacon protocol was used because of its open design and license, offering 20 bytes of payload [18]. As suggested by the AltBeacon authors, the first 16 bytes are used as an advertiser ID, which will be fixed for all beacons. The remaining 4 bytes are used to reference transparency information (3 byte DC ID, 1 byte device/policy ID). The smartphone app of the bystander scans continuously for new beacon signals, which are detected within 1s, just like the absence of them (out of range). The AltBeacon library¹⁷ offers the possibility to filter for certain beacon features, which we utilize to find beacons based on the chosen advertiser ID. If data is received, the payload of the advertisement package is extracted, included IDs are queried in a database, which returns the privacy information of the IoT device. We implemented this communication channel using separate Bluetooth dongles from Nordic Semiconductor.

4.2.4 Analysis of Bystander Perceptions regarding the Issue and the Audit TET

In the following, we present insights into how the bystander issue as well as our transparency solution is perceived. In particular, the goal was to investigate the following questions:

RQ 1: Is there a general desire for transparency regarding IoT devices in everyday life? Since related studies on concerns and perceptions of bystanders (see Sec. 4.2.1) mainly focus on participants from the United States, and privacy perceptions differ across cultures [322], the first question that arises is whether

¹⁶<https://tosdr.org>

¹⁷<https://github.com/AltBeacon/android-beacon-library-reference/>

these findings can be applied to participants of the German culture. Once it has been determined whether people are concerned regarding surrounding audiovisual smart devices and what they are concerned about, the question arises as to whether they need transparency to address the perceived concerns.

RQ 2: What does transparency concretely mean for people? In addition, we want to find out what people consider as a transparency solution. This part therefore refers to required transparency information and its priorities, as well as whether there is an understanding of given information.

RQ 3: Does our solution match the theoretical and practical desire for transparency? Finally, we wanted to know to what extent our implementation meets the requirements and needs of the participants regarding a transparency solution. As well as whether it would be generally usable.

4.2.4.1 Method

To investigate the bystander's view regarding the research questions, we conducted semi-structured interviews, using a guideline, supplemented by standardized questions. In this way, on one hand, consistent questioning of the participants and comparability of their answers was ensured. On the other hand, the interview's openness enabled sufficient flexibility to capture previously overlooked aspects. Moreover, the conversational nature facilitated an atmosphere of openness concerning the answers, mitigated the risk of misinterpretation of the questions, and supported a deeper understanding and exploration of the participants' answers [293]. In addition, interviews were supplemented with a think aloud usability testing session.

The environment for the interviews was an ordinary office. To simulate a real IoT environment and to investigate perception and detection towards IoT devices, a smart speaker (Amazon Echo Dot - clear visible) and a webcam (Google Dropcam - more hidden) were placed as surrounding IoT devices.

The interview guideline was pretested three times. As a result, the guide was enhanced with concrete scenarios instead of abstract terms (e.g. "IoT device in everyday life") to avoid misinterpretations by the respondent [180]. Additionally, suggestive questions were replaced with neutral ones. The last interview tested the course of the final guide.

Interviews were recorded acoustically and subsequently transcribed using the selective protocol transcription system. Finally, answers and reactions are compared and analyzed, using the inductive coding approach according to Mayring [303].

Participants The recruitment and corresponding interviews were conducted in early 2020. In order to obtain the highest possible applicability of the results we selected participants by quota sampling [293], based on characteristics of gender, age and previous technical knowledge. Overall a total of eight (German) participants were selected (Tab. 4.7).

4.2.4.2 Interview Guideline

The interview guideline was tailored to the research questions. Before the actual interview, the general topic was briefly introduced. To avoid priming, the introduction only mentioned that the interview would be about intelligent devices, such as smart speakers, which use voices or cameras to communicate with their users. In addition, a privacy statement and an informed consent form for recording the interview were given.

Table 4.7: Participant distribution.

younger	male	female
has technical knowledge	(M1) 21 / Student (computer science)	(F1) 24 / Student (computer science)
no technical knowledge	(M3) 25 / Student (educational science)	(F3) 18 / Student (teaching)
older		
has technical knowledge	(M2) 51 / Professor (electrical engineering)	(F2) 58 / System Administrator
no technical knowledge	(M4) 61 / Geologist	(F4) 62 / Interpreter

User's Attributes & Desire for Transparency Within the first part, we wanted to investigate general perceptions regarding audiovisual IoT devices. Participants were asked about their experiences; where the experiences were derived from, e.g., through ownership, friends or advertising; the reasons for (not) owning smart devices; and perceived (dis-) advantages concerning these devices. Further, we investigated participants' knowledge about the devices' general functionality. Either via a well-known device to the participant or by using the Amazon Echo Dot as example by asking for the weather forecast. Participants then drew their idea of how it arrives at an answer on a piece of paper. Subsequently they assessed if the imagined process also applied to other IoT devices. To reinforce the setting and a realistic feeling of (not) desired transparency, it was asked whether they have initially recognized the smart devices that were present in the room. In the follow-up to this intervention, participants assessed an 11-point Likert scale how important it would be for them to be able to recognize such devices in everyday life (0 - not important at all, 10 - very important; see Fig. 4.19 (a)). On the one hand, we thereby gained an unbiased impression of the participants' felt importance of transparency, free of communicative skills. On the other hand, it allowed to rank participants and compare them with each other.

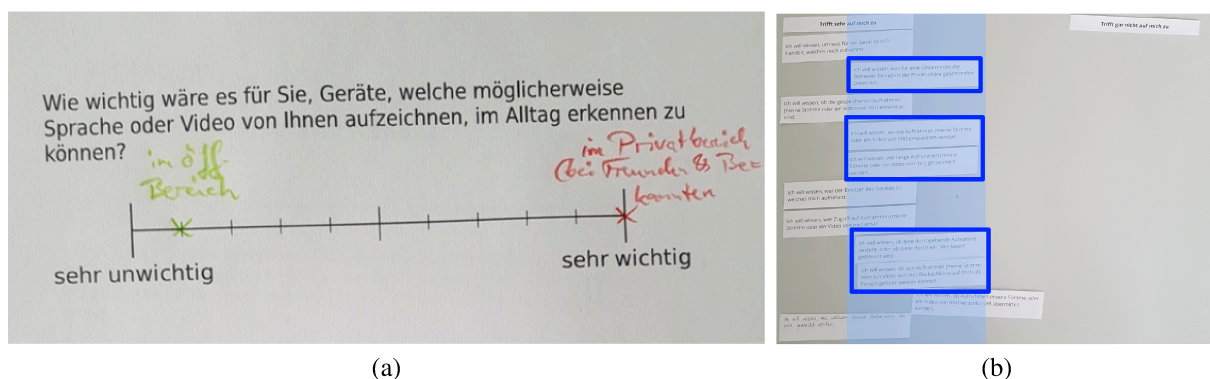


Figure 4.19: (a) Eleven-point Likert scale regarding importance of transparency.

(b) Rankings of different transparency information, written on strips of paper. "Rather applies to me" section - Score 4 - is shown in blue.

Meaning of Transparency The second part of the interview addressed the users' understanding of a transparency solution and their expectations of essential aspects. In addition to the open answers, two specific questions assessed the importance of the recording channel ("How important would it be for you to know about what is recorded from you by a device (e.g., voice or video)?") and of further

information regarding the processing of recordings ("How important is it for you to know what happens to the recordings?"). Both questions could be answered on a 11-point scale (0 - not important at all, 10 - very important).

To investigate which further details regarding recordings are important to participants, statements, written on strips of paper (e.g., "I would like to know how long the recordings (my voice or a video of me) are stored.", "I would like to know who has access to the recordings.", see Tab. 4.8) were ranked according to their relevance to the participants. To do this, they arranged the statements on a 5-point scale (1 - "Does not apply to me at all" and 5 - "Applies to me a lot"; see Fig. 4.19 (b)). In addition, they were encouraged to add their own statements on additional blank slips.

Thus, the general interest in the details of recordings could be compared between participants and associated with open answers.

Transparency App Usability The third part examined the reception and usability of the app. For this, the issue and the idea of our TET solution were explained to the participants. In addition, participants were provided with a smartphone (Motorola Moto X) and -watch (LG Watch R), both equipped with the transparency app. The app showed mocked-up details for both surrounding IoT devices and two smartphones (Samsung Galaxy S7, S4 mini). First, participants tested the smartphone by moving around within and beyond the room. When they left the room, devices disappeared on the app and were displayed when entering the room again. This reflected the way the app works in everyday life. Second, participants tested the app on the smartwatch analogously. If the participant had no experience with smartwatches, the functionality was explained briefly. During the test participants were asked to think aloud. To capture the answers and the using situation, audio recordings were supplemented by screen recordings and observations by the researcher during the experiment. Thus, the participants' preferences could be identified in relation to the presented transparency information in the app. We assessed app perception and usability concerning the following questions: How is the displayed information received? Is the app intuitively navigated? Are all functions recognized? What is the participant's impression of the app? And: How is the overall score and the associated popup received? Finally, participants should assess if they would use the app in their daily life and if they would recommend it to others. Thus, our solution could be evaluated, and parallels and discrepancies to the statements of the first parts could be drawn

4.2.4.3 Experience, Trust and Fear

Described **Advantages** of IoT devices were related to convenience, practicability, efficiency, remote control, multitasking, facilitation for the blind or visually impaired, and plain fun or curiosity. Although they mentioned many advantages, participants actually used their IoT device, if they used any, for limited functions.

M1 *"I use Siri quite often, but actually only for a few cases, setting timers, reminders."*

F1 *"I use it for weather queries or app control from a distance, for example when I don't have a hand free."*

Regarding **disadvantages**, some participants showed a general discomfort with being recorded. The main reason for this was the fear of targeted observation. Two participants (M2,F2) also drew parallels to the Stasi¹⁸ of the former GDR.

M1 *"My feeling is that in a way I don't want everything to be recorded where I am."*

¹⁸<https://en.wikipedia.org/wiki/Stasi>

M2 *"I simply feel uncomfortable with it, because there is also the potential ... because I am not completely unfamiliar with the telephone being listened in on."*

Moreover, many participants expressed specific **uncertainties** and **distrust**, which could be categorized as loss of control, misuse of recordings, and fear of consequences:

Loss of Control Some participants described uncertainty, due to unknown devices and especially due to concealed processing (what and how is recorded), resulting in a felt loss of control over their vicinity as well as over the recorded data. One participant thought that such devices can not be controlled at all (M3).

M4 *"That's why I always keep an eye on the risk that things can happen here that I can't even recognize or control. Maybe because I don't notice them at all."*

F3 *"I find that as soon as data like that leaves you, it's in a kind of cycle, and you can no longer control who has it anyway."*

M3 *"It makes me a bit uncomfortable when I have such a thing installed, because I've already read articles where employees have leaked that they can permanently plug into these devices, eavesdrop on conversations under the label of secret market survey, and that's a really intimate invasion of privacy, and that was a stronger reason not to get me such a thing."*

Misuse of Recordings Participants also showed mistrust concerning the handling of recordings, doubting that these are just used for necessary processes, but also for, e.g., hidden advertising or for sharing with political agents. Some were afraid that hackers may be intruding their smart devices, e.g., to find lucrative burglary targets. Others expressed uncertainties about foreign device owners.

M3 *"And I also know in any case that politics or something like that can somehow force companies to give out data, [...] by the fact that there can simply be a connection between the economy that says of itself: 'It's only product optimization, the whole data collection', nevertheless a lot of data is sold to politics, e.g. used for election campaigns."*

M4 *"I wouldn't want to have a smart home either, simply because of the danger of being hacked or spied on, and then they find out when you're not there, etc."*

F1 *"If [cameras are] hidden, there must be an ulterior motive of some kind on the part of the person who put them there."*

Fear of Consequences Finally, some expressed fear of consequences due to recordings of them. One participant, who otherwise showed higher confidence, expressed concerns if, for example, her employer had access to recordings (F3). Some expressed even a higher severity of consequences, e.g., permanent surveillance or manipulation of their opinion.

4.2.4.4 Previous Knowledge

Nearly all participants used the Amazon Echo Dot to explain how smart devices work. Participants with previous technical knowledge (Tab. 4.7), knew that recordings had to take place and that these are processed on external computers. Among the other participants, three showed basic understanding, but none of them were aware about external storage and processing of audiovisual recordings (Fig. 4.20). One respondent could not answer the question at all.

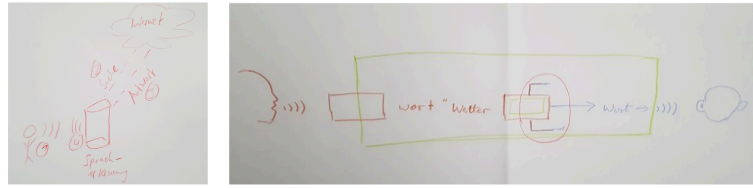


Figure 4.20: Sketches of the IoT device functionality of participants without technical background. This drawings show voice recognition directly on the device. Left: M3, Right: F4.

- F4 *"Sound to word, word to ... well, but that's all a computer [green circled]. [...] In any case, with the Internet must be ... no computer can have so much information, you have to search the internet, with these words [red circled]."*

4.2.4.5 Desire for Transparency

All participants predominantly desired to recognize IoT devices in everyday life, with ratings from 6 to 10 (Md 8.7) on the first 11-point scale. Participants with lower ratings generally trusted such recordings more. Whereas those with higher ratings expressed either more concerns and a higher severity of consequences, or a general rejection of recordings of their voice or video.

- F3 *"Well, I don't think there's necessarily bad stuff happening or anything."*

- F4 *"I want to detect them immediately. I want to know who's been watching me, that's what I think ... I don't want to be watched."*

However, a demand for transparency information was highly context-sensitive. Some participants distinguished between private, public and professional contexts. Desired transparency in the *private* context (e.g. conversations with friends, colleagues or even financial consulting) was related with higher intimacy and exchanged information. However, if there was a higher focus on the human factor, such as trust in friends and acquaintances, and the ignorance on technological data collection and processing, transparency in the private context was not important in the same extent (F2,F3).

- M2 *"[...] that always depends on the context. I don't want the possibility that if I have a private conversation in the cafeteria, that there is a potential there, [...]"*

- M4 *"In private situations [...] I want to be sure that nothing is recorded."*

- F3 *"I'd know that most of the time. I'll figure it out. I might be with people and I know them, they are usually trustworthy."*

In contrast, *public* spaces were less problematic due to the expectation of being observed anyway. Thus, transparency could here even be perceived as annoying. Except if people were afraid of public surveillance, loss of anonymity, or blurring boundaries of explicit and implicit information.

- M4 *"In the public sphere, I'm relatively uninterested. [...] As far as what everyone can see on the street, I don't care if a camera is allowed to look at it."*

- M2 *"[...] there I am anonymous in the masses, so to speak, and that's how it should stay."*

- M2 *"[...] while e.g. at an official event it is already clear to me anyway, it rather bothers me the other way round, that I am reminded of a situation that is to be expected anyway."*

- M4 *"Whether [the device] can interpret, e.g. the emotional state; state of anxiety, from the person of whom speech/video is made. I wouldn't want to have such interpretations."*

The *professional* context, was located between private and public in terms of transparent importance by M4. Especially action interests and premises were regarded as partial intimate information. Thus, a possible analysis of his state of mind would make him feel uneasy. In addition, one participant made a contextual distinction into specific recordings of her as an individual (very important) and recordings of her in the mass (less important) (F2).

4.2.4.6 Meaning of Transparency

Participants rated transparency about recordings in general, the specific recorded channels (i.e. audio or video) as well as the handling of these recordings as equally important. The recording channels itself were perceived ambivalent. Whereas some considered voice recordings as a higher risk to their privacy, due to a higher value of vocal information. Others viewed video recording as more problematic, due to the possibility of identification and an effect on their behavior. Only one participant expressed that the recording channel, compared to data handling and general recording, was not important at all since the recording itself would be the worrying thing (F4).

M1 *"If I [...] know that it only records video, then I would certainly behave differently than if I know that there is a device that listens to what I say all the time."*

M3 *"For video, you could probably identify the person more clearly."*

M4 *"I guess you behave differently when you are visually recorded."*

F4 *"But I think it's bad enough when there's already a device, no matter what, but it bothers me."*

However, participants expressed the need for more information depending on the data properties, such as whether the data is processed and stored locally; whether it is possible to draw conclusions about the identity of the recorded person; or on the acceptance for certain purposes.

F3 *"If this would be stored only locally, then I find some things even less important here, because then it simply - then it doesn't matter."*

M1 *"If it can't be traced back to me at all, I wouldn't care if it was recorded."*

F1 *"If it's for security reasons, I'm thinking 'okay, fine.' But if it's in a private room, then I would like to know why, for what, for how long."*

M1 *"Where it stays on the device, or is only used for certain purposes, like a Tesla, I can live with that. But now, for all devices that record speech and audio, I would like to know what happens to it."*

Concerning the importance of various transparency information, different needs and preferences occurred (Tab. 4.8).

Table 4.8: Ranking of statements by participants (M1,F1,M2,...), sorted by Median & Mode, on a scale from 1 (not important) to 5 (very important). Statements are shortened for legibility.

I want to know, ...	Statements	Md	Mo
how long recordings are stored	5,4,3,5,5,5,4,4	4.5	5
where the recordings are stored	5,4,3,5,4,5,5,4	4.5	5
whether recordings allow identification	4,4,5,5,5,4,3,5	4.5	5
whether stored recordings can be viewed	4,5,5,5,5,3,3,4	4.5	5
who has access to recordings	3,5,3,5,5,2,5,4	4.5	5
whether transmission is encrypted	3,3,3,5,4,4,4,4	4	4
if recording is controlled by a wake word	4,4,4,3,4,3,1,4	4	4
what is processors overall privacy grade	2,4,5,3,4,2,3,3	3	3
who is the device owner	1,5,3,1,3,1,5,4	3	1
what specific device records me	3,5,3,1,2,1,1,4	2.5	1

Participants were primarily interested in the data handling, independent of who owns the recording device or the type of device. The principal needs were to know how long and where the data is stored, as well as who can access the data.

F1 *"It's [...] good to know if it will stay in the database forever, like on Amazon, or only for a certain period of time."*

M3 *"It would just be cool to know, to see ... different countries do have different data protection laws."*

Also important, but not to the same extent, were the possible encryption of the data and whether it is recorded permanently or punctually by using a trigger. An overall privacy grade for the data processor was rated ambivalent, which could result from not being sure whether the grade originated from a trustworthy institution:

M3 *"The question is again who creates this mark. Well, that's cool in itself, if you knew you could somehow rely on it ..."*

4.2.4.7 Reception of the Transparency App

Within the scope of the practical test, the majority liked the concept and the implementation of the presented transparency solution. When asked for a potential use in daily life, 6 out of 8 participants would recommend and install the app. One noted that she would recommend the app only, if there is an unexpectedly large amount of surrounding devices in everyday life (F3). Accordingly to the theoretical desire, people were willing to use the transparency app preferably in the private and professional and less in the public sphere. If the app would be used in public, it would be mainly out of curiosity or simply boredom.

M4 *"I would of course use it privately, simply to check whether everything is as I imagine it to be in my house. And also in relation to neighbours, acquaintances. [...] I would also want to use it in the professional area. [...] especially in government buildings, to find out how well it is monitored, what do they record?"*

F1 *"I wouldn't use it hourly, daily, because it would make me paranoid. Just using it out of interest in how much is really recording me, how many devices are around."*

However, some would refrain from installing/frequently use the app because of the fear of becoming too paranoid; the already existing information overload (F1); a lack of perceived personal affectedness (F4); or own coping strategies (M2). The latter two could be based on the misjudged number and bandwidth of such devices, which was directly mentioned by M3.

F4 *"I would rather not install them. That doesn't mean that I find them useless, it's just that I don't really need it. Maybe it's just me, because my radius of movement is only at home, in the city, or at most a doctor's visit. Otherwise nothing, so it's not all that important to me."*

F1 *"'Oh God there are so many cameras recording me.' I think it might be freaking me out a little bit."*

M2 *"If I wasn't sure in any environment, I wouldn't have taken that approach. Instead, I would have said: Let's go somewhere else."*

M3 *"... so what I now found surprising was that it is such a bandwidth of devices that determines data at all."*

Most users attach importance to the app interfering as little as possible with their daily flow, which could be achieved by the passive notification regarding the number of surrounding devices and their recording channels. This was also positive mentioned by the participants (for both, smartwatch and smartphone) with regard to its use in everyday life. One exception are devices posing a particular threat to data protection, which should be actively notified. This feature is currently missing, however, individual desires and situations must be taken into account in its implementation.

F3 *"The fact that you only have the notification on top I think it's really cool. [...] And I think it's good that you're not so flooded with information at the beginning, but you can look relatively focused for what you want to know."*

M1 *"And it would be really cool to get a notification when a nasty device is near me: Hey, watch out, there's a class D device there right now"*

Interestingly, in comparison to the theoretical need for transparency stated by participants, their practical need was less pronounced. When asked about other notification options such as smart glasses or headphones, acoustic notifications were considered rather annoying, but some could imagine a visualization in glasses if the notifications were kept within limits.

Reception of provided Transparency Information The given transparency information was accepted well, with most participants being satisfied with our selection. In accordance with the theoretical desire, the most important aspect for participants was the notification about nearby recording devices. Although other information were before considered as equally important, F4 was explicitly indicating, that in practice, the notification about devices may be sufficient.

F4 *"[...], I don't need to know anything else. I see at most, here is one [smart device], and where maybe. As soon as I know that, then I am already careful."*

The majority had no difficulty understanding the presented textual information. Solely the description regarding the *wake-up word* was misunderstood by one (F4). Additional requested information were the exact location and/or the distance to recording devices and range of the app, the location of data processing, details of the distribution of recordings (who has access) and the opportunity to intervene in recordings. In addition the question arose as to the trustworthiness of the information presented.

M1 *"And of course there is the question of how much you can trust what is written there."*

The historic function was deemed as "good in principle, but uninteresting" (M3). One stated that the function would be more interesting with more features (M1).

M3 *"It's quite cool that you can see the history, although I don't know now whether this will bring you so much benefit, because you can't change what was there"*

M1 *"[...] for example, the app could learn which devices are located in which places and show them on a map."*

Given the limited space on the smartwatch, the amount of information displayed was mainly considered appropriate. However, some expressed the wish to see more information about the individual devices, similar to the pop-up on the smartphone. Interestingly, there were also concerns about the smartwatch itself.

M4 *"With smart watches, whether the data from the watch is transmitted somewhere? Your step counts, ... that's highly personal information"*

Usability and Comprehension The general structure of the app and the underlying hierarchical representation was perceived as intuitive; all participants quickly achieved an overview of the surrounding devices.

M3 *"I found it relatively easy to use, which is really cool because it is low-threshold and I don't have to search for all the different settings."*

M4 *"I have very little experience with smartphones, but that was understandable."*

However one described the UI as generally too “*boring*” (F3). Some icons (such as third party sharing, which was misinterpreted as an indicator for the smart device location) needed contextualizing text. Although the division of devices according to the recording channel was intuitively understood, the icon of combined audio and video recording was not immediately recognized (esp. on smartwatch). The abstraction of data handling to a privacy grade of a device (by using letters and the color code) was misunderstood by some. Misinterpretations ranged from reversed reading of the scale (F being the best, and A being the worst rating), to misunderstanding the letters as those of the device’s owner or name. After an explanation, the principle was perceived as positive and interesting. As a result, an explanation of the icons’ by a tutorial, and information about the privacy score’s calculations were identified as essential additions. The view to details of the devices was not always found quickly, since it was not intuitively clear that this could be opened by touching. The history function sometimes was not discovered (“too hidden”).

4.2.4.8 Discussion of the Interview Results

In this chapter we summarize the reactions with regard to our research questions.

RQ1 In accordance with previous research in other cultures [504, 333], German people predominantly desired to recognize IoT device in everyday life. They primarily wanted to know, if they were recorded due to perceived adverse effects, like feared targeted observation, loss of control, misuse of recordings, or abstract negative consequences. The highly context-sensitive interest in transparency [495, 333, 298, 316], was found to be a particular issue of the difference between the private, public and professional sphere. In addition, the need depended for some on the purpose of the recording [333], whether it is a direct recording or a recording in the crowd, and sometimes simply on pure interest or curiosity. We uncovered that the private and professional (although not in the same intensity) spaces were perceived to be especially vulnerable regarding intimate information. However, if there was a higher focus on the trust in on site people or less focus on technological data collection and processing, transparency in the private context was not important in the same extent [299]. In contrast, public spaces were less problematic due to the expectation of being observed anyway (e.g., through other people)—except if people were afraid of public surveillance, loss of anonymity, or blurring boundaries of explicit and implicit information. Accordingly, people were willing to use a detection app preferably in the private and professional and less in the public sphere. However, information overload appeared to be serious issues for not using a transparency solution. Hence, the impact of the amount and the timing of given transparency with respect to the specific contexts needs to be explored further. The justifications for the various needs were in part well-founded on the basis of actual technological developments and own needs. However, they also depended in part on a lack of awareness, for example, the underestimation of possible ambient devices or by the fact that for some the trust in the device owners (strangers vs. friends) overrode actual functionality.

The mentioned underestimation of the potential existence of recording devices and of the bandwidth as well as the feared paranoia if knowing about it, indicates that increased awareness through transparency should be accompanied by literacy regarding how to deal appropriately with the acquired knowledge.

RQ2 People rated transparency about recordings in general and the recorded channels as equally important. Although video recordings inhabit the most extensive data, for some people, voice recordings

meant a higher threat to their perceived privacy. The need for more information depended on the recording properties. The need was higher if data was not stored locally, if the recorded person could be identified, or if the purpose of recording was not considered normative. While the first issue relates to technical and design decisions of IoT devices, the second relates to the scope and handling of the data, and the third is located in the social area (namely the acceptance of certain recordings for certain purposes), emphasizing the need for privacy literacy.

If the need for transparency was given, it was most vital for people to know about the handling of their recorded data (duration and location of the storing, access of third parties, and who these parties are). Also important, but not to the same extent, were the possible encryption of the data and the duration or triggering of the recording. However, which device was involved or to whom this device belonged was rated as rather unimportant. Additional desired information related to the exact location of surrounding devices, the purpose and the possibility of intervention. The abstraction of data handling to a privacy grade was perceived as positive and interesting, with the note that the method of preparation must be transparent.

RQ3 The given transparency information was accepted well, with most participants being satisfied with our selection. The usability analysis revealed that the app is intuitively navigable, and given information is easily understood. Many participants commended an increased feeling of data protection and awareness about possible recording devices in everyday life. The desired context sensitivity, could be achieved through the passive notification, which allows the user to get information about his environment in contexts that are important to him, without being disturbed in other situations. However, active notifications in contexts that are important to the user are missing.

Interestingly, in comparison to the theoretical need for transparency stated by participants, their practical need was less pronounced. Although the majority would install the app, they indicated a limited use to specific scenarios such as the private sphere, and some even denied a practical need due to the assumption of less pronounced progress in the field of IoT. This emerging discrepancy between theoretical and practical interest in transparency may, on the one hand, been caused by the presented transparency solution in combination with the specific experimental situation. On the other hand, this discrepancy could be related to the privacy paradox [339]. Thus, the perceived risks influence the general intention to provide personal data, but neither the perceived risk to the own data, nor the actual behavior. That is, some recognize possible risks in general, but do not perceive these basically abstract risks as having any practical influence on them. This is supported by the difficulty of describing the value of one's own personal data, which is also shown in other work [5].

F4 *"But I mean, my information has no value actually, no monetary value."*

Limitations Although the apps were received well overall, an assessment of daily usage is only possible to a limited extent, as the usability analysis was carried out within half an hour in an office room [406]. Due to the qualitative approach of this pilot study and the small amount of participants, representativeness and distribution of the found issues and relations need to be verified in further studies. There is also the possibility that the participants indicated a higher interest regarding the solution within the experimental setting than they would in reality, e.g. through a novelty effect, a possible social-desirability bias or the focus of the study could have had a priming effect.

4.2.5 Concluding Remarks regarding Bystanders in the IoT

The ongoing realization of the IoT and the associated integration of tiny, networked sensors into any type of physical objects emphasizes new risks, esp. with regard to privacy. Thereby, not only device owners are affected, but ultimately anyone who enters the recording radius of such IoT devices.

Therefore, in this chapter we introduced our transparency concept for bystanders, to enable insights into surrounding, audiovisual IoT devices in everyday life. The concept is based on the regulative assumption that such devices need to identify themselves. They are equipped with BLE beacons and continuously broadcast information about recordings and their privacy properties. A digital mobile device of the bystander detects these signals and visualizes the current privacy state regarding surrounding smart devices. This concept was prototypically implemented for smart phone and -watch.

In addition, we conducted a semi-structured interview to analyze how the bystander issue is perceived by participants of the German culture, their desire for transparency as well as whether our transparency approach is considered a sensible solution at all. We could confirm results of previous studies, which stated a high hypothetical demand for transparency regarding surrounding IoT devices in everyday life. The in-depth interviews uncovered that this desire related primarily to a feared loss of control, misuse of recorded data and abstract negative consequences. Hence, the data handling was more important than which device is recording or who owns it. However, the desire for transparency was contextual and almost limited to private or intimate situations, whereas public spaces were seen as less problematic. It was less pronounced if the recordings were stored locally, were assumed to not allow the individuals identification, or their purpose was socially accepted (e.g. security). Nonetheless, we could find the privacy paradox, resulting in our study from an underestimation of the data's value and of the societal IoT penetration. Overall, the prototype was received well, and included transparency information were felt largely sufficient. Most of the participants stated that they would install and use the app in their everyday life. However, they preferred to use it primarily for the private sphere, which was enhanced by the desire to locate and block recording devices. Based on the results, we have taken initial steps for improving the transparency solution and identified further open challenges in Appendix II-C.

5

Digital Information Consumption and its Risks for Democratic Societies

In addition to its impact on privacy, the advance of digitization has also drastically changed the way we consume information. This is due, in particular, an increased speed of information flow, the enormous amount of information, and the democratization of its content. Thereby, online social networks (OSNs) have established themselves as tools for keeping track of, filtering, and making sense of the flood of information. In recent years, large audiences have embraced OSNs, such as Facebook, Instagram, Youtube, or Twitter, as their primary channel for information exchange. These networks have almost ubiquitous reach. The information circulating in these networks is manifold and comes from various sources. In particular, news providers are making great efforts to publish and disseminate their articles on multiple social platforms to reach a wider audience [131]. Moreover, politicians have adapted to the digital environment by utilizing social media for campaigning and connecting with their target audience [423]. As a result, an increasing number of citizens consume their daily news and political information directly on these platforms [336, 337, 401]. The availability of social-media mobile apps amplifies this effect and increases exposure in various everyday situations.

The positive aspects, such as the constant and immediate availability of information and its free, democratized distribution, are also accompanied by new challenges. Consumers had to learn and judge some, to a few dozen newspapers, TV stations, and other media channels in the past. The plethora of news sites, blogs, and video platforms that have emerged in the meantime complicate this endeavour [483]. The representation of external information in OSNs as short article previews allows users to get a quick overview, but also leads them to feel better informed about the content being covered than they actually are [21]. Moreover, this content is far from consisting solely of investigative, researched information. Additional actors have emerged. Next to news, satire, commentary, or opinions, some are distributing misinformation, conspiracy theories, and propaganda with agendas ranging from the commercialization of click-bait, over political influence, to establishing opinion platforms as hidden distribution channels for marketing of all types of products [499]. Research underlines that social media users are more exposed to populism, propagated by political actors from both extreme ends of the political spectrum, than individuals without social media [132]. A balanced news selection has to give way to a choice of posts and topics reinforced by the user's chosen neighborhood in this sheer mass of information. Individuals who put

more trust in information shared by friends, likely regress to consume news from narrow contexts [49]. Moreover, visiting social media rather for entertainment and distraction than information acquisition, the users consume the presented information rather incidental than targeted [49].

Overall, this development increases the difficulty of evaluating the credibility of information [483, 437], promotes political polarization and ideological division [29, 53], reduces political education [23] and, therefore, arguably represents a risk for democratic societies. This is reflected, e.g., in the dissemination and impact of false and misleading information during democratic elections [15, 189] or in crisis situations like the Covid-19 pandemic [227].

In the first part of this chapter, we therefore want to understand the type of content circulating in the German-speaking Twitter Community and measure the extent and impact of (anti-democratic) news-related content, which contributes to the forming of political opinions. The basis of our approach is the data acquisition and enrichment strategy (i.e., obtaining an automatically labeled, virtually complete data set). Overall, we report on the media-consuming behavior of the German Twitter population and investigate the prevalence and impact of well-researched as opposed to polarizing, or populist content as well as influential content provider. We base our analyses on a representative Twitter snapshot (77M Tweets, 6.9M users), collected during the two months surrounding the 2019 European Parliament election.

In the second part of this chapter, to mitigate negative impacts of changed news consumption, we investigate effects of contextualization: augmenting news article previews in social media with additional information, to support the reader in assessing the credibility of the content. Thus, create transparency for information consumption. Hence, we propose a 5 point rating-scale, called Nunti-Score, which represents the information quality of the corresponding news article. This rating is based on automatically extracted background information. Testing a mock-up with 455 participants, we investigate the short-term impact on the perceived credibility of news feed items (or: how accurate users can evaluate the credibility of its content, with which certainty), recall at a later time, and overall to which extent the users accept support by such an automatic rating.

5.1 News Consumption within the German-Speaking Twitter Community

If you don't read the newspaper, you are uninformed, if you read the newspaper, you are misinformed.

Mark Twain (there is no evidence Twain ever said anything of the sort¹)

To investigate the impacts of digital information consumption, we want to shed light on the media consumption within Twitter. Although this platform changed its name to X in mid-2023, we will use the traditional name Twitter in the following. Due to data collection limitations, we have to strike a trade-off between sample size and data quality. We base our studies on a concise, well-defined, and virtually complete Twitter community, concentrating on the German-speaking Twitter Community (GTC). Selecting tweets by language allows for a detailed observation of such a specified population. Often,

¹<https://marktwainstudies.com/the-apocryphal-twain-if-you-dont-read-the-newspaper-youre-uninformed-if-you-do-youre-misinformed>

culturally and geographically diverse groups speak the same language. The GTC, however, represents a large, geographically well-defined population of around 7 million active users. The majority are from Germany, Austria, and adjacent parts of neighboring countries, and they all share a relatively homogeneous political landscape and corresponding media outlets

We want to investigate if and how users of this well-defined sub-group utilize the platform to consume media sources that contribute to the forming of political opinions. We want to find out which types of information they consume and distribute within the network as well as the impact of news-related content. Further, we want to study the reach and impact of anti-democratic content within this group as well as its supporters. Thus, we define controversial and non-controversial content. Controversial content combines articles from providers that contribute to misinformation, conspiracy theories, political propaganda, and similar democracy decomposing elements.

First, we propose a reproducible method for uniform data collection of tweets published in specified target languages. In addition we propose an automated data augmentation strategy to facilitate data enrichment on large, real-world data sets. We leverage shared external content, hashtags, and its categories to get a high-level understanding of discussions in an automated manner.

The content of this chapter was developed in very close collaboration with Jan Reubold and the results were jointly published in the journal *Online Social Networks and Media* under the title 'Dissecting chirping patterns of invasive Tweeter flocks in the German Twitter forest'. This publication was also used in Jan Reubold's PHD thesis. While he focused mainly on the community detection part and its analysis (understand user behavior and interactions, analyze existence and spread of echo chambers), this part is completely omitted here. Here, we focus on the analysis of information distribution within Twitter of the GTC and its categorization. Thereby, we get a (i) high-level understanding of what is shared/discussed based on the automated categorization of content, a (ii) measure on the share of news related discussions within the network, and (iii) can identify influential actors, (iv) measure the presence of established news providers within the network, (v) analyze user engagement w.r.t. different types of news, and (vi) measure the influence of controversial content.

Overall, we report on the media-consuming behavior of the German Twitter population and investigate the prevalence and impact of well-researched as opposed to polarizing, or populist content as well as influential content provider. We base our analyses on a representative Twitter snapshot (77M Tweets, 6.9M users), collected during the two months surrounding the 2019 European Parliament election.

5.1.1 Preliminaries and Related Work

Online social networks (OSNs) offer researchers the opportunity to investigate human interaction on a large-scale [271]. Based on such rich data, we address various topics of technical and theoretical nature. In the following, we provide a brief overview of current research and preliminaries on topics related to our approach.

5.1.1.1 Acquisition of Twitter Data

The goal was to capture a virtually complete snapshot of the German Twitter traffic. From a data mining perspective, a complete data set on the usage of an OSN is comprised of all user-generated data within a specific time frame.

In the past, the academic community leveraged methods towards collaborative data collection [115]. However, according to Twitter's policies, the public sharing of its particular contents is prohibited [471]. Therefore, researchers started to develop customized data crawling techniques that fit their particular research scope. A reliable data collection process should also be transparent and reproducible for evaluation through future research.

In 2010, Kwak and others [264] crawled the entire Twitter platform. Utilizing 20 machines operating with different IPs that crawled tweets via the Twitter Search API over several weeks to bypass Twitter's rate-limit, they obtained 41.7 million user profiles, 1.47 billion social relations, 4 262 trending topics, and 106 million tweets. Following the growth of Twitter over recent years, this approach is prohibitively costly and time-intensive, and it can be considered infeasible to collect a complete data set.

Analysing a specific subgroup of Twitter users (German users in our example) in a complete data set requires thorough pre-processing, as there are 500+ million new tweets generated every day. Recently published studies avoid this overhead by using Twitter's Streaming API, which allows researchers to obtain a limited number of real-time tweets that match a specific word-filter. Accordingly, the size of the acquired data set depends on the prevalence of the determined search terms (e.g. event-related hashtags) [175, 36, 25]. A downside of the Streaming API is that Twitter restricts the total number of tweets that can be crawled per day to 1% of all data. If the number of tweets matching a word-filter exceeds the limit, the stream will return a random sample of all matching tweets. For research purposes, this is an undesired outcome, as studies have revealed that the Streaming API provides a non-representative sample tweets [325]. Furthermore, it is not sufficient to fix the sample discrepancy by using multiple machines to combine simultaneous samples from the Streaming API [229]. Scheffler captures a representative snapshot of the German Twitter traffic despite the limitations [385]. She configured Twitter's Streaming API based on an exclusively German word-filter list. Since the number of German tweets within the whole Twitter-sphere is considerably small, the number of captured tweets only slightly exceeded the 1% limit. Thereby, she minimized the effects of Twitter's downsampling. By collecting every tweet that matched at least one word from the word list, Scheffler also collected a great number of tweets that were not German. She used a language detection algorithm to filter for German tweets in consequence. However, due to insufficient labeled data, the algorithm had to be evaluated manually on a small subset of the captured tweets. Regarding the effects of Twitter's downsampling, Scheffler concluded that these were negligible, as they accounted for under 3% of missing data. Nonetheless, it must be noted that the data collection process was conducted in 2013. Therefore, a future-proof data collection method should consider the possibility of a rising number of German tweets.

5.1.1.2 Analysis of Media Consumption

A large body of work analyses Twitter as a novel news medium. Recent work reports how news providers are utilizing social media to increase their reach [131, 130]. They show that the providers permanently adapt to changing online demands, and Lasorsa et al. [270] suggests that the journalists embrace social media, stating their personal opinions more freely. Other studies have investigated information consumption on Twitter. Malik and Pfeffer analyzed the corresponding actors within a random sample of 1.8 Bn tweets from 2014 [297], and Cinelli et al. analyzed almost 400 000 tweets posted by 863 accounts during the 2019 European Parliament election [82]. Neudert et al. analyzed a random sample of almost 1m tweets generated by 149 573 users in 2017, taking junk news and bots into consideration [335]. To the best of

our knowledge, there exists no study that analyzes the entirety of tweets of a concise population defined by a political eco-system with its corresponding media landscape. Scheffler reports that the German twitter community exposed some interactivity and was reluctant to share their geographic location [385]. However, they do not investigate political content and the corresponding engagement.

5.1.1.3 Political Orientation

Studies on the matter of OSNs rely on large and rich data sets. Depending on the objectives, data often has to be augmented with further information. Our studies rely on information about the political interests of users. In this context, [88] worked on the complete Twitter-sphere [264] to investigate the political homophily of Republicans and Democrats across the entire network. Using linguistic features extracted from annotated tweets and news texts, they utilized a supervised classification approach. While a common approach in the area of user classification on Twitter [349], research showed that the prediction of political affiliation is not reliable in multi-class scenarios, e.g., in the context of the broader political spectrum of German parties [86]. Additionally, textual features of tweets are not stable over time.

Therefore, an algorithm inferring user characteristics and interest from context-specific activities is more promising for the German Twitter user base. In this context, several attempts rely on Wikipedia articles to infer the interests of users [177, 57, 113]. Wikipedia and its broad range of categorized articles, including people, events, and locations, can be utilized to build a reliable knowledge database. Faralli et al. [150] approximated user interests by finding “*friends*” they could link to Wikipedia articles. For example, if a user followed a famous basketball player, her interests included sports and basketball. The researchers proposed a hierarchical representation of user interests and conducted a large-scale homophily analysis on Twitter. Their methodology offered a compact, tunable and readable way to examine user interests.

For a more thorough understanding of user interests, Himelboim et al. [207] leveraged frequently shared hyperlinks, user mentions, and hashtags and, thereby, analyzed users based on domain-related interests and hashtags. We deploy a similar approach for inferring user attributes.

5.1.1.4 Promotional Profiles

Besides manually controlled accounts, there also exist orchestrated and automated ones. Several guidelines recommend creating social media profiles for improved public relations and dissemination. To increase the distribution of news content in social media, Orellana-Rodriguez et al. [347] propose best practices. They suggest creating employee accounts to promote their corresponding tweets. Such accounts should contain a statement about their affiliations. News providers establish Twitter profiles to further the distribution of their articles [131]. News agencies, such as Reuters or AFP, instruct journalists using their accounts for work to include a disclaimer. The disclaimer identifies them as employees of a specific news agency [372, 166, 362]. It should also include a declaration that they speak for themselves and not their employers.

5.1.1.5 Controversial Users and Behavior

Related work reported on political echo chambers from the extreme ends of the political spectrum [97, 52, 155]. A common assumption regarding users within these chambers is that they only inform themselves based on a small and narrow set of information sources. McPherson et al. [306] reported this biased information consumption in social networks, called selective exposure. Boutyline and Willer [53]

observed that conservative and politically more extreme individuals showed a more pronounced tendency to form segregated user groups than liberals. While Barbera et al. [29] report similar results consistent with psychological theory and research bearing on ideological differences in epistemic, existential, and relational motivation, they conclude that previous work may have overestimated the degree of ideological segregation in online social networks.

Bor and Petersen [50] examined the question of why online discussions seem more hostile than their offline counterparts. They examined eight studies using cross-national surveys and behavioral studies and concluded that it is not that people are more hostile online, but that hostile people gain greater visibility online. Additionally, other studies report that emotion triggering posts [54], especially posts about political opponents are substantially more likely to be shared [366]. Combined, these effect seems to be amplified by the fact that moderate users turn away from discussions because of this hostile behavior [204]. This inevitably leads to the behavior of the few receiving a disproportionate amount of attention. In the U.S., this seems to be compounded by the fact that the most extreme left- and right winged political groups not only attack users with opposing views, but are particularly hostile to moderates who espouse their beliefs [204, 173, 327]. “Those who express sympathy for the views of opposing groups may experience backlash from their own cohort.” ([204]) This behavior undermines discussion between people with different opinions and even causes social media to have a detrimental effect on democratic societies [287].

5.1.2 Acquisition and Enrichment of the Twitter Data Set

We want to shed light on the news consumption of German-speaking Twitter users. The basis of our approach is the data acquisition strategy, i.e., obtaining an automatically labeled, virtually complete data set. We assume that measurements of the shared external content allow us to approximate statistics on news consumption. The classification into categories allows for an automated high-level understanding of its content. Additionally, hashtags (related to shared external content) provide further semantic understanding. The approach avoids biases due to inaccuracy during the pre-processing. An example here is utilizing NLP techniques for semantic understanding.

In the following, we introduce the various parts of the data engineering process. Therefore, we summarize Twitter functionalities before presenting our data collection strategy and explaining the automated data enrichment.

5.1.2.1 Twitter OSN and Functionalities

Twitter offers its users different types of *Tweet-Objects* to generate content on the platform. *Original tweets* are the standard way of posting. As of 2019, a user can write a message to his timeline, also known as a status update. The timeline of a user represents a roster of posts to record activities and make them visible to followers. Furthermore, the timeline displays activities of followed others (e.g., news providers, celebrities, and friends), to whom the user has subscribed. Therefore, the system provides a news-feed-like overview tailored to the user’s choice.

Retweets represent another type of post, which allows a user to copy a tweet from another user to his timeline. Therefore, it is visible to his respective followers and visitors. Users can also *quote* other tweets (except retweets). Thereby, they can re-post a user’s message with a comment of their own. Lastly, there are *replies* to comment on any given tweet, except retweets.

Besides textual content, such a *Tweet-Object* can also contain multimedia content (*images, videos, animated GIFs*), interactive content (*hashtags, user mentions*), places (*geolocation*), or links (*URLs* linking to external sources, which commonly are visualized as *Twitter Cards*). In addition to manually embedded user mentions (@username), Twitter automatically adds *mentions* in front of content that implies an interaction between users (retweets, replies, and quotes). Further, every Tweet-Object has an attribute (*source*) that describes the service used to post the tweet. Besides official Twitter clients, there are also third-party services. These services allow accounts to post tweets in an automated manner.

User-objects provide a variety of meta-data. It contains multiple free-text fields (e.g., name, description, URL), statistics about the social links of a user (e.g., follower-, and friend count), and statistics about her activities (e.g., favorites-, and tweet count).

Users can interact with others via direct User Mentions within a tweet or indirectly via connected tweet types, such as retweets, quotes, and replies. Compared to static follows, interactions allow capturing relationship dynamics over time.

5.1.2.2 Data Acquisition

Our goal was to capture a virtually complete snapshot of the German Twitter traffic. Therefore, we propose a comprehensive data collection scheme, i.e. an extension of Scheffler’s approach [385] (cmp. Fig. 5.1).

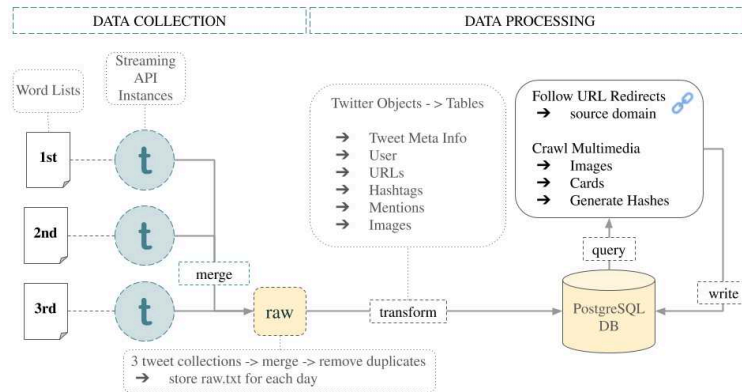


Figure 5.1: Data collection pipeline with three parallel Twitter Streaming API instances; each with a separate stop word list, including 400 frequently used terms in the German language; output of streams is merged; duplicated entries are dropped; raw Twitter-Objects are extracted from the files and parsed into a PostgreSQL database.

Our evaluation of different collection methods confirmed Scheffler’s findings. Geolocation-based filters only capture tiny amounts of German tweets. We hence decided to utilize word lists for our purpose. In contrast to Scheffler, we do not collect-then-filter to remove tweets in other languages, but we leverage the built-in language identification of Twitter. We thus created word filters, encompassing the 1200 most frequent German words. We base our choice on multiple text corpora, provided by the Leipzig Corpora Collection [184] and one corpus of frequently used words from OpenSubtitles.org². The latter encompasses terms that are more prevalent in informal conversations.

Twitter enforces a maximum of 400 keywords per instance, so we divided our word filter into 3 different lists and used three individual, parallel data streams. All streams obtained a high number of tweets from

²<https://github.com/hermitdave/FrequencyWords/>

600k to 1.2M on average. Thus our approach does not exceed the rate limitations of 1% ($\approx 5M$ tweets). We drop duplicated entries and merge the stream outputs. Findings in [326] suggest that German tweets are sufficient to capture political debates of the German-speaking population as non-German Tweets are ignored by the community. So, relying on Twitter’s language detector, we exclusively capture German tweets. Therefore, we sidestep Twitter’s rate limitations and, thereby, avoid down-sampling. While the detector lacks thorough documentation, research showed that, in some cases, it outperforms established alternatives such as Google’s Compact Language Detector [352].

We enrich recorded tweets with additional data. Besides the attributes, we further extracted child objects (original tweets, replies, quotes) from collected Tweet-Objects. The latter may entail collecting additional (non-German) Tweet-/ User-Objects. We argue that we need to include users who do not tweet in German but interact with German tweets. Further, we developed an algorithm that resolves shortened URLs to reveal their source domains.

5.1.2.3 Data Enrichment

After acquisition, we have to understand the content to analyze news consumption on Twitter. In the following, we propose an automated, sophisticated, and comprehensive data enrichment strategy evolving around shared external content (e.g. Fig. 5.2).

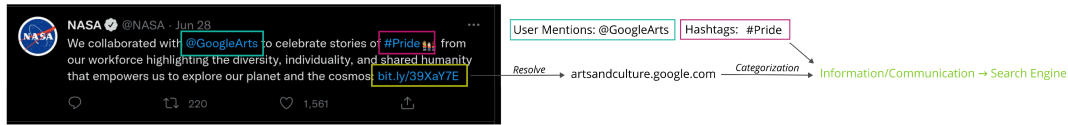


Figure 5.2: Identification of meta-information of tweets.

We focus on embedded news: shared external links presented as a preview within the social media platforms (for instance, Twitter cards with a headline, thumbnail, and summary on Twitter). Our analysis terminally requires to extract the category and type from the shared tweets as well as additional meta-information, which we perform in the following ways:

Categorization of Domains: Functional Groups (FGs) To obtain a comprehensive understanding of the external content shared by the German Twitter users, we categorize domains leveraging McAfee’s TrustedSource³ (2019). We tested different categorization services, and McAfee’s TrustedSource successfully identified the highest number of domains. Further, it provides a fine-grained set of 100 hierarchical categories (e.g., News, Lifestyle, Political Opinions, or Spam). McAfee also provides semantic subsets that split the categories in 12 so-called Functional Groups (FGs). Using TrustedSource, we categorized 98.3% of the URL-tweets in our data set. In the remainder, we sort URLs based on their domains into **FGs** and its related **categories (FG→category)**.

The News Group To identify all domains that influence the forming of political opinions, we manually investigated the most-shared websites from every category in our data set. Based on this research, the following set of domain categories, distinguished by the objectivity of reports (from **moderate-**, over **tendentious-** to **extreme** views), comprises the **News Group**:

³<https://trustedsource.org>

- **Information/Communication → General News:** Domains that generate daily news, political opinion sections, and educational content. (Top 5: *spiegel.de*, *welt.de*, *bild.de*, *sueddeutsche.de*, and *zeit.de*)
- **Society/Education/Religion → Education/Reference:** Web pages that relate to educational content, for example, classic literature, history, art, and other academic-related content. (Top 5: *de.wikipedia.org*, *spektrum.de*, *fridaysforfuture.de*, *kurierdeswissens.de*, *danisch.de*)
- **Society/Edu./Religion → Non-Profit/Advocacy/NGO:** Web pages run by charities and or educational groups or campaigns. (Top 5: *change.org*, *correctiv.org*, *peta.de*, *deutschland-kurier.org*, *mimikama.at*)
- **Society/Education/Religion → Government/Military:** Web pages provided by governmental or military organizations, including national branches as well as supranational entities, such as the United Nations or the European Union. (Top 5: *bundestag.de*, *polizei.bayern.de*, *auswaertiges-amt.de*, *bundeswahlleiter.de*)
- **Society/Education/Religion → Major Global Religions:** Web pages that provide information about major religions (e.g., Buddhism, Chinese Traditional, Christianity, Hinduism, Islam, Judaism, etc.) and include discussions and non-controversial commentary. (Top 5: *katholisch.de*, *catholicnewsagency.com*, *kath.net*, *vaticannews.va*, *evangelisch.de*)
- **Society/Edu./Religion → Politics/Opinion:** Web pages that cover political parties and opinions on various topics such as political debates. (Top 5: *tichyseinblick.com*, *jungefreiheit.de*, *achgut.com*, *politikstube.com*, *volksverpetzer.de*)
- **Lifestyle → Controversial Opinions:** Web pages that share extreme opinions, which are offensive to political or social sensibilities. Examples include xenophobic, fundamentalist viewpoints, and disinformation campaigns. (Top 5: *journalistenwatch.com*, *pi-news.net*, *philosophia-perennis.com*, *anonymousnews.ru*, *der-dritte-weg.info*)
- **Risk/Fraud/Crime → Discrimination:** Web pages that provide content that explicitly encourages the oppression or discrimination of a specific group of individuals. There are only a few domains that McAfee classifies as discrimination and only a few found in our data. (Top 5: *metapedia.org*, *theeuropoibe.org*, *renegadetribune.com*, *vanguardnewsnetwork.com*, *nordfront.se*)
- **Risk/Fraud/Crime → Historical Revisionism:** Web pages that spread misinformation, or offer divergent interpretations of, significant historical facts (e.g., Holocaust denial). (Top 5: *renegadetribune.com*, *vho.org*, *altright.com*, *dailystormer.name*, *johndenugent.com*, *codoh.com*)

Accessing OSN Links In addition to external sources referring to news content, we want to gain insight into content included in links to posts on other social platforms. Thus, we developed web crawlers for *YouTube*, *Facebook*, and *Instagram*, the most shared platforms in our data set. By utilizing the *YouTube Data API v3* and the *HTML* and *JavaScript* sources from Facebook and Instagram, we identify corresponding external profiles and their influence on news distribution on Twitter.

Political Hashtags To investigate user discussions about shared news content, we also consider the hashtags they contain. We automatically categorize the corresponding tweets by leveraging the co-occurrence of hashtags with URLs, as classified above. For example, if the hashtag #CDU appears in a tweet that also shares an article from *Spiegel*, we assign the #CDU hashtag to the category General News. This approach allowed us to assign categories to 60% of the hashtags in our data set. Note, however, a hashtag is assigned to several categories depending on its usage w.r.t. URL-tweets.

User Engagement Besides understanding content, we also want to study its distribution and impact. To measure how *users engage* with news or external political content, we define *reaction-tweets* in addition to simple tweeting and retweeting. Reaction-tweets contain direct responses (replies and quotes) and their retweets. We attribute them to the original tweets they are referencing. To measure content popularity, we leverage related reactions.

Promotion Profiles and Automated Accounts We also identify promotional profiles to measure their impact on the distribution of news. We base our automated detection of self-promotional profiles on the guidelines of news agencies such as Reuters or AFP (see Sec. 5.1.1.4). This process yields two types of promotional profiles: (i) journalists and (ii) feeds (Tab. 5.1).

Table 5.1: Sample of self-promotional Twitter-profiles from Spiegel.

Screen Name	@SPIEGEL_Politik	@joleffers
Name	SPIEGEL ONLINE Politik	Jochen Leffers
Description	Hier twittet das Politik-Ressort von @SPIEGELONLINE. Datenschutz: http://spon.de/afemu	ist bei SPIEGEL ONLINE im einestages- Ressort, twittet hier aber-so-was-von-privat
URL	spiegel.de	spiegel.de
Journalist	✗	✓
Feed	✓	✗

We identify a journalist’s profile by checking if it stated a news source in the free-text *URL-field* (e.g. *spiegel.de*) as well as the respective news domain in the *description text* (e.g. *Spiegel*) of their profile. Feeds, we identify using the above and check if their screen name contains the respective news domain (e.g., @spiegelonline). These feeds often act as the official publisher of articles. They generate automated content and rarely interact with other users. Websites often create multiple feeds solely to disseminate their articles.

This approach has obvious limitations. We cannot automatically detect promotion profiles that do not follow the journalistic guidelines. Therefore, we conducted a manual search for additional promotion profiles for the 30 top content providers. As it did not yield any additional profiles, we are confident that our findings below are representative in this regard.

Besides official automated profiles, malicious bots exist. With regard to these bots, we pursue a different route. In general, bot detection is an unsolved problem. For this reason, scientists resort to heuristics. Often, suspended accounts are interpreted as bots. However, a recent study [295] reports that less than 1% of the suspended accounts were suspected or potential bots. In line with other research, they found that suspended accounts pursued specific polarizing political agendas. Another approach to identify bots is to use tools such as the BotOrNot service. While often used by scientists, research shows how limited this approach is [122, 367]. With Twitter adding that binary judgments have real potential to poison our public discourse⁴. Based on this evidence, we argue that using these heuristics to exclude bots from our study provides no guaranteed benefits while seemingly introducing significant amounts of noise.

User Interests So far, our data enrichment strategy allows us to understand the content and distribution of tweets. However, we also want to gain insights into the political attitude of users. Prior work [150] identified user interests based on language processing and augmented this information into the friendship graph. This approach yields a more static assignment and relies on potentially error-prone text extraction. We aim to capture the dynamics of interest more accurately. Therefore, we identify it according to the hyperlinks the users interact with and share to avoid language processing and ambiguities. Using our approach, we leverage the categories of shared domains and hashtags. Briefly, we consider a user who regularly shares or replies to a specific news domain interested in related topics.

⁴https://blog.twitter.com/en_us/topics/company/2020/bot-or-not

Controversial Users The majority of studies classify users based on a political spectrum. Expressing opposing views in the political landscape of the U.S., researchers often label users as either Democrats or Republicans. Since the U.S. has a virtually two-party system, this is a justified and sensible approach. The political landscape in German-speaking countries, however, is more diverse. The political agendas of parties, e.g., tend to overlap. Also, deducing opinions based solely on hashtag information does not distinguish between support and opposition. Therefore, we do not rely on party references in tweets for estimating political affiliation.

The ‘Hidden Tribes’ study [204] took a more nuanced approach to analyze America’s political landscape. Surveying 8000 Americans, they identified seven groups based on shared beliefs and behaviors. Interestingly, the groups furthest to the right and left of the political spectrum were similar in surprising ways (e.g., color and wealth) and, most importantly, these two groups are the driving force behind the widening of the gulf between the two political factions. Therefore, we distinguish between moderate and extreme users, labeled as non-controversial and controversial.

Based on McAfee’s TrustedSource database, our domain categorization approach identifies domains that produce extreme political content and misinformation. While the category *Politics/Opinion* already contains domains with extreme and inflammatory content, categorizing users as controversial based on a shared article of these domains would lead to imprecise labels. Hence, we only include domains with extreme political views that, e.g., deny the Holocaust or encourage the oppression or discrimination of specific groups.

In this context, we assume that retweeting indicates an interest in a topic or even agreement with the sentiment of a message [309]. Therefore, after investigating all of the domains, we posit that people, who support these contents by sharing them in the network and contributing to its distribution, are likely to hold extreme political views. The categorization in our database classify these domains as **Controversial Opinions**, **Discrimination**, and **Historical Revisionism**. We define a group of **Controversial Users** comprised of users that shared at least one of these URLs. Accordingly, we specify users who share non-controversial content as **Non-Controversial Users**. While we cannot deduce their political affiliations, we assume they manifest less extreme views.

5.1.3 The German-Speaking Twitter Community (GTC)

To obtain a representative sample of the Twitter-sphere of the German-speaking user base, we collected tweets between the 2nd of April and the 2nd of June 2019. Thus, we obtained a virtually complete snapshot of the GTC, collected during 2 months surrounding the European Parliament Election in 2019 (26th May 2019). The sample contains 77 million tweets and 6.9 million user profiles users with 18.3 million shared external sources (see Tab. 5.2).

5.1.3.1 Tweet Types

Categorizing these Tweet-Objects by tweet type (i.e., original tweet, retweet, reply, quote) revealed that the most frequent action was retweeting. The majority of activity in our sample was reactive. Retweets account for 38% of all tweets in our corpus and are used to distribute content from other users, Replies for 31%, and original tweets, creating novel content or initiating conversations, account for only 27%. Quotes are rarely used at all (3.7% of the sample). Interestingly, we observed fewer users in our sample using

Table 5.2: Twitter-Objects captured during the data collection process.

Object-Type	Count	Tweets (%)	Users (%)
Tweet	77 390 122	-	-
User	6 919 206	-	-
Mention	85 155 158	72	80
URL	18 358 074	23	25
Hashtag	39 197 019	22	29
Multimedia	19 702 261	19	56
Place	1 189 696	2	2

replies (23%) than retweets (64%).

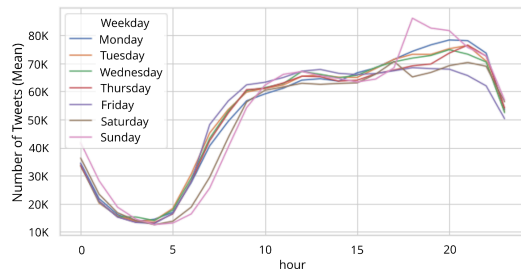
Besides investigating tweet types, we also analyzed their interactions. Table 5.3 shows that most often original content was retweeted (66.8%), followed by replies (21.7%) and quotes (11.5%). Looking at quoting, the distribution is very similar. Regarding replies, however, most of these tweets react to other replies (71.7%).

Table 5.3: Distribution of Tweet variants when performing actions.

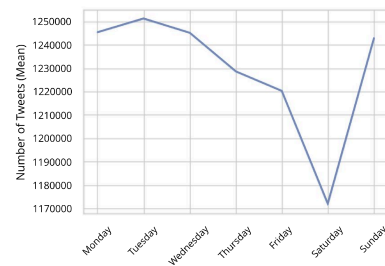
Action	Tweet Variant		
	Original Tweet (%)	Reply (%)	Quote (%)
Retweeting	66.8	21.7	11.5
Replying to	24.7	71.7	3.6
Quoting	76.5	14.5	9.0

5.1.3.2 Tweets over Time

The volume of daily captured tweets varies from 1M to 1.6M messages with an average of 1.2M. By examining the average collection of tweets by weekdays, we observed that German-speaking Twitter users were more active from Sunday to Tuesday and had a decreasing interest in Twitter from Wednesday to Saturday, with the lowest activity on Saturdays (Fig. 5.3 (a)).



(a) Twitter activities over the course of every weekday.



(b) Twitter activities over the course of a week.

Figure 5.3: Tweets over time.

The overall daily usage (Fig. 5.3 (b)) is moderate in the morning, increases during after-work hours, and drops to its lowest point at night between 1am and 5am. At the weekend, Twitter usage naturally starts a few hours later in the morning. The oddly shaped peak on Sunday evenings results from high volumes of tweets during the night of the 2019 European Parliament election. The daily Twitter activities match Central European Time and the working schedule of people from Germany and Austria.

5.1.3.3 Tweet Content

The content of each tweet can consist of text and additional, interactive content. Table 5.2 shows statistics on the usage of different content types. The most prominent type is *user mentions* (85M). Since every retweet, reply, and quote contains at least one mention to the originator, these automated user mentions make up for 35%, 29%, and 3%, respectively. Therefore, 33% of the 85M mention-objects (28M) are user mentions, which are added manually into a tweet (@*username*). *URLs* (18M) are the second most prominent objects found in 23% of all tweets. There were 6 667 962 distinct *URLs* shared that originated from 275 078 different domains. Since $\frac{1}{4}$ of all users in our corpus actively shared at least one *URL*, it seems typical for the German user base to consume and share content from external sources.

Beside these external sources we extracted 19.7 million (5 874 013 distinct) multimedia-objects. The majority of the multimedia contents shared are photos (82%), followed by videos (12%) and animated GIFs (4%), shared by a total of 56% of the users. Note that we can only obtain multimedia content from text tweets, as at least a single word is needed to identify a tweet to be German. Further, 29% of the users in our data set shared 39 million hashtags in 22% of all tweets. However, while we observe more tweets with hashtags than multimedia content, more users share multimedia content (56%) than hashtags (22%). Users using hashtags are about two times more active on Twitter than users sharing multimedia content, which, to some extent, explains this effect. A feature almost entirely neglected by users in our data set is the submission of geolocation data (*Places*), which supports the results of Scheffler [385]. Only 2% of the users share their location when tweeting.

5.1.3.4 Hashtags and Captured Events

In addition to external sources, users produce a high amount of hashtags. By examining popular hashtags shared during unusual high peaks in daily usage, we could identify the related influential events (see Tab. 5.4). Twitter users were most active at the end of the campaign for the 2019 European parliament elections (May 26). All top hashtags shared on Twitter during this period can be attributed to the election and the associated debate about the election results. The *Christian Democratic Union* (CDU) leads in both the hashtag ranking and the actual election (28.9%). The controversial *Alternative for Germany* (AfD) party follows next (#AfD), although it only came fourth in the election with a total of 11%. In comparison, the *Bündnis 90/Die Grünen* party, which came in second in the election with almost twice as many votes (20.5%), only appears in 11th place in the hashtag ranking. The *Sozialdemokratische Partei Deutschlands* (SPD) landed in third place in the election and also managed to attract more attention on Twitter (#SPD) than *Die Grünen*. However, the popularity of the hashtag #CDU could also be a side effect of Annegret Kramp-Karrenbauer's (#AKK) controversial comments on the political comments of Youtube influencer Rezo (#Rezo), which triggered a general discussion about censorship in online forums (#Censorship). Looking at the context of the hashtags, one concludes that the popularity of the hashtags is the result of lively discussions rather than a reflection of political party affiliation. The beginning of the Rezo controversy can be seen in the spike in tweet volume on May 22 and 23. Youtube influencer Rezo (#Rezo) posted a viral video (#RezoVideo) to express his concerns about the CDU's political course. The video received widespread media coverage and led to a reaction video (never published) by CDU politician Philipp Amthor (#Amthor). Another political controversy occurred on May 18. Austrian politician Heinz-Christian Strache (#Strache) was the main character of a compromising video (#StracheVideo) that caused

Table 5.4: Most shared Hashtags during busiest days of data collection; here GTNM stands for Germany's Next Top Model.

Period	Hashtag	Count	Category	Event
May, 26 th – 28 th	#Europawahl2019	105 498	politics	European Parliament Election 2019
	#CDU	42 570	pol. party	European Parliament Election 2019
	#AfD	36 038	pol. party	European Parliament Election 2019
	#EUWahl19	25 562	election	European Parliament Election 2019
	#AKK	24 697	politician	European Parliament Election 2019
	#Europawahl	24 185	election	European Parliament Election 2019
	#Rezo	23 498	controversy	European Parliament Election 2019
	#SPD	21 519	pol. party	European Parliament Election 2019
	#Zensur	16 087	controversy	European Parliament Election 2019
	#Grüne, #Grünen	14 290	pol. party	European Parliament Election 2019
	#Sachsen	11 272	election	European Parliament Election 2019
June, 2 nd	#NichtOhneMeinKopftuch	124 218	politics	Political campaign
	#Nahles	17 991	politician	Politician resignation
	#SPD	14 698	pol. party	Politician resignation
	#뷔	9 237	music	Campaign by BTS
	#태형	9 234	music	Campaign by BTS
	#태태	9 066	music	Campaign by BTS
May, 18 th	#EurovisionSongContest2019	67 188	music	Eurovision Song Contest 2019
	#Eurovision	45 210	music	Eurovision Song Contest 2019
	#Strache	30 218	politician	Ibiza Affair
	#esc2019	23 038	music	Eurovision Song Contest 2019
	#strachevideo	12 911	controversy	Ibiza Affair
May, 22 nd – 23 rd	#rezo	30 742	controversy	European Parliament Election 2019
	#GNTM	22 980	entertainment	TV Show
	#CDU	20 216	pol. party	European Parliament Election 2019
	#Amthor	19 006	politician	European Parliament Election 2019
	#EuropaWahl2019	15 595	election	European Parliament Election 2019
	#RezoVideo	14 515	controversy	European Parliament Election 2019

the Austrian government coalition to collapse.

Based on the popular hashtags, we see that a high number of political topics are discussed. In addition, previously announced political campaigns on Twitter were also able to generate high volumes of tweets. The hashtag #NichtOhneMeinKopftuch was the most dominant hashtag on June 2, with 124 218 tweets. For comparison, the second most shared hashtag that day was mentioned in only 17 991 tweets.

In addition to political events, pop culture events also dominate Twitter (e.g. #GNTM, #ESC2019). There are also some non-German hashtags that refer to a Korean pop band called BTS, which reached high rankings in the music charts in Germany for several weeks. During our data collection, they released several singles and generated trending hashtags. Most popular events also dominated the news in Germany during the data collection period. Therefore, we can conclude that our data collection correctly collects German-language tweets.

5.1.3.5 External Media Usage

Given the collected and complemented URLs, our approach automatically categorized 98.3% of all shared URLs. Table 5.5 details the distribution volumes of the Top 10 FGs and their categories.

While the users generate most of their traffic in the *Information/Communication* FG (47% tweets), the FG with the maximum user base is *Entertainment/Culture*. The majority of the users of this FG are interested in multimedia content, such as videos and photos (Streaming Media: 36%; Media Sharing: 33%). Most of the content of the *Information/Communication* FG is related to news (General News:

Table 5.5: FG and categorical distribution of the 17 million URL-Tweets (multiple assignments per domain possible) and form of distribution, such as Original Tweets (OT), Retweets (RT), Replies (RP), and Quotes (QT); statistics on third-party services (Third) are included; FGs and categories with less than a 1% share of all tweets are excluded; FGs containing categories of the *News Group* are depicted in **brown**.

Category	Tweets %	Users %	URLs %	OT %	RT %	RP %	QT %	Third %
Information/Communication	47	35	39	46	52	2	1	29
General News	32	21	23	40	57	2	1	23
Blogs/Wiki	10	16	10	52	44	3	1	36
Public Information	2	3	2	68	29	3	1	59
Portal Sites	2	5	2	46	52	2	1	20
Technical/Business Forums	1	2	1	66	31	2	0	52
Forum/Bulletin Boards	1	2	1	64	33	3	0	43
Entertainment/Culture	15	43	13	44	50	5	1	21
Streaming Media	10	36	8	42	52	6	1	17
Media Sharing	8	33	6	39	54	7	1	14
Entertainment	4	10	4	56	41	2	1	38
Internet Radio/TV	1	1	0	69	29	2	1	53
Art/Culture/Heritage	1	2	0	37	60	2	1	21
Lifestyle	12	23	17	65	34	1	0	55
Social Networking	7	18	12	69	30	1	0	64
Sports	3	4	3	66	32	1	1	49
Controversial Opinions	1	1	0	29	70	1	0	11
Travel	1	1	1	84	13	3	0	73
Society/Education/Religion	9	13	7	35	59	6	2	17
Politics/Opinion	3	5	1	27	68	4	2	14
Education/Reference	2	5	2	43	46	9	3	23
Non-Profit/Advocacy/NGO	2	6	2	35	60	4	3	10
Government/Military	1	3	1	30	61	8	4	16
Health	1	1	1	52	41	5	2	35
Purchasing	8	10	10	75	22	3	1	58
Marketing/Merchandising	3	5	4	73	24	3	1	59
Online Shopping	3	4	3	71	25	3	0	52
Auctions/Classifieds	1	1	1	91	9	1	0	57
Business/Services	6	10	8	71	26	2	1	52
Business	4	8	5	65	31	3	2	42
Finance/Banking	1	2	2	78	20	2	1	63
Job Search	1	1	1	92	8	0	0	86
Information Technology	5	12	7	69	28	3	1	56
Internet Services	3	7	4	73	24	2	1	60
Software/Hardware	1	3	2	83	15	2	0	72
Pornography/Nudity	4	5	3	43	56	1	0	43
Pornography	2	2	2	49	51	0	0	63
Incidental Nudity	2	3	1	35	64	1	0	19
Games/Gambling	3	5	2	57	42	1	1	38
Games	3	5	2	57	42	1	1	37
Risk/Fraud/Crime	1	1	1	65	33	2	0	77

32%) and personal blogs (Blogs/Wiki: 10%), mainly consisting of content from online news media and personalized political websites. Based on the high number of retweets in this group (52%), news and blog content seems to be well-received by the user base. We observed the same popularity of political domains in the FG *Society/Education/Religion*, comprised of even more elaborate political content. Most of the URLs captured during our data collection are from domains within the *Information/Communication* FG, which means a high number of different news articles are generated and distributed on Twitter. Despite this variety of articles, the Twitter community still reacts to these links by spreading them via retweets and replies. In contrast to news and political content, lifestyle-related content (FG: *Lifestyle*) results in fewer retweets (34%), which indicates a lower acceptance by the German community. An exception is the category *Controversial Opinions*, which includes domains that share highly opinionated political content (e.g., journalistenwatch.com, philosophia-perennis.de, pi-news.net). McAfee's Trusted-Source grouped *Controversial Opinions* into the FG *Lifestyle*. The number of retweets in this category

is 70%, further supporting the assumption that political content on Twitter is widely distributed and acknowledged. The FGs with the most original tweets are related to marketing campaigns (*Purchasing*: 75%), business advertising (*Business/Service*: 71%), and online technologies (*Information Technologies*: 69%). Third-party services generate a majority of these tweets. Therefore, we assume that most of these domains conduct an automated distribution of their products. The lack of retweets within the respective categories indicates that this distribution approach is not overly effective in the German Twitter community. Further, the share of spam and inappropriate content is relatively small (overall 4% to 7%). Just a tiny margin of users is involved in the distribution process. Spam URLs found in our data were mainly shared via original tweets (97%) and distributed via third-party services (92%), which suggests automated distribution in the context of spam and marketing. Based on the low number of retweets, users recognize spam content and do not distribute these any further in the network. Links disguised by users via link shortening services make up 14% of URLs shared on Twitter. Resolving these links, we discovered that news providers and bloggers use marketing services to distribute their content in an automated manner.

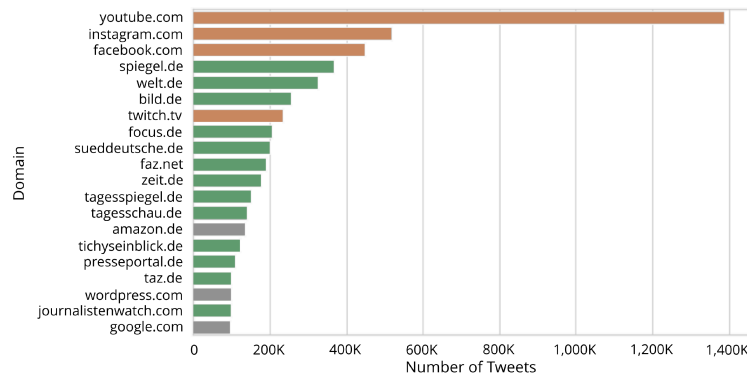


Figure 5.4: Tweet volume of top 20 external sources (orange: OSN, green: news content, gray: other)

A closer look at the 20 most shared external sources (Fig. 5.4) revealed that the top domains are external OSNs, led by YouTube, followed by Instagram and Facebook (Fig. 5.4). These platforms have a significantly higher distribution and more users sharing content from these platforms than any other domain. They are platforms for a variety of content providers. Therefore, we resolved links to *YouTube*, *Facebook*, and *Instagram* to identify popular *YouTube Channels*, *Facebook Pages*, and *Instagram profiles* (see also Sec. 5.1.2.3).

YouTube Overall, the data set contains 1 402 441 tweets that shared 74 414 distinct YouTube-URLs originating from 97 680 different YouTube Channels. Regarding the type of media shared via these platforms *YouTube links* seldom contained other content than video links (97%). The content of shared videos varies from music, gaming, and political opinions to educational content (see Table 5.7).

We identified single videos accounting for large chunks of the YouTube links on Twitter. For example, a newly released single of a Korean pop band (BTS) or a video of a channel called *Rezo* belonging to a person who was at the center of a political controversy surrounding the 2019 European Parliament election. He published a video with the title “Die Zerstörung der CDU” (Engl.: the destruction of the CDU) that went viral, expressing concern regarding the political course of the *CDU*. In general, there is only a small number of frequently shared content providers from YouTube (see Tab. 5.6). Half of these Channels are

Table 5.6: Top external social media profiles (Brown: Political Emphasis).

Provider	Tweets #	Users #	URLs #	Category	Description
YouTube					
<i>체코리아</i>	284 980	282 903	1	music	South Korean singer
<i>Rezo ja lol ey</i>	50 277	30 802	29	political cont.	Rezo controversy
<i>AfD Kompakt TV</i>	14 323	3 413	99	political party	Political party: AfD
<i>Rammstein Official</i>	11 793	8 664	70	music	German band
<i>ProDogRomania e.V.</i>	10 265	617	637	activism	Dog rescue Romania
<i>AfD-Fraktion Bundestag</i>	8 201	2 351	309	political party	Political party: AfD
<i>ibighit</i>	8 094	7 328	36	music	Korean pop band
<i>Joko & Klaas</i>	7 540	6 231	35	entertainment	German entertainers
<i>RT Deutsch</i>	7 138	2 321	1 003	news/politics	Russian news media
<i>Gottfried Curio</i>	6 904	2 330	50	politician	Politician from AfD
Instagram					
<i>@zkdlin</i>	16 845	6 933	202	music	South Korean singer
<i>@oohsehun</i>	9 752	7 102	91	music	South Korean singer
<i>@ksh7909</i>	4 342	3 878	4	music	South Korean singer
<i>@sooyoungchoi</i>	3 579	1 720	9	music	South Korean singer
<i>@daniel.k.here</i>	3 093	3 038	9	music	South Korean singer
<i>@taeyeon_ss</i>	2 666	1 155	22	music	South Korean singer
<i>@saulami1g</i>	2 453	7	2 443	gaming	Gaming/Streaming
<i>@svchicas</i>	2 113	219	2	nudity	Explicit Content/Spam
<i>@stephenathome</i>	1 957	1 953	5	politics	Late night show host
Facebook					
<i>@aliceweidel</i>	16 433	3 867	327	politician	Party member of AfD
<i>@alternativefuerde</i>	11 762	2 930	190	political party	Facebook page of AfD
<i>@Prof.Dr.Joerg.Meuthen</i>	8 149	2 496	85	politician	Party member of AfD
<i>@Bjoern.Hoecke.AfD</i>	3 498	1 523	54	politician	Party member of AfD
<i>@Pazderski.Georg</i>	2 408	910	59	politician	Party member of AfD
<i>@Academia-Para-C...</i>	2 131	222	1	nudity	Explicit Content/Spam
<i>@Deutschland3000</i>	2 012	1 988	15	education	Educational (politics)
<i>@GegenDieAfD</i>	1 740	825	140	activism	Activism against AfD
<i>@GottfriedCurio.AfD</i>	1 694	1 138	10	politician	Party member (AfD)
<i>@app: rossmann.de</i>	1 472	1 079	1	advertisement	Facebook app (shop)

Table 5.7: YouTube URLs: Most shared video categories.

Category	Share (%)	User (%)	Video Count
Music	20	34	47 018
News & Politics	19	18	18 484
Gaming	14	10	40 427
People & Blogs	13	20	29 610
Entertainment	12	21	21 982
Education	4	7	10 274
Science & Technology	4	8	8 332
Film & Animation	3	7	7 967
Nonprofits & Activism	2	4	3 868

related to political topics. Moreover, they show a specific political affiliation. Channels belonging to the right-wing political party AfD are shared more often than channels of any other party. This observation indicates a high activity during their election campaign and shows a trend towards utilizing multimedia content to reach a broader spectrum of users.

Instagram Although the number of shared Instagram-URLs (520 466) amounts to only a third of the distributed YouTube-URLs, they show a similar number of distinct URLs (370 510). This data suggests that Instagram posts go less often viral compared to YouTube content. Most of the content shared via Instagram links are images (71%), direct posts (12%), which also contain multimedia content, and profile pages (10%). Overall, shared Instagram links are mostly apolitical and dominated by profiles from the entertainment industry (Tab. 5.6). Therefore, we conclude that only an insignificant number of users consume the content of Instagram profiles as a source to get news updates.

Facebook The distribution ratio of Facebook links shows a high similarity to Instagram shares. In total 454 128 Facebook-URLs were distributed, with 292 316 distinct destinations, originating from 96 221 different Facebook profiles. The content from *Facebook-links* is mainly textual (post: 53%; story: 13%) and less multimedia-based (photo: 10%; video: 8%). There are only a few events and groups shared within our corpus. Looking at the most shared Facebook profiles (Tab. 5.6), we observe a relatively small user base that only supports a handful of Facebook pages or profiles, with a low distribution factor. We notice, however, that most Facebook profiles are politically motivated and shifted towards the right-wing party AfD. One exception to this rule is a frequently shared page that directly opposes said party (@GegenDieAfD).

A portion of 55% retweets and 7% replies when sharing YouTube-URLs suggests that users distribute YouTube videos to support the content and communicate with other users. On the other hand, the Twitter user base widely ignores Facebook and Instagram URLs. These links get mostly shared via original tweets (Instagram: 72%, Facebook: 67%). While users distribute YouTube links primarily via the official mobile and web clients from Twitter, they share most Facebook and Instagram content via third-party services. We assume that most Facebook and Instagram users share their content passively while actively using Facebook and Instagram clients. They share content on these platforms and forward them to their Twitter profiles to extend their reach. However, the low number of retweets (Instagram: 27%, Facebook: 31%) indicates that this strategy is not very effective. Consequently, we conclude that Facebook and Instagram content is perceived less distinctly than YouTube or other shared media content.

Overall, the top content providers from YouTube and Facebook are mostly related to political parties and activism. We observed that the German political party *AfD* was highly active on social media. Regarding shared links from Facebook (6 out of 10) and YouTube (3 out of 10), *AfD*-related topics dominated this content. However, in general, compared to news media sources that distribute their articles directly on Twitter (see Fig. 5.4), content providers that operate from other social media networks attract considerably less attention. URLs of the 10 most shared content providers reach an average shares per unique URL (S/U) ratio of 12.58. YouTube- (S/U: 3.75), Instagram- (S/U: 1.40) and Facebook links (S/U: 1.55) were shared less often, and, hence, failed to generate reach and impact. With regard to the total number of tweets, only *BTS* and *Rezo* were able to generate reach comparable to other popular content providers on Twitter.

5.1.4 News Consumption within the GTC

We established that 13 of the 20 most shared external sources (see Fig. 5.4) link to popular German news providers such as *Spiegel*, *Welt* or *Bild*, as well as to smaller news/opinion blogs, such as *Tichy's Einblick* and *Journalistenwatch*. To further our understanding of news distribution within the German-speaking Twitter community, we analyze the subset of shared external sources that link to news-related content. We defined the *News Group* as a collective term that comprises external domains related to news, political/controversial opinions, and educational content (Sec. 5.1.2.3). We compare the popularity, reach and impact of categories and content providers within the News Group by analyzing the volume of tweets they generated, the number of users they mobilized, and the number of reactions they prompted. Further we investigate the distribution of political hashtags and controversial content.

5.1.4.1 News Content

Table 5.8 shows the volume of tweets, users, and URLs within the News Group.

Table 5.8: News Group Volume: Number of URL- and Reaction-tweets

Data Set	Tweets (#)	Tweets (%)	Users (#)	Users (%)	URLs (#)	URLs (%)
URL-tweets	17 478 261	100	1 720 752	100	6 667 962	100
News Group	7 247 843	41	454 381	26	1 903 133	29
Reaction-tweets	9 582 682	100	1 222 863	100	1 193 232	100
News Group	5 660 382	59	391 139	32	515 883	43

Approximately 41% of all URL-tweets distribute content that belongs to this group. However, only 26% of the users sharing URLs belong to this group. The ratio between URL-tweets (41%) and distinct URLs (29%) in the News Group implies that the average URL is shared 3.81 times. Compared to the average distribution of non-members with a distribution factor of 2.15, the News Group is more active in sharing the content of interest. Therefore, URLs shared on Twitter predominantly link news-related content.

Table 5.9: Categorical usage and distribution of 7M URL-tweets (450k users), 6M reaction-tweets (390k users) within the News Group (URL/reaction); categories are distinguished by political views from: *moderate*-to *extreme*.

Category	Tweets %	Users %	Distribution (%)				
			OT	RT	RP	QT	Third
General News	77 / 82	79 / 85	41	57 / 41	2 / 49	1 / 22	23 / 3
Politics/Opinion	8 / 8	19 / 20	27	68 / 47	4 / 44	2 / 25	14 / 3
Education/Reference	5 / 4	20 / 15	43	46 / 36	9 / 54	3 / 22	23 / 4
Non-Profit/Adv./NGO	5 / 3	21 / 14	35	60 / 48	4 / 40	3 / 32	10 / 4
Controversial Opinions	3 / 3	2 / 3	29	70 / 68	1 / 25	0 / 14	11 / 1
Government/Military	3 / 3	13 / 14	30	61 / 43	8 / 46	4 / 34	16 / 3
Major Global Religions	1 / 1	3 / 3	42	54 / 35	4 / 54	2 / 20	20 / 3
Discrimination	< 1 / < 1	< 1 / < 1	42	32 / 51	24 / 41	2 / 12	5 / 1
Historical Revisionism	< 1 / < 1	< 1 / < 1	59	19 / 60	22 / 32	< 1 / 34	2 / < 1

The *News Group* comprises nine categories. These categories allow us to examine the reach w.r.t. different types of news. Table 5.9 gives an overview of the sharing behavior viewed by category. Most URL-tweets originate from moderate domains (General News: 77%). Besides religious- (20%) and educational content (23%), general news is with 23% on the top of the list w.r.t. the distribution via third-party services. Regarding support via retweets, we observe that news sources that offer tendentious to extreme views on politics (i.e., Politics/Opinion and Controversial Opinion) are the most supported domains (retweet factor: 68 – 70%). However, the average distribution of URLs via retweets is consistently high in almost all categories. An exception is URLs propagating extreme political views, i.e., discrimination and historical revisionism. With a retweet factor of 19 – 32%, such content experiences significantly less support via retweets. Interestingly, however, these links seem to be often used within discussions, resulting in a 22 – 24% URL-tweet share via replies (others: 1 – 9%).

The data suggests three different support patterns: (i) highly shared and discussed articles, (ii) highly distributed articles via retweets (68 – 70%), and (iii) articles supported via replies (22 – 24%) but mainly ignored by the general public ($\leq 3\%$ of all users). These support patterns correlate strongly with the subjectivity level of shared content, i.e., *moderate* domains are supported by (i), *tendentious* outlets by (ii), and *extreme* domains by (iii). Overall, we rarely observe extreme external content. A share of $< 4\%$ of URL-tweets, actively shared by $< 4\%$ of the users, and rarely replied to, extreme content seems to be

widely ignored by most Twitter users.

Table 5.8 shows the number of tweets commenting on or referencing URL-tweets. We found that most reaction-tweets (59%) occurred in the News Group. Furthermore, 43% of the URLs that prompted reactions on Twitter originated from the News Group. The proportion of users (32%) and tweets (59%) indicates a highly active News Group.

We analyzed the distribution of reaction-tweets considering each category of the News Group (see Table 5.9). In contrast to the distribution of URL-tweets, we registered almost no Reaction-tweets from third-party services. Regarding discussions, users commented on moderate content actively (replies + quotes: 71 – 80%), followed by tendentious articles (69%). The more extreme the content, the more “discussions” via retweets (extreme content: > 50%) with an active discussion ratio (replies + quotes) of 53 – 66%. These results suggest that some users continue to disseminate and support controversial opinions, while others are less likely to respond to such content (reply rate: General News 49% vs. Controversial Opinions 25%).

In terms of activity levels, it can again be seen that users discussing extreme content are the most active, with a ratio of tweets per user of 8.93. Users discussing tendentious content are this time more similar to users discussing moderate content, with 5.43 and 4.84 respectively. We find that users are more active in discussing than sharing moderate content 4.84 versus 3.91. The reverse is true for tendentious (5.43 vs. 10.87) and extreme (8.93 vs. 12.20) content.

Throughout, we observe many replies and quotes. Therefore, we assume that users heavily engage in political discussions. Note that the cumulative percentages of retweets, quotes, and replies exceed 100% because retweets may contain nested quotes and replies, which we counted as a retweet of each instance in this case.

5.1.4.2 News Content Provider

Next, we compare the popularity, reach and impact of content providers within the News Group. The comparison also considers the amount of self-promotional Tweets produced by feed-profiles and corresponding journalists. Figure 5.5 depicts the tweet volume broken down by provider. Further, table 5.10 (right column) shows additional data regarding tweet distributions.

The user/tweet ratio reveals two distinct user types. Followers of domains such as *tichyseinblick.de*, *journalistenwatch.de*, or *philosophia-perennis.com* (tendentious to extreme views) have the most active users with a user to tweet ratio of 10.83 (tendentious) and 12.20 (extreme). In comparison, readers of traditional news outlets such as *Spiegel* or *Zeit* only have a ratio of 4.54 and 3.21, respectively (traditional German news providers: 3.91).

Also noticeable, Twitter users sharing less moderate outlets retweet more often, with *tichyseinblick.de*, *jungefreiheit.de*, *taz.de*, and *philosophia-perennis.com* as top domains in this category and retweet counts ranging from 81% to 91%. Note that similar to *Bild*, while *taz* is part of moderate news media, in the past several articles with tendentious, disputable content were rebuked by the German Press Council⁵. Traditional media sources, in contrast, reach a broader spectrum of users, but their popularity partially depends on the number of articles they publish. Users neither share the links from moderate nor tendentious media via replies, indicating that users less often reference such content within discussions. In this category,

⁵https://de.wikipedia.org/wiki/Die_Tageszeitung#Presserats%C3%BCgen

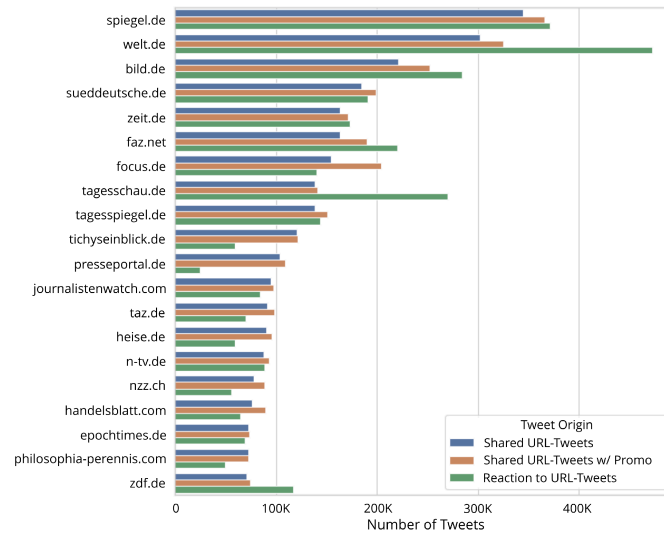


Figure 5.5: Tweet volume of the top 20 news domains, comparing URL-tweets (with and without promotion profiles) and Reaction-tweets.

the governmental outlet *bundestag.de* shows the highest reach in this context with distributions via replies and quotes of 6%, each.

Figure 5.5 depicts the URL-tweet volumes and the number of Reaction-tweets concerning each content provider. Traditional news media, such as *Spiegel*, *Welt*, and *FAZ*, trigger a high amount of Reaction-tweets, exceeding the number of tweets that share their articles. It suggests that users actively discuss their content. Of note are the statistics of *Tagesschau*. While showing only moderate amounts of URL-tweet shares, it prompted the 4th-highest number of Reaction-tweets. Not reaching the number of their respective URL-tweets, tendentious and extreme media providers, in contrast, receive fewer reactions. Table 5.11 gives a more detailed overview of the reactions prompted by the 30 most shared domains in the News Group.

For instance, *Spiegel* received 371 725 Reaction-tweets to 10% of tweets that shared a *Spiegel* article. In total, 72 881 users reacted to 17 884 distinct URLs from *Spiegel*, encompassing 35% of all unique *Spiegel* URLs shared on Twitter. Thereby, only 14% of the users that shared a *Spiegel* article received any reaction. The outlets of *Welt*, *Bild*, *ZDF*, and *Tagesschau* trigger a significantly larger user base discussing their content than the user base that shares it. Other outlets show more similar ratios. Overall, traditional media providers attract more users than controversial sources. However, the smaller user base of the controversial news providers generated the highest Reaction-Tweet per user rate. This further shows that the user base of controversial, political content providers are more active than consumers of traditional media. While publishing a relatively small contingent of articles, 37-64% of distributed URLs receive reactions.

Promotional Profiles We complement information on the reach of news providers by studying promotional profiles, i.e., we consider self-promotional tweets produced by feed-profiles and corresponding journalists. Table 5.10 (left column) details the results of our promotional profile detection process for each of the 30 content providers. For instance, we identified 98 promotional profiles from *Spiegel*, comprised of 68 journalist-profiles and 30 feed-profiles. Over two months, these profiles produced 20 850 tweets, which constitutes a daily average tweet volume of ~ 342 tweets (per account: ~ 3.49). These accounts mainly distributed content via original tweets (86%) and shared them via third-party services (72%). In

Table 5.10: Statistics on identified promotional profiles (left) and reach (right) of the most shared content providers within the News Group (U = Users, J = Journalists); well-respected traditional German news providers (acc. to [475]) are highlighted in blue.

Content Prov.	U #	J #	Feeds #	Tweets #	URLs %	OT %	RT %	RP %	QT %	3rd %	UTweets #	Users #	U/T #	URLs #	OT %	RT %	RP %	QT %	3rd %
<i>Spiegel</i>	98	68	30	20 850	24	86	13	0.42	0.18	72	344 946	76 028	4.54	47 805	31	66	2	1	12
<i>Welt</i>	72	53	19	23 017	39	88	12	0.23	0.08	90	302 259	47 139	6.41	43 842	26	71	3	0.40	7
<i>Bild</i>	138	59	79	31 059	52	97	3	0.05	0.01	94	221 394	30 547	7.25	24 526	21	78	1	0.21	5
<i>Sueddeutsche</i>	76	38	38	14 231	27	91	8	0.43	0.16	81	184 966	55 572	3.33	20 990	24	74	2	0.43	9
<i>Zeit</i>	80	59	21	7 724	24	88	11	1	0.27	82	163 648	51 052	3.21	21 314	23	73	4	1	9
<i>FAZ</i>	59	34	25	26 322	35	94	6	0.16	0.08	89	163 531	40 784	4.01	32 909	28	70	2	1	8
<i>Focus</i>	20	5	15	49 952	53	81	19	0	0	81	154 124	23 751	6.49	25 050	27	71	2	0.25	11
<i>Tagesschau</i>	10	8	2	2 463	21	98	2	0.04	0.04	95	138 522	39 178	3.54	10 770	27	71	2	0.46	14
<i>Tagesspiegel</i>	77	57	20	12 559	38	45	53	1	2	0	138 274	38 231	3.62	13 590	18	79	2	1	6
<i>Tichys Einblick</i>	2	1	1	1 483	29	88	9	1	1	15	120 412	9 344	12.89	2 315	8	91	1	0.34	3
<i>Presseportal</i>	3	1	2	5 550	10	100	0	0	0	100	103 418	10 401	9.94	52 888	55	44	1	0.48	43
<i>Journalisten...</i>	1	0	1	2 151	56	100	0	0	0	100	95 225	6 470	14.72	3 789	22	78	0	0.04	3
<i>taz</i>	43	26	17	6 754	44	91	8	1	0.07	83	91 674	29 327	3.13	8 432	16	82	2	1	7
<i>Heise</i>	26	14	12	4 946	23	76	24	0.22	0.14	93	90 829	25 667	3.54	15 330	38	58	3	2	22
<i>n-tv</i>	12	5	7	5 177	24	95	4	0.19	1	85	88 036	21 254	4.14	19 304	32	66	2	0.39	13
<i>NZZ</i>	82	55	27	10 395	35	58	38	4	0.14	20	78 256	22 146	3.53	14 883	33	64	2	0.39	11
<i>Handelsblatt</i>	66	60	6	13 090	36	74	25	0.34	0.11	60	76 174	22 765	3.35	23 212	38	60	2	0.44	19
<i>Epochtimes</i>	2	1	1	192	2	100	0	0	0	100	73 086	6 898	10.60	8 313	25	74	1	0.13	6
<i>Philosophia p...</i>	0	0	0	0	0	0	0	0	0	0	72 926	6 408	11.38	1 457	19	81	1	0.23	2
<i>ZDF</i>	64	35	29	3 877	29	80	14	5	0.48	19	70 492	27 887	2.53	9 914	18	79	2	1	6
<i>Der Standard</i>	33	28	6	8 131	39	96	4	0.33	0.09	89	66 846	15 745	4.25	14 495	37	60	3	1	13
<i>change.org</i>	11	8	4	169	0	51	31	14	4	3	61 586	27 915	2.21	34 674	59	39	3	1	3
<i>Junge Freiheit...</i>	2	1	1	672	26	83	17	0	0	1	56 823	7 026	8.09	1 864	14	85	1	0.07	5
<i>Deutschlandf...</i>	4	3	2	3 824	31	86	9	5	0.18	4	49 799	19 815	2.51	9 178	25	71	3	1	9
<i>WDR</i>	38	24	16	3 966	36	72	15	9	5	15	44 134	18 345	2.41	5 218	18	80	2	1	7
<i>bundestag.de</i>	4	1	3	71	1	100	0	0	0	8	43 280	19 733	2.19	4 233	17	76	6	6	9
<i>BR</i>	59	40	20	8 937	63	87	10	2	1	47	42 108	16 190	2.60	7 805	23	74	3	1	9
<i>Stern</i>	17	10	7	5 588	29	98	2	0.18	0.07	93	41 611	14 892	2.79	16 231	43	55	2	0.46	27
<i>NDR</i>	37	22	15	3 860	32	64	34	2	1	9	40 851	16 025	2.55	6 851	28	70	2	1	13
<i>RT</i>	10	0	10	2 334	40	91	1	7	0	0	39 262	6 683	5.87	5 189	31	67	2	0.31	7

Table 5.11: Reaction-tweets towards the 30 most distributed content providers from the News Group

Provider	Tweets		Users		URLs		Distribution (%)			
	#	%	#	%	#	%	RT	RP	QT	Third
<i>Spiegel</i>	371 725	10	72 881	14	17 884	35	36	56	20	2
<i>Welt</i>	472 260	11	62 868	15	24 592	46	37	56	19	2
<i>Bild</i>	283 968	10	43 158	13	14 790	48	40	52	15	1
<i>Sueddeutsche</i>	190 727	9	53 591	12	9 088	40	37	53	25	3
<i>Zeit</i>	173 291	9	45 878	12	9 024	41	31	60	21	2
<i>FAZ</i>	220 151	10	46 895	13	14 820	40	34	57	23	2
<i>Focus</i>	140 549	7	23 089	15	8 960	19	46	46	15	1
<i>Tagesschau</i>	270 323	7	50 758	10	4 623	43	41	53	18	2
<i>Tagesspiegel</i>	144 017	8	37 048	12	7 176	50	37	53	27	2
<i>Tichys Einblick</i>	59 531	3	10 331	9	983	42	42	48	20	1
<i>Presseportal</i>	24 369	5	9 297	9	5 233	10	41	49	24	5
<i>Journalistenw...</i>	84 145	8	7 616	10	2 450	64	68	25	14	1
<i>taz</i>	69 589	8	24 225	9	4 547	50	38	51	32	3
<i>Heise</i>	59 171	10	18 783	12	5 024	31	63	30	15	11
<i>n-tv</i>	88 275	11	23 658	13	8 027	39	44	48	16	2
<i>NZZ</i>	56 114	10	18 572	13	5 306	31	39	51	23	2
<i>Handelsblatt</i>	64 816	10	22 968	13	7 456	26	38	50	27	2
<i>Epochtimes</i>	68 923	7	7 542	14	3 144	37	63	30	18	0
<i>Philosophia p...</i>	49 726	6	7 558	11	721	49	68	26	12	1
<i>ZDF</i>	117 548	8	31 993	8	4 559	44	33	58	23	2
<i>Der Standard</i>	60 219	13	14 356	14	6 770	37	47	44	23	3
<i>change.org</i>	17 808	6	9 697	8	3 277	9	55	38	20	3
<i>Junge Freiheit</i>	45 106	5	8 928	10	752	40	41	50	21	1
<i>Deutschlandf...</i>	49 611	11	18 625	11	4 290	45	31	57	24	3
<i>WDR</i>	40 341	9	16 336	9	2 525	45	38	50	30	3
<i>bundestag.de</i>	26 863	6	12 334	8	1 135	27	45	44	40	3
<i>BR</i>	43 720	11	16 508	10	4 325	38	44	46	28	3
<i>Stern</i>	46 682	13	16 232	9	5 411	29	32	60	15	2
<i>NDR</i>	36 454	10	15 287	11	2 948	41	40	50	22	3
<i>RT</i>	36 089	12	7 524	11	3 193	58	58	34	18	1

the process, they actively distributed 27% of the distinct URLs from *spiegel.de* shared during the two months.

In general, we observed that predominantly traditional news media sources, such as *Spiegel*, *Welt*, *Bild* and *FAZ* disseminate their articles via third-party services to extend their reach on Twitter. In particular, *Focus* utilizes a sophisticated feed-profile network that produces a massive volume of tweets. Besides *Bild* with 52%, *Focus* also covers (53%) most of their articles circulating on Twitter, only topped by *Journalistenwatch* (56%) and *BR* (64%). In contrast, non-commercial public news media such as *Tagesschau* and governmental news providers such as *bundestag.de* only generate small amounts of such tweets utilizing significantly more diminutive Feed-networks.

We observed that tendentious to extreme outlets, such as *Tichys Einblick*, *Philosophia perennis*, *Journalistenwatch* and *Epochtimes*, generate much less self-promotional tweets than traditional media. Note, however, that these findings could also be an artifact due to our detection approach, i.e., the corresponding promotional profiles could not adhere to media best practices (see Sec. 5.1.2.3).

5.1.4.3 Political Hashtags

Next, we analyze hashtag-usage w.r.t. categories to further our understanding. Table 5.12 gives an overview of popular hashtags within news categories and compares URL- and Reaction-tweets.

Table 5.12: Popular Hashtags within the News Group.

Category	Hashtags (URL-Tweets)	Hashtags (Reaction-Tweets)
General News	AfD, SPD, Berlin, CDU, ots, news, Europawahl2020, Merkel, FridaysForFuture, EU, Europawahl, Deutschland, NotreDame, Klimaschutz, Polizei	AfD, SPD, CDU, Europawahl2019, Merkel, FridaysForFuture, EU, Berlin, NieMehrCDU, Deutschland, Europawahl, Rezo, Strache, FPÖ, Klimaschutz
Politics, Opinion	AfD, Europawahl2019, EU, Europawahl, PIRATEN, SPD, Europa, Bundestag, EP2019, CDU, Prüffall, Deutschland, FridaysForFuture, Liebe, ReconquistaInternet	AfD, Europawahl2019, CDU, SPD, NieMehrCDU, PIRATEN, Europawahl, TERREG, EU, Piraten, NieMehrSPD, Uploadfilter, FridaysForFuture, FDP, CSU
Education, Reference	FridaysForFuture, Rezo, Europawahl, Europawahl2019, Klimaschutz, FFFfordert, actnow, Digitalisierung, OSTSTEINBBEKKER 音, GrimmsWort, OTD, Berlin, DOYOUNG, KI, Stellenangebot	FFFfordert, actnow, wespoke, OER, GoBlue, kangdaniel, 김종현, WelcomeBackDaniel, 임영민, ABSOLUTE6IX, Marburg, noplanetB, twitterlehrerzimmer, wählegehen, Twitterlehrerzimmer
Non-Profit, Advocacy, NGO	Europawahl2019, Rezo, Zensur, Homöopathie, Meinungsfreiheit, Klimaschutz, Europawahl, Europa, FridaysForFuture, Klimakrise, EU, Uploadfilter, AfD, Berlin, Transsexuellengesetz	Lifeline, Scientists4Future, Florida, unteilbar, Atheisten, AlleGegenRWE, Weimar, OperationSophia, SafePassage, GrandTheftEurope, Economists4Future, Garzweiler, Thema, Upskirting, GamerGate
Controversial Opinions	FFD365, AfD, anonymous, anonymousnews, NotreDame, Merkel, EU, SPD, Antifa, EU19, Berlin, Grüne, CDU, Papst, Migration	anonymous, OliverFlesch, RRG, anonymousnews, MiloYiannopoulos, ramadan, Sperre, Obdachloser, MeinungsfreiheitAuchFürDumme, Schönleinstraße, FFD365, Grosz, einschönesOsterfest, pädophil, homophob
Government, Military	Bundestag, AfD, keinluxus, Klimaschutzgesetz, Feuerwehr, Polizei, Klimaschutz, Europawahl2019, FridaysForFuture, ParentsForFuture, Petition, Fahndung, EU, Braunkohle, Urheberrechtsreform	Urheberrechtsreform, Feuerwehr, Urheberrechtsreform, BVerfG, Protokollerklärung, Fahndung, KeinAber, copyright, 1919LIVE, SPC_Watch, Vermisstenfahndung, txwx, NRWE, Barcelona, Rossell
Major Global Religions	Kirche, AfD, NotreDame, Frauen, Europawahl, Missbrauch, ZdK, Sternberg, Karwoche, Ostern, PapstFranziskus, Woelki, Europa, GehtWählen, Papst	Karwoche, BenediktXVI, Glaube, Ratzinger, Benedikt, Maria20, kirche, Tagesevangelium, klerikal, Kirchenkrise, Kirchenaustritt, Sexualität, berührende_Erzählung, Gründonnerstag, Freitagsworte
Discrimination	ISIS, falseflag, Churchill, H8Front, H84U, Weltkrieg, PeterPadfield, KJM, RudolfHess, niemehrCDU, niemehrSPD, sydney, kalergiplan, Gunskirchen, IMMIVASION	falseflag, ISIS, AfD, leftwing, Gruene, Gewalt, H8Front, H84U, Ibizagate, Linke, Podcast
Historical Revisionism	Churchill, Weltkrieg, Grundgesetz, PeterPadfield, RudolfHess, GG70, Freimaurerei, Verfassungsschutz, Kommunismus, 1Mai, Kühnert, Verfassung, Nationalsozialismus, Sozialismus	Grundgesetz, GG70, Euro, Verfassung, Verfassungsschutz, Verfassungsrichter

A variety of content providers report on the same events. Therefore, many popular hashtags, such as #AfD, #Europawahl2019, and #Rezo, appear in multiple categories. We also observe that most of the hashtags in General News exhibit a political background. Based on the reaction-tweets prompted by these categories, we observe that users often reply to political news concerning the CDU with hashtags

that dissent the party and its coalition partner (e.g., #NieMehrCDU, #NieMehrSPD). Reaction-tweets indicate that users discuss the shared news and use hashtags to express their opinion. While many news articles express less extreme opinions, reaction-tweets express their views more directly. Tweets that distribute controversial news content also receive attention from users with opposing opinions, observable by the usage of hashtags like #MeinungsfreiheitAuchFürDumme (Engl: free speech even for idiots) and #homophob within the reaction-tweets. Users sharing discrimination sources use hashtags opposing the CDU and SPD. In contrast to other categories, reaction-tweets contain fewer opposing hashtags. Only a minor fraction of the user base discusses content from discrimination sources without attracting much attention from users opposing their views.

5.1.4.4 Controversial News Content and its Users

We complement our studies, exploring the distribution of controversial, anti-democratic content. According to Section 5.1.2.3, we label users as either Controversial or Non-Controversial Users. Even though traditional news providers dominate news-related content within the GTC, actors spreading and supporting controversial opinions are also part of the landscape.

In the top 30 news providers on Twitter there are also three which spread anti-democratic content (see Tab. 5.10). *Epoch times*, supported by 6 900 users, *Journalistenwatch* supported by 6 471 users and *Philosophia perennis* supported by 6 408 users. The group of users supporting at least one of these three domains includes 10 694 accounts. 3 555 of which share articles from each of these three sources. Furthermore, it can be observed that a large part of these users also share articles from politically right-winged platforms that we do not consider to be extreme (*Tichy's Einblick*, *Junge Freiheit*), e.g. there are still 2 922 users who share articles from each of the five platforms (*Epoch Times*, *Philosophia perennis*, *Journalistenwatch*, *Tichy's Einblick* and *Junge Freiheit*).

In terms of responses to URL tweets from these providers, 12 809 users participated in the discussions (*Epoch Times* 7 542, *Philosophia perennis* 7 558, *Journalistenwatch* 7 616). Combined, this results in a group of 15 811 users who share or discuss these articles. Including *Tichy's Einblick* and *Junge Freiheit*, this figure grows to 22 334 with 19 043 users that responded to these URL tweets.

On the one hand, content providers that distribute tendentious to extreme political content play only minor roles in the network (see tab. 5.9). On the other hand, their supporters are highly active. While popular domains that share anti-democratic content only have roughly $\approx 6\,500$ supporters (combined: 10 694) and an active audience of $\approx 7\,500$ users, *Journalistenwatch* (Position in Top30: 12th with 95 225 Tweets supporting and 84 145 Tweets discussing the content), *Epochtimes* (18th: 73 086 / 68 923), and *Philosophia perennis* (19th: 72 926 / 49 726) are among the 30 most shared news providers in the GTC. For example, 9 088 articles of the renowned newspaper *Süddeutsche Zeitung* were discussed by 53 591 users in 190 727 tweets (*Tweets/audience*: 3.56, *Tweets/article*: 20.99), while only 721 articles of the anti-democratic domain *Philosophia perennis* were discussed by 7 558 users in 49 726 tweets (*T/audience*: 6.58, *T/article*: 68.97). So, 7-times more people discussed articles of the moderate outlet in comparison to articles of the anti-democratic domain. The moderate discussions, however, only generated 3.8x more tweets with only 2x the number of retweets involved in the anti-democratic discussions. Here, 68% of the 'discussions' were in form of retweets.

Overall, we have a group of 11 129 users who support anti-democratic content. While these users prefer controversial information sources, they also share many articles from traditional news providers. In

particular, articles from the large daily newspapers *Welt* and *Bild* (both conservative) attract a considerable attention from Controversial Users. Overall, 92% of the Controversial Users shared a traditional news provider at least once. They also use a wider variety of content providers (\varnothing 6) than Non-Controversial Users (\varnothing 3) to inform themselves.

Regarding the reactions towards domains, we discovered that Controversial Users mainly react to traditional news sources. Articles from *Welt* caught the attention of many users in this group. Primarily, these users commented on political news articles that voice critical opinions about the *AfD* or reports about topics like immigration or political and climate activism. A closer look at related articles revealed several instances of comments on misinformation content (e.g., faked statistics) in favor of critical views about immigration. Moreover, Controversial Users fiercely commented against news articles that were generally positive on Islam and immigration. In contrast, extreme information sources received virtually no attention from Non-Controversial Users.

In addition, Controversial Users produce a large share of political hashtags (Sec. 5.1.2.3). For example, the #AfD hashtag appears in 469 987 tweets shared by 49 883 users. While Controversial Users only make up for 15% (7 239) of these users, they generated 55% of these tweets. We made similar observations for most of the other tweets regarding political hashtags, such as #Merkel, #Islam, #Flüchtlinge (en: refugees), and #Migranten (en: immigrants). Despite their small numbers, Controversial Users, on average, distribute 42% of the tweets that contain political hashtags.

At first glance, this high frequency of interactions with various users contradicts the assumption that people with more extreme political ideologies tend to form echo chambers [53]. Taking their high activity, hashtag usage, content, and shared URLs into account, a picture similar to the findings in [204, 50] emerged. A minor group of extreme users – formed according to standard echo chambers – spread out to aggressively support their opinions in public. Most of their interactions with external political content are responses to Tweets from Non-Controversial Users. They use the Reply-function (22-24%) to inject their content into discussions (see Tab. 5.9). Non-Controversial Users, on the other hand, tend to remain in their moderate area of political discussions, ignoring external content that supports extreme political ideologies.

5.1.5 Limitations

To cope with a large data set, we formulated several assumptions. Thereby, we accepted certain limitations of our approach. Studying the content discussed on Twitter via shared external content seems a rough estimate in the first place. However, curated third-party services significantly reduce the complexity of content understanding. Looking at a handful of domains to understand an FG and, thereby, thousands of articles/domains helped us cope with the sheer amount of data. Also, due to the restrictions on tweet length, URLs offer themselves an easy way to share opinions. Statistics on our data set support and confirm our abstraction approach. Alone $\frac{1}{3}$ of all tweets discussed news-content. Including other discussed content, the method allows understanding large parts of discussed topics.

We collected users active during the collection phase. Therefore, we missed all inactive users, even if these users passively consumed content on Twitter. Follower-information would have provided data on passive users (having other drawbacks). The information would also have allowed for more detailed approximations of reach and impact of content. However, concentrating on a virtually complete snapshot of the targeted community made it impossible to collect this information (request limitations).

We only labeled users as *controversial* that actively shared an article from extreme, anti-democratic

domains. This probably leads to the fact that we have an uncertainty in the group of non-controversial users. However, we argue that the imprecision introduced in this way has a smaller impact (arguably none) because it affects by far the larger group of users to a much smaller extent.

Finally, our crudest approximation regards promotional profiles. To rely on voluntarily provided information from the relevant account carries some risks. Especially the striking difference between traditional news providers (where we identified plenty of promotional profiles) and tendentious to extreme news-content providers (almost none) needs further investigations. To mitigate these uncertainties, one could use shared external content information.

Further research on the detection of automated accounts is also needed. We decided to ignore the noise introduced by bots, because recent reports question current detection solutions. According to Majo-Vazquez et al. [295], e.g., accounts we focused in our research are especially prone to get suspended due to their behavior rather than bot activities.

Social media platforms are very dynamic. Thereby, changes in platform functionalities could potentially also change directly user behavior. Our data collection approach also depends on the API functions offered by the platform. Since mid 2023, the filtered stream endpoint of API v1.1, which we used, is deprecated⁶. The new API v2 also offers a filtered stream, but has a more restrictive limit for capturing tweets⁷. Thus, meanwhile our data acquisition approach for Twitter, to capture a complete new GTC sample, works only with an enterprise access which starts with \$42k per month.

5.1.6 Concluding Remarks regarding News Consumption within the GTC

This chapter focused on Twitter's German-speaking user base and their behavior. We emphasized external information sources contributing to the forming of political opinions. The goal was to estimate the extent and impact of news-related content, as well as influential actors and user engagement regarding this content. We captured the Twitter traffic of German-speaking users over two months during the 2019 European Parliament election. By utilizing the Twitter Streaming API for our systematic collection approach, we obtained a representative snapshot of the German-speaking Twitter community, comprised of 77M tweets from 6.9M users.

Evaluations yielded detailed insight into the daily and weekly routines of the German Twitter population. In particular, we found that political events, such as several controversies and the election, lead to significant activity increases. To further study the news consumption of users, we automated categorized external content. We believe that such an approach provides a powerful tool for identifying meta-information in large-scale networks.

Overall 41% of all shared tweets, containing at least one link, were related to news content. Accounting only for unique URLs, we observe that news-related content makes up 29% of these. The discrepancy, a share of 29% unique URLs making up for 41% of all links, further emphasized the popularity of news-related content. Most popular traditional news providers, such as *Spiegel*, *Welt*, *Bild*, *Sueddeutsche*, *Zeit*, and *FAZ* are at the top of all content providers within the GTC. Further, established German TV stations (e.g. *ZDF*, *WDR*, *BR*), related content (e.g. *Tagesschau*), and traditional news outlets from Switzerland (*NZZ*) and Austria (*derstandard.at*) were also part of the top 30 content providers. Traditional content providers also put much effort into creating sophisticated promotional networks within Twitter.

⁶<https://devcommunity.x.com/t/announcing-the-deprecation-of-v1-1-statuses-filter-endpoint>

⁷<https://developer.twitter.com/en/docs/twitter-api/tweet-caps>

News-feed accounts and journalists contributed to the distribution of articles and became involved in political discourse. We also identified political blogs (e.g. *Tichyseinblick*, *Journalistenwatch*, *Epochtimes*, *Philosophia-perennis*) with a tendency towards tendentious to extreme content within the top 30. In this context, we also observed that the German political party *AfD* was highly active on social media. Regarding shared links from Facebook (6 out of 10) and YouTube (3 out of 10), *AfD*-related topics dominated this content. However, in general shared external sources from other OSNs (YouTube, Instagram, and Facebook) failed to generate reach and impact within the GTC, with minor exceptions of special events like *BTS* and *Rezo*. Due to the election period, official governmental information providers also received much attention and were referred to by users on the platform.

Regarding the impact of news content, especially, the high number of replies w.r.t. tweets sharing and discussing such content suggests that users are willing to discuss or comment on others' content. The ratio of retweets of shared content (news related: 3.81; others: 2.15) reaffirmed this observation. Statistics on Reaction-tweets depicted a similar picture. News-related content and, especially, news providers triggered many Reaction-tweets. These figures suggest a high interest and participation in news consumption and political discussions. Only self-promotional profiles seemed to fall short of their intended goals, yielding minor to no effects concerning user engagement.

Further, we studied the scale and influence of anti-democratic news content. Thus, we defined the groups of Controversial and Non-Controversial Users. Even though established news providers dominate news-related content within the GTC, actors spreading and supporting controversial opinions are also part of the landscape. Comparing their tweet behavior, we found striking differences. Mostly, established news providers received significant attention from the user base and contributed to a lively discussion culture. In contrast, people who consumed tendentious to extreme politically opinionated blogs were overly supportive. These small blogs, supporting extreme political ideologies, had a small but loyal user base that distributed their content extensively via retweets in the network and inject their content into discussions via replies. Despite their small size (overall 11 129 users), they generate large amounts of tweets. However, information on used hashtags suggests that these users propagate their opinions rather than discuss topics. Due to their high activity, this small group of users is overly influential and visible in the GTC. Our findings contradict the assumption that users with extreme views tend to form closed systems only reaffirming each other's beliefs. Most of their interactions with external political content are responses to Tweets from Non-Controversial Users. They actively engage in many discussions and confront people with opposing views. In summary, we conclude that a similar, overly active, behavior from users of the extreme ends of the political spectrum as reported in the 'Hidden Tribes' study [204] can be observed in the GTC. Interestingly, however, it seems that these users are largely ignored in discussions by the moderate majority of users in the GTC.

Based on our data, we believe that the German Twitter user base consumes daily news and is highly interested in interacting with political content directly on the platform. Our work is a first step towards enhancing our understanding of the German Twitter population. We highly suggest that researchers apply similar methods to conduct their studies on large-scale snapshots rather than small network samples.

5.2 Nunti-Score: Supporting users in the Assessment of News Article Previews in OSNs

Falsehood flies, and the truth comes limping after it.

Jonathan Swift, 1710

So far, our analyses on shared content and engagement showed that political content produces the most activity within the German Twitter user base [337, 401]. The high number of reactions to news-tweets suggests keen interest in such content and that users use Twitter as a platform for political discourse. Thereby, established news outlets were the basic source for discussions. This is also reflected in studies which have determined a recovered and increased trust in traditional news outlets within the German user base in recent years [223]. Only a very small proportion of German-speaking Twitter users distributed extreme controversial news content. In comparison: around a quarter of American adults admitted to sharing misinformation, knowingly or unknowingly⁸.

However, this small group is overly active, engage in many discussions and confront people with opposing views. Thus, also their content becomes visible in the GTC and users are confronted with. Moreover, results only reflect Twitter. In addition, established media have lost journalistic quality due to the speed of information distribution and actuality pressure. Media became dependent on their advertisers, which also influenced their content [359]. News has to sell, thus bad news (disasters and catastrophes) are preferred, which leads to "negativism" [195]. Moderate news outlets that practice subjective or tabloid journalism do not follow unbiased objective journalistic criteria, but explicitly aim to arouse emotions [388]. For example, in the past, several articles with tendentious, disputable content of moderate outlets like Bild or taz were rebuked by the German Press Council. Moreover, even traditional media were observed, e.g. within the German election campaign 2017, to be involved in the distribution of misinformation to varying degrees [434]. Journalists tend to express their opinions more freely on social media [270]. Social media, like Facebook, incentivizing polarizing content which produces user-engagement, rather than according to qualitative criteria⁹.

Thereby an issue of news consumption in OSNs is the representation of external content as short article previews (summary, headline, image). Users tend not to read the underlying information, but only these previews [181]. At the same time, users who only read article previews feel better informed about the content covered than they actually are [21]. Moreover, these linked news previews are mixed with a huge amount of information of all kinds within the users news feed. Thus, presented news information are consumed rather incidental than targeted [49], reducing political learning [178]. In this context, Moravec et al. [323] found that users tend to exhibit a 'hedonistic mindset' when consuming news content in this way, where entertainment is the primary focus and critical engagement with news content is foregone. Overall, even if extreme anti-democratic content is not widespread within the GTC, the difficulty of verification and assessment of social media posts in the abundance of information remains as well as its impacts mentioned in the introduction.

Therefore, in this chapter, we present and examine a possible approach to support users in the assessment of news in OSNs and thus to minimize negative impacts. The idea is to enrich news information in the

⁸www.pewresearch.org/fact-tank/2017/12/28/key-trends-shaping-technology-in-2017

⁹<https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-60-minutes-polarizing-divisive-content/>

user's social media with automatically extracted background information to support them in the evaluation and classification. Based on the gap between utility and usability in daily usage, we decided to prepare such information in a hierarchical structure. The lowest level should thereby visualize the objective background information, whereas the first level is intended to be an abstraction of the information quality of the news information for everyday use, i.e. it indicates how suitable this information is as a news source. This information quality label was named Nunti-Score ("nuntii" latin: news). In the evaluation we mocked up a possible 5-point scale Nunti-Score implementation and evaluated how well such a solution would be accepted by the German-speaking community and whether it has an impact on the perceived credibility of news feed items.

5.2.1 Counter the Impacts of Misinformation - Related Work

On the one hand, there are measures that try to reduce the exposure of misinformation, e.g. through deletion, down ranking, demonetization or the introduction of limits for sharing (see also Sec. 2.2.2). On the other hand, there are measures that attempt to reduce the potential for deception and improve the ability to assess the credibility of information [418]. These methods range from generally raising media literacy [235, 376], over manual¹⁰ and automated fact-checking [505, 397], to contextualization.

Contextualization, i.e. the enrichment of news article previews with additional information, represents an interesting approach to directly support the user in the assessment of its content, instead of invasive manipulation of the users' news feed. A first approach in this category is to display related articles next to news posts. The goal is that users get a better picture of a topic. Especially Facebook was promoting this approach in 2017¹¹. Studies show that this approach is not very effective and could even backfire. The 'backfire' effect appears when correcting content not only fails to reduce the credibility of misinformation, but actually reinforces it [484, 39, 248, 323]. Thus, Facebook switched to approaches based on independent third-party fact-checkers and warning labels¹².

Several studies have shown that warning labels related to misinformation posts can reduce their perceived credibility [323, 248, 84, 353, 236]. In addition, Mena [307] showed that warning labels reduce the sharing intention of such posts. Kirchner and Reuter [248] evaluated different warning-based approaches and concluded that warning flags, combined with an explanation of how and why it is labeled, achieve high acceptance among users and are also most effective in terms of reducing the perceived credibility of false news.

To label misinformation posts as such it is necessary to recognize them. As mentioned above, even Facebook relies on human fact checkers. The drawbacks of manual detection approaches are the need for experts, associated costs, low coverage rate and the slow reaction time. In addition, the use of third-party services and even crowd based solutions carries a potential risk of subjective and biased results, which makes it necessary to resort to several independent services, to reveal a good reflection of the information. Contrary, automated detection solutions allow to cover a wide range of topics, directly. However, these approaches usually cannot explain their decisions why it was detected as false. In addition to the positive effects of warning labels, Pennycook et al. [353] revealed a new problem they called the *Implied Truth Effect*. Merely labeling posts confirmed as false can lead to increased credibility of all unlabeled posts,

¹⁰e.g. <https://www.snopes.com> or <https://www.politifact.com>

¹¹<https://money.cnn.com/2017/08/03/technology/facebook-related-articles>

¹²<https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works>

which could still contain misleading information.

Next to labeling shared external news information as true or false, some approaches try to contextualize their publisher¹³ [42, 243], author¹⁴, or the OSN message content itself [202, 191]. Required information is also extracted either manually (crowd or expert) or machine-based. Besides warning labels, there are approaches which use ratings instead [243, 191]. Kim et al. [243] investigated source ratings, based on human assessments (expert and crowd). They found that such ratings directly affect the perceived credibility of news items and that the mechanism behind the ratings mattered. Cahoots¹⁴ (unfortunately discontinued), does not provide any abstraction at all but enriches directly with information about the author and e.g. his or her affiliations. This way, the reader gets insight into possible influences therefore helps to judge the trustworthiness of the article.

5.2.2 Assisting the Crowd

This chapter focus on contextualization of news article previews in social media. Recent studies showed fact-checking labels and ratings are effective approaches to reduce perceived credibility and distribution of misinformation [248, 243]. However, certain human cognitive biases have diminishing effects [350]. Warnings are not necessarily believed [417], where the acceptance of such approaches varies based on the source of the assessment [243, 484]. In general, a system with a human in the loop, a non-transparent factor, allows individuals to discredit these systems, e.g., as propaganda machines disguising themselves as journalism¹⁵. Transparency and comprehensibility of a corresponding solution are therefore enormously important for a trustworthy use of such an rating of information. We think, that augmenting news article previews in social media with automatically extracted meta information can bridge this gap. Instead of labeling warnings, we propose to use automatic extracted textual properties of the corresponding article and other meta data, which provide indications of its information quality. These can be visualized transparently, and thus insights into the rating can be given. Considering various properties of a linked article allows for finer and more comprehensible statements, instead of validating solely, e.g., its publisher. Additionally, automated contextualization enables a holistic, independent and direct coverage of news content in social media. Thus, issues like the implied truth effect or the slowness of manual procedures are avoided.

The categorization parameters of our TET approach could be find in Table 5.13.

Parameter	Value
Target Audience	data-subject
Data Types	content, author, meta-data
Execution Environment	hybrid
Application Time	realtime
Interactivity Level	read-only
Delivery Mode	push

Table 5.13: Classifying the envisioned TET according to Section 2.3.2.

¹³<https://www.newsguardtech.com>

¹⁴<https://github.com/getcahoots>

¹⁵<https://www.allesaufdeutsch.tv/faktenchecker.html>

5.2.2.1 Automated Context Extraction

Certain background and meta-information can help assessing credibility of an article. Analysing text or media quality of an article can provide useful evidence. A good source of information should follow journalistic quality characteristics. These include for example the independence of the author, source transparency, balance, neutrality, comprehensibility, factuality, topicality, or the placement of facts in a larger overall picture [388].

Text quality of an article can, e.g., be measured by statistics on bias [197], use of hate-speech and abusive language [70, 164, 381], stance [250], the emotion [13], framing [33, 332, 331], (aspect) sentiment [284], click bait [8] or the use of AI for generation [502]. This information could help to automatically examine the balance, neutrality or comprehensibility of the textual content. The analysis of embedded media elements (*media quality*) provides, e.g., the detection of (unmarked) deep fakes [441, 317], synthetically generated images and videos, or the detection of discrepancy between media elements and text content. Perceptual image hashing algorithms (see also Appendix A) could be used for reverse image search¹⁶ to investigate whether images have already been used at another time for another event. Other valuable information for the assessment are, for example, the type of the article (e.g. satire, sarcasm score, opinion, ...) [154, 375], its *transparency* regarding authorship, sources, and publisher, or if the date of publication corresponds to the content of the article [460]. Also conceivable is the comparison of an article with the media landscape, e.g., with regard to the presentation and framing of the content [135]. Factors of factuality (*fake measures*) of the content could also be taken into account, either through automated fake-news detection [321] and/or by querying several fact-checking platforms. Meta data concerning the author and publisher can be aggregated by analysing their whole set of published articles regarding, e.g., text and media quality. While this is an incomplete snapshot of the state-of-the-art in NLP, it shows the amount of background information accessible through data mining techniques. Using speech-to-text algorithms these analysis methods could also be applied for video contextualization [7].

5.2.2.2 Meta Data Visualization

The above context information is useful for assessing credibility, but too complex for intuitive representation in direct interventions. A contextualization that covers it completely would most likely be ignored and suffer from bad acceptance. It hence is necessary to develop a suitable abstraction.

As a first approach, we decided to abstract possible objective contextual information to a subjective rating. The respective context information would be evaluated, accumulated, and expressed by an overall rating. This rating does not label an article as false or true, but rather indicates how suitable its content is as a source of information, i.e. it indicates the *information quality* of the article content. Constructing such a metric — reducing objective criteria to a subjective measure of rating — is anything but trivial, especially in the case of information, and is here treated as future work. In addition, the abstraction to a subjective value is not entirely unquestionable, but is assumed to be important with respect to the discrepancy between utility and usability.

To offer users both an intuitive abstraction, but also the opportunity to investigate the reasons for the algorithm to choose the corresponding rating, we propose a hierarchical approach. This allows transparent insights into the rating algorithm, and disclosing the details shall merit higher trust and acceptance by the

¹⁶<https://tineye.com>

users, than rating based on subjective, human decisions.

The first layer presents the user an abstract score based on the background information in form of an overall rating. The score is rendered directly onto the article preview within the news feed. More information is provided by the second layer of information. The idea here is to divide collected context information into categories which represent the various aspects of information quality. It provides a score for each category of background information. The overall rating of information quality should be comprehensibly composed of the ratings in these subcategories. Here, we mocked up a first example of possible second layer categories:

- Text quality (text neutrality, writing style, clickbait detection, ...)
- Topicality (age of the article, discrepancy between date of publication to the content of the article, ...)
- Media quality ((undeclared) use of deep fakes, contradictions between content and images/videos, ...)
- Transparency (Are details of publisher, author and sources available?)
- Trustworthiness (Have the publisher, author, or sources published false reports in the past? What are their subject areas / expert fields compared to the content? ...)
- Comparison to the media landscape (Is there other reporting on this topic? Are the facts presented similarly there? ...)

Finally, the third layer breaks down the categories into their corresponding automatically extracted contextual information. Overall, it should provide detailed information on all aspects of mined information, the rating mechanism and justifications.

5.2.2.3 The Nunti-Score

We considered the first layer of such a solution to be the most crucial part of the design, especially in terms of everyday use and the impact on the assessment of information. Thus, this study focused on the evaluation of the rating. For this purpose, various mock-ups were designed, with different visualizations, scales, colors and positioning of the rating elements (see Fig. 5.6). Designs were aimed at desktop users based on a Twitter-like user interface. To analyse the perception, comprehensibility, and appearance of the

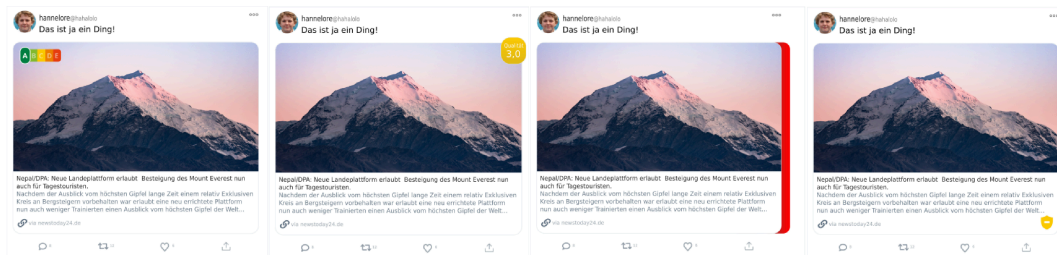


Figure 5.6: Different rating mock ups.

mock-ups as well as to get a sense of user needs, we conducted individual interviews with 7 participants in an age range of 25-32 years. The results were used to design the information quality rating (Fig. 5.7), consisting of a five-point rating scale, coded in color and letters (i.e. green A = best rating, red E = worst rating). The pretests yielded that it is best positioned at the upper right corner of the teaser image. The used color scheme is based on the well-known traffic light system. Overall the final design resembles the European Nutri-Score quality-label for food, whose effectiveness and better readability compared to



Figure 5.7: Final design of the Nunti-Score rating scale after incorporating user feedback

other food labeling systems has already been studied [74]. Thus, we call our quality-label for information *Nunti-Score* (from 'nuntii', latin 'news').

Based on this design, the second layer was mock-uped. Here the idea was, that the user could interact with the first layer (e.g. click, mouse over), if required. Thus, a more detailed breakdown of the overall rating will be presented. For this purpose, an overview of the categories of information quality is visualized and their sub-ratings are shown. For the sub-ratings, the same 5-level scale is used as for layer 1 (Fig. 5.8).

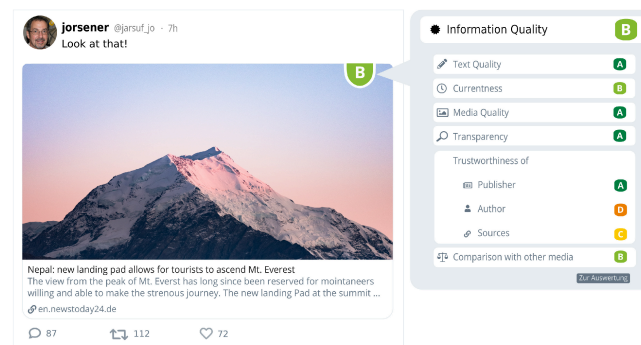


Figure 5.8: The first two layers of the contextualization mock up.

5.2.3 Nunti-Score Analysis - Hypotheses and Experiment

In the following we describe our research questions, as well as our developed quantitative experiment in order to investigate our hypotheses.

5.2.3.1 Research Questions

Before implementing the contextualization, it is important to examine whether such a rating will be accepted. Therefore, we investigated the following research questions: First, we questioned whether participants will use the Nunti-Score rating to assess the content of news article previews? Second, does the use of the Nunti-Score rating influence the perceived credibility and the certainty of the evaluation of the news feed information?

The work by Kim et al. [243] on human based source ratings revealed influences on user beliefs in article previews but also demonstrated that the rating mechanism matters. Thus, we were interested whether the Nunti-Score rating, based on automated extracted background information, influences the perceived credibility of news items. The following hypotheses were predicted:

H1 *Contextualizing a news item with a Nunti-Score rating impacts the perceived credibility of its content.*

More precisely, we presume that the respective score of the rating influences this impact:

H1.1 *The more distinct the rating (i.e. very high or low), the stronger the impact of the Nunti-Score on the perceived credibility.*

Such a finding would be in favour of the Nunti-Score approach, where low ratings (E) would indicate misinformation, while high ratings (A) would indicate good quality information. More ambiguous ratings

would indicate that additional scrutiny on the content is required. The impact of the Nunti-Score should furthermore be influenced by the *difficulty* of the assessment decision of a news item. This possible relationship seems interesting especially with regard to emerging topics. For example, at the beginning of the Covid-19 pandemic, information about this type of virus was rather unknown to the society, i.e. assessing the credibility of news on this topic is more difficult which can result in a higher susceptibility to misinformation. Therefore, the following hypothesis were predicted:

H1.2 *Higher difficulty of the assessment of news items leads to a stronger impact of the Nunti-Score on perceived credibility.*

The second question to be evaluated is whether the Nunti-Score rating of article previews influences the perceived certainty of users in their assessment of news items.

H2 *Contextualizing a news item with a Nunti-Score rating has an impact on the perceived certainty in assessing its credibility.*

Our assumption here is, that the type of rating as well as the difficulty of the decision have an influence on the impact of this effect:

H2.1 *The more distinct the rating (i.e. very high or low rating), the stronger the impact of the Nunti-Score on perceived certainty.*

H2.2 *Higher difficulty of the assessment of news items is associated with a stronger impact of the Nunti-Score on perceived certainty.*

In order demonstrate a benefit for the Nunti-Score rating, we are also interested in long-term effects and whether we can find influences beyond the moment of reading:

H3 *Contextualizing news items with a Nunti-Score rating has an impact on the recall rate of the news items content.*

While one aim of such a contextualization would be that poor quality information would not stick in memory, it could also happen that poor and ambiguous ratings are accompanied by a more critical scrutiny of these information items. Such a manifestation of ambiguously rated information in memory could take place because it is perceived as an unsolved problem, described as Zeigarnik effect [501]. Directly related to the question above, we were finally interested whether the rating of the information is remembered:

H4 *The Nunti-Score rating of an news item is retained when its content is remembered.*

Finding support for H1 but not for H4 would imply that the Nunti-Score rating is only effective at the moment of reading, but in memory only the information is taken away, and later assessment is again made intuitively.

5.2.3.2 Participants

Participants were recruited mainly via the student mailing list of the Technische Universität Dresden, in addition to the dissemination in social media channels. After preprocessing the data, where subjects with very short or very long response times for single items and subjects with continuously the same answer were excluded, a total of $N = 455$ participants remained for further analyses. Of these, 55,6% indicated female as their gender, 43,1% male, and 1.3% diverse. We gathered age in the following groups: 18 to 29 years (86.2%), 30 to 39 (9.7%), 40 to 49 (2%), 50 to 59 (0.9%), and 60 to 69 years (0.9%). The majority of educational attainment were academic high school diplomas (66.6%), followed by university/college degrees (27%). 6.1% have completed vocational training and 0.2% have a secondary school diploma.

5.2.3.3 Stimuli

News feed items were created in order to evaluate the effects of the Nunti-Score rating. In accordance with typical social media posts, these items contained information about the author (avatar, display name, account name), date of the post, social features (likes, comments, shares), an author comment and an embedded news article preview consisting of image, headline, teaser text and source.

Forty article preview contents (headline and teaser) were created, based on rather neutral topics instead of existing news, current affairs topics or polarizing content, in order to minimize the influence of possible user bias. The writing style mimicked news article previews in OSNs and we tried to be consistent across all items.

Of the 40 items, 20 contained true and 20 contained false statements. Furthermore, difficulty was manipulated, where half of the true and half false statements were either of common knowledge (simple to judge) or considered to be difficult. Accordingly, there were 4 categories with 10 items each: *true-simple*, *true-difficult*, *false-simple*, and *false-difficult*. The classification of difficulty was based on the probability for a correct judgement, i.e. for simple items a probability of 80 – 100% for a correct judgment was assumed, for difficult items this probability should be in the range of 20 – 80%.

These assumptions were analyzed in pre-tests. Therefore, 20 participants (m:6, f:14, age:18-29) were presented with all items in random order and had to indicate whether they considered an item as true or false. The analysis for correct categorizations revealed 75%. The high accuracy for the false-difficult items required a revision of the item set. Subsequently, a second pre-test with 16 participants (m:8, f:8, age:18-39) was conducted. As a result of the two preliminary studies, we obtained 24 items, with six items per category (see Appendix B). The selected items met the previously defined probabilities for classification in terms of difficulty (Fig. 5.10). Afterwards, suitable images for each preview were collected¹⁷.

To minimize other potential influences on the perceived credibility of an article preview, the further post content was created as described below: For each news item, a news source was invented that sounds like a typical news domain but did not exist. The author of the social media post (Hannelore), her avatar¹⁸ and her comment was consistent across all feed items. The date as well as the social features (e.g. number of likes) for the posts were varied within a certain range. Hence, the further post content information did not provide relevant information about the credibility of individual items. Two examples of the used news feed items are shown in Figure 5.9.

5.2.3.4 Procedure

The quantitative survey was conducted in February 2021 over a period of three weeks using an online application (Lime Survey). At the beginning, participants were introduced to the general problem of spotting trustworthy news information in OSNs, and to the idea of automated rating of such information. An example was provided to explain the 5-point Nunti-Score rating together with the criteria which would have been used for the automated decision making of the rating (Fig. 5.8).

Participants were instructed about their task; the assessment of the credibility of article previews, some of which were labelled with the Nunti-Score. There was no explicit instruction that the rating should be

¹⁷based on unsplash.com and NASA

¹⁸created with thispersondoesnotexist.com

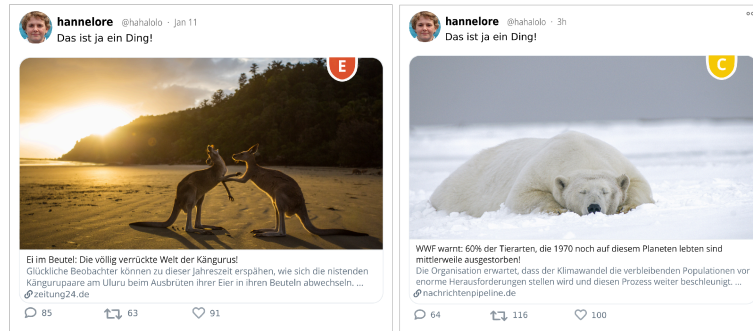


Figure 5.9: False-Simple Feed Item with unambiguous Nunti-Score rating, and True-Difficult Item with ambiguous rating

included in the participants' decisions. In addition an informed consent was given. While completing the survey, participants could not return to the previous item. In addition to the participants' responses, the time to answer to each of the items was recorded. Response data were stored anonymously. The survey consisted of three blocks: *Assessment of News Items*, *Demographics*, *Experiences and Feedback*, and *Memory for News Items*. We describe each block in the following.

Assessment of News Items In this part the effect of the Nunti-Score rating on the perceived credibility of news feed items as well as on the perceived certainty of the assessment was examined to answer H1 and H2.

Therefore, the 24 news items were presented in random order. For each item, participants were asked to indicate on a 5-point Likert scale how credible this information (i.e. agree/disagree) is and how certain they are in their assessment. The answers given by the participants could be compared with the classification of the news items (i.e. true/false and simple/difficult). Accordingly, the assessment of agree/disagree with respect to the credibility of a true/false item, is treated as a correct assessment. The deviation from the correct assessment is then used to measure the impact of the Nunti-Score and is henceforth referred to as *accuracy*.

Overall, three different types of contextualization were considered. For this, two items of each item category were contextualized with unambiguous ratings (i.e. *A* for true and *E* for false items), two with an ambiguous rating (i.e. *C*) and two items without any rating, serving as control group (Fig. 5.10). This resulted in a total of eight items per contextualization type, which are further subdivided into four simple and four difficult items.

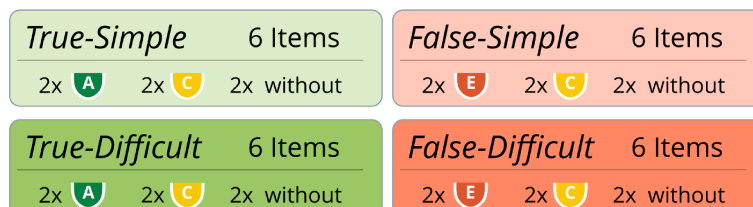


Figure 5.10: News Feed Item Categories and their different Nunti-Score Ratings

In order to measure the effects independently of the item content, participants were randomly assigned to one of three user groups at the beginning of the study. This random assignment resulted in 168 participants for group 1, 141 for group 2, and 146 for group 3. For each group, the items were contextualized differently.

Thus, each item was assigned with each of the three contextualization types (ambiguous, unambiguous, none). An example of this distribution for items of the true-simple category could be seen in Table 5.14. The distribution within the other item categories was exactly the same (Appendix B).

Table 5.14: Distribution of contextualization types across items for the 3 different participant groups, using true-simple items as an example.

	Item 1,2	Item 3,4	Item 5,6
User Group 1	A	C	–
User Group 2	–	A	C
User Group 3	C	–	A

To evaluate the impact of the Nunti-Score on perceived credibility, the answers of the different groups were merged according to the type of contextualization and difficulty. Using the example in Table 5.14, for simple items assigned with unambiguous ratings, the responses of items 1&2 of Group 1, items 3&4 of Group 2, and items 5&6 of Group 3, for both the true and false items, were considered. For each participant, the accuracy (distance to factual correctness) they achieved in assessing the credibility of the items on *average* was calculated. In total, six mean values of achieved accuracy are calculated from four item assessments (unambiguous-simple, unambiguous-difficult, ambiguous-simple, etc.). Two single 3 (Nunti-Score) by 2 (difficulty) repeated-measures ANOVAs were conducted to estimate the effect on credibility and certainty.

Demographics, Experiences and Feedback The second block of the survey served to obtain further demographical information but also to include delay between the first and the last block of the study. The demographic questions included gender (male, female, diverse), age (in increments of 10) and (professional) qualification. Furthermore, there were questions about the participant's use of OSNs. Which social networks does the participant use in everyday life (eight multiple choices, none and other as text field) and on which of these channels does the participant come across news content or linked article previews. Finally, it was asked how often the participant was uncertain about the credibility of article previews in OSNs when reading them in everyday life.

In addition, further insights about the use and meaningfulness of the Nunti-Score rating from the participant's perspective were explored. First, it was asked which features were used for assessment of the article previews. If a participant had not used the rating, he was additionally asked why (multiple choice of given answer options and free text field). Finally, participants should indicate whether (i) the rating was helpful for the assessment, (ii) the design was appropriate, (iii) they understood the visualized rating, (iv) they could imagine using the tool in everyday life, and (v) they would recommend it to a friend.

Memory for News Items The final block was included to analyze whether participants can remember the news items presented in the first block. Furthermore, we were interested in the recall of the credibility. Therefore, participants received twelve statements in random order, each of which is intended to evoke the memory of a specific item. The statements for each participant were chosen in such a way that four of them were in relation to items with unambiguous rating, four in relation with items with ambiguous rating and four were in relation with items without rating. Of these, two are again associated with simple and two with difficult content. They are then asked if they can remember a news item that fits the statement and if so, give an assessment on how credible they would consider the contained information. This was

done in order to analyze whether there are differences between the contextualization types with regard to the recall of information (H3). Furthermore, this allows to examine whether there are deviations in the perceived credibility between the first and second assessment of the items (H4).

5.2.4 Impact of the Nunti-Score - Experiment Results

In this section we report results of our quantitative study.

5.2.4.1 Usage of OSNs and the Problem of Information Assessment

The vast majority of the participants (96.7%, $N = 440$) use several social media in their daily lives (Fig. 5.11). Thereby, they come across news content primarily on Instagram, YouTube, Facebook, Twitter, and Reddit. Additionally mentioned platforms regarding news content were e.g. Whatsapp, Jodel or Telegram. When asked how often they were unsure regarding the credibility of news previews when reading them ($N = 403$), for the vast majority (48.9%) this was the case at least occasionally. One third of the participants even stated that this was often (28.5%) or very often (4.2%) the case. Only about one-fifth of the participants reported that this was rarely (11.7%) or very rarely (6.5%) the case, and only one participant stated that he never felt any uncertainty in this regard. This shows that more than 80% of our participants perceive it difficult—at least from time to time—to assess news content in social media.

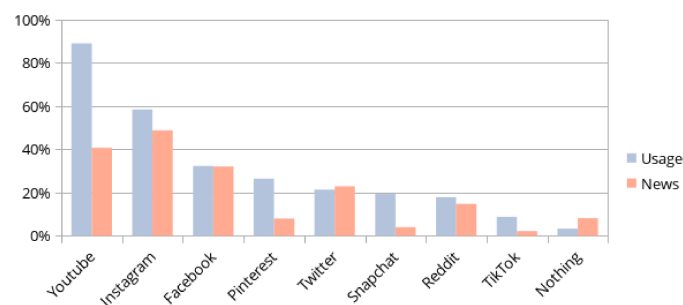


Figure 5.11: Use of social media (blue) and contact with news information (red).

5.2.4.2 Features used by Participants for Assessment

With regard to the features of the news items that participants used to assess the credibility (multiple answers were possible), we obtained the following feedback: Almost all participants (97.5%) stated that they had used the teaser text of the article preview for assessment. Also, a large majority (75.8%) of the participants used the headline of the article and about half of them (48.5%) used the source. Although all other features of the news items were kept almost identical, some participants referred to the author (18.9%) or to the date (5.3%) of the post; only very few referred to the presented social features (shares, likes, comments) for this purpose (1.3%).

"Each article seemed highly implausible to me insofar as 'Hannelore' had commented in each case: 'That's a thing!'. This sentence alone makes me believe nothing more, because this corresponds to my idea of classic clickbaiting."

However, almost half of our participants (47.3%, $N = 215$) stated, that they had used the Nunti-Score for assessing the credibility of the article previews. Accordingly, the usage of the Nunti-Score was considered as between-subjects-factor for the following analyses.

5.2.4.3 Impact on Perceived Credibility

Testing for the impact of the Nunti-Score on participants' ($N = 455$) perceived credibility of corresponding news items revealed significant differences between the three different contextualization types (unambiguous, ambiguous, none), with an estimated effect size $\eta_p^2 = 0.121$, suggesting a medium effect (Tab. 5.15).

Table 5.15: Results for Impact on Perceived Credibility.

Cases	Sum ²	df	Mean ²	F	p	η_p^2
N	31.69	2	15.85	62.25	<.001	0.121
N * U	12.92	2	6.46	25.38	<.001	0.053
D	345.16	1	345.16	1286.27	<.001	0.740
D * U	0.4	1	0.4	1.49	0.223	0.003
N * D	1.09	2	0.54	2.84	0.059	0.006
N*D*U	1.06	2	0.532	2.777	0.063	0.006

N–Nunti-Score, D–Difficulty, U–Usage of Nunti-Score

As expected, for those who reported that they had used the rating a significant effect for the impact of the Nunti-Score was found. There were no differences between the contextualization types among the participants who indicated that they had not used the rating (Fig. 5.12). This indicates the rating was fully ignored and did not produce any subconscious impact on the assessment.

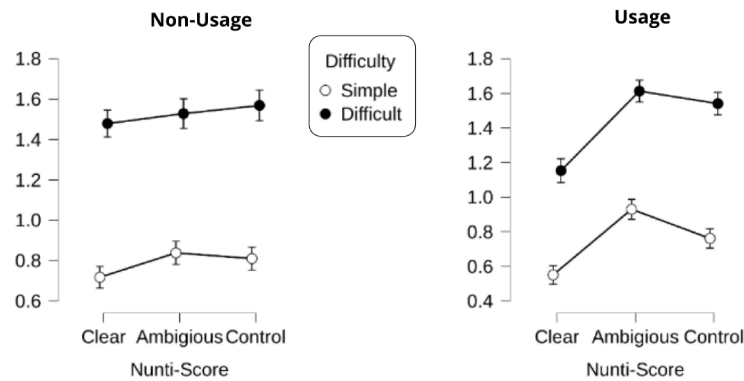


Figure 5.12: Descriptive plots regarding the impact on perceived credibility measured with achieved accuracy

Considering those participants who used the rating, Bonferroni-Holm corrected post-hoc tests revealed significant differences between all three conditions (Tab. 5.16).

Table 5.16: Post Hoc Comparisons - Usage*Nunti-Score*Difficulty - Impact on perceived credibility through different context types for participants who used the rating.

		Mean Diff	SE	t	p_{holm}
Simple					
A/E	C	-0.379	0.046	-8.322	<.001
	None	-0.210	0.046	-4.621	<.001
C	None	0.169	0.046	3.702	0.004
Difficult					
A/E	C	-0.459	0.046	-10.084	<.001
	None	-0.387	0.046	-8.501	<.001
C	None	0.072	0.046	1.583	1.000

Unambiguous Nunti-Score ratings showed statistically significant lower mean scores (distance to factual correctness) than the other two context types. Thus, participants achieved significantly higher accuracy in

assessing item content with an unambiguous (A/E) contextualized rating, i.e. it generates a higher impact on perceived credibility of news items. Overall, this supports our first hypothesis *H1* as well as *SH1.1*. It is striking that a more ambiguous rating (C) results in a slightly reduced accuracy compared to items without a rating. This initially leads to the conclusion that an ambiguous rating of an item results in a poorer assessment of credibility. Our assumption here is that, for difficult items, and especially for simple items on which the user could draw on prior knowledge, participants were unsettled by the rather ambiguous rating, and therefore were less clear in assessing the credibility. Overall, the impact of the rating on perceived credibility was measurable for both difficult and simple items with similar effects. Thus *SH1.2* is not supported by our findings.

5.2.4.4 Impact on Perceived Certainty

The impact of the Nunti-Score on the perceived certainty in the assessment of news items, shows similar results. There is also a significant effect between the contextualization types as well as with regard to the use of the rating (Tab. 5.17). The effect size of $\eta_p^2 = 0.063$ suggests a medium effect.

Table 5.17: Results for Impact on Certainty.

Cases	Sum ²	df	Mean ²	F	p	η_p^2
N	11.46	2	5.84	30.32	<.001	0.063
N * U	5.96	2	3.04	15.77	<.001	0.034
D	80.82	1	80.82	344.08	<.001	0.432
D * U	1.04	1	1.04	4.43	0.036	0.010
N * D	1.58	2	0.79	4.51	0.011	0.010
N*D*U	0.30	2	0.15	0.85	0.428	0.002

N–Nunti-Score, D–Difficulty, U–Usage of Nunti-Score

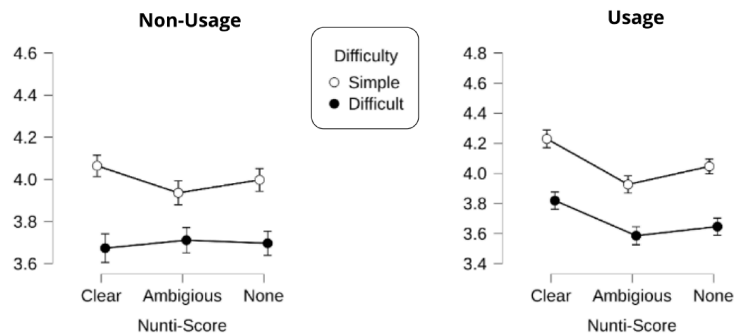


Figure 5.13: Descriptive plots regarding the impact on perceived certainty (5 very certain, 1 very uncertain).

Based on the post-hoc tests (Tab. 5.18) and depicted in the descriptive plots (Fig. 5.13), the influence of an unambiguous Nunti-Score rating has the highest impact on certainty for participants who used the rating. Thus, *H2* and *SH2* are also supported. However, there are no reliable differences between more ambiguous rated items and those which contained no rating. There is a slightly higher uncertainty for ambiguous ratings—similar to the perceived credibility.

Also here, the contextualization had no influence on the certainty of participants who did not use the rating. There is only the exception of simple items with a clear rating. A very careful explanation could be in the direction of a slight subconscious influence (Tab. 5.18). Interestingly, the change in perceived certainty is independent of the difficulty of the items, i.e. the effect is similar for both difficulty levels, with even slight increase for simple Items. This, however, requires to reject *SH2.2* based on the present findings.

Table 5.18: Post Hoc Comparisons - Usage*Nunti-Score*Difficulty - Impact on perceived certainty through different context types.

		Mean Diff	SE	t	<i>p_{Holm}</i>
Simple-Usage					
A/E	C	0.303	0.041	7.371	<.001
	None	0.184	0.041	4.462	<.001
C	None	-0.120	0.041	-2.909	0.091
Difficult-Usage					
A/E	C	0.234	0.041	5.676	<.001
	None	0.173	0.041	4.208	<.001
C	None	-0.06	0.041	-1.469	1.000
Simple - Non-Usage					
A/E	C	0.128	0.039	3.288	0.028
	None	0.067	0.039	1.711	1.000

5.2.4.5 Impact of Difficulty

The difficulty of an item content showed a significant effect on the perceived credibility as well as on the perceived certainty. The effect size of $\eta_p^2 = 0.740$ for accuracy and $\eta_p^2 = 0.427$ for certainty, indicates a large effect that was stronger than contextualization. This demonstrates lower accuracy and higher uncertainty for those assessments where prior knowledge presumably does not exist.

Thus, the classification of the items into their respective difficulty levels seems to confirm a valuable representation of the actual difficulty of correct assessment. Accordingly these items are valid for use in follow-up studies. In addition, the Nunti-Score rating does have an impact on this effect (unambiguous ratings increases certainty and accuracy, ambiguous slightly reduces both), but only to a certain extent. I.e. the own intuition is influenced (strengthened or unsettled) to a certain extent, but without switching it off and blindly following the rating.

5.2.4.6 Effects on Memory

The recall performance of the different items showed no significant differences. Participants were able to remember about 94% of the items regardless of the contextualization type and whether the rating was used or not (Fig. 5.14, left).

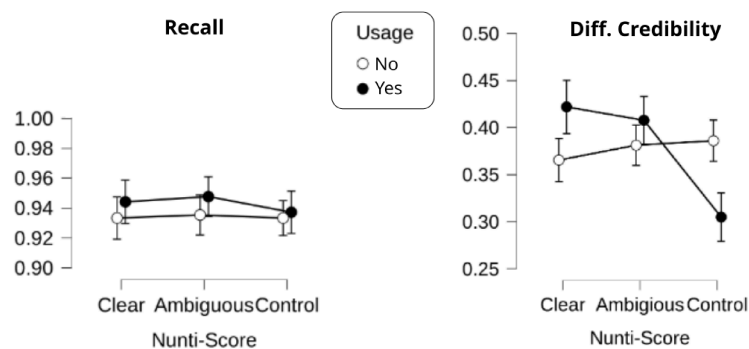


Figure 5.14: Descriptive Plots of, Left: Recall of Information Content (0 no memory, 1 complete memory), and Right: Deviations from the previous item assessment regarding perceived credibility (0: same assessment)

However, this finding could also be attributed to the rather short delay between the presentation of the last item in the first block and the recall in the final block. This assumption is supported by the fact that

participants needed on average about 172 seconds for answering the questions in the second block. Such a short period of only 3 minutes might not be a sufficient interval of time to let participants forget about the information they just have been confronted with.

The perceived credibility of the information content assessed in the follow-up compared to the first assessment also showed no significant differences between the contextualization types ($F = 2.69, p = 0.069, \eta_p^2 = 0.009$). However, there was a small effect regarding the use of the rating ($F = 4.464, p = 0.012, \eta_p^2 = 0.015$). The descriptive plot (Fig. 5.14, right), shows the difference especially between unambiguous rated and none rated items (Mean Difference=0.117, $t = 3.324, p_{holm} = 0.014$). For the participants who used the Nunti-Score, the perceived credibility of items without rating corresponded more to the previous decision, whereas the items with rating corresponded to a stronger deviation afterwards. This could be an indication that the rating is effective at the moment of reading, but that the participant's own intention is used after remembering the information content. However, the question then arises why participants who do not use the rating do not arrive at this control level.

Overall, the influence of information contextualization on recall performance and specifically on decisions about the credibility of information in the aftermath should be investigated in depth in subsequent studies.

5.2.4.7 Reasons for Non-Usage

In order to investigate why more than half of the participants ($N = 240$) ignored the Nunti-Score in their assessments, they were asked why this was the case. Multiple answers were possible. They stated that they did not perceive the rating (50.4%), found that the rating was not trustworthy (39.6%), or had not understood its functionality (9.6%). The former may have been due to errors in the study design, like a too short introduction of the contextualization or an obstructive presentation of the survey on mobile devices, or also on an unbecoming placement and/or visualization of the Nunti-Score label, which requires further investigation in future work. However, some of the participants provided specific reasons why they not used the rating for assessment. This feedback included more details for the selection criteria, e.g. visibility ($N = 6$), or the lack of trust in the rating ($N = 5$). Some participants ($N = 17$) wanted to form their own opinion and so deliberately ignored the rating in order to decide independently and without influence. For instance:

"I didn't want to let my assessment be influenced by the tool and form my own opinion."

One participant followed the Nunti-Score only for negative ratings and his own opinion for positive ones. Three participants stated that they had decided habitually

"I assessed credibility based on the image and caption out of habit, [...]"

and two indicated that they were not entirely sure that they had used the rating subconsciously.

Further concrete reasons were found in the evaluation of the final feedback of the survey. For 19 participants the rating was too unobtrusive. However, the main issue mentioned was the not yet existing trust towards Nunti-Score and therefore especially the condition/wish for transparency ($N = 41$), i.e. a presentation of a comprehensible justification for the rating. Additional, there was also a general rejection of information ratings, even with transparent decision-making ($N = 7$).

"Such tools are fundamentally untrustworthy as you can simply use them to push your own agenda and opinions as is already attempted with other aids of this type."

"[...] it would ultimately end up as a means of influence/censorship. Similar to the 'fact checkers' on Facebook and co. In the end, there is always an agenda behind it, even if it is well-intentioned. Better an unrestricted freedom of opinion than doubting the maturity of the citizen."

To be honest, I consider it, [...] a paternalism of self-thinking people."

Two participants had specific concerns about automated decision making:

"Such a tool cannot possibly assess the seriousness of articles in an automated way."

"Automatisms and algorithms have their limits. [...] I am skeptical that this tool can be even halfway trustworthy without continuous human intervention."

Some ($N = 7$) also described the fear of misuse and possible implications (e.g. who is responsible for the ratings, articles might try to "please" the rating, how to ensure that ratings are not faked, or users might get lazy in assessments). There was also general criticism of using social networks for news consumption ($N = 4$). One participant stated to have lost confidence in the Nunti-Score when there was a discrepancy between the rating and his prior knowledge. In addition, few participants ($N = 6$) stated that they did not use the rating because they already had enough prior knowledge about the topics and that it was therefore not necessary to use it.

5.2.4.8 General Valuation of the Nunti-Score

With regard to the concrete perception of the Nunti-Score (Tab. 5.19), the perceived benefit of the rating showed a rather neutral estimation, with a slightly positive tendency.

Table 5.19: Overview of the participants' perception of the rating. (1: strongly disagree, 5: strongly agree)

	ø All	ø Rating-Used
Rating is helpful in assessing the items	3.27	3.82
Design fits well into the user interface	4.11	4.40
No problems understanding the rating	3.92	4.28
Can imagine to use it daily	3.62	4.08
Can imagine recommending to a friend	3.47	3.87

In addition, participants also showed a rather neutral attitude with regard to a potential use or recommendation of the rating. However, the potential willingness to use such a contextualization solution is high in the group of participants ($N = 215$) who have used it within the study. Despite partial criticism of the visualization and perception, the Nunti-Score was perceived as well integrated and the rating as well understood.

5.2.5 Concluding Remarks & Discussion of the Nunti-Score

To mitigate the impacts of misinformation, we propose an approach that enriches news article previews in social media with a rating of its information quality. Its called Nunti-Score and is based on automatically extracted background information. This meta data contains impartial information targeting the article content (e.g. text or media quality), the author, and publisher.

The results of an online experiment with $N = 455$ participants indicates that the Nunti-Score influenced participants' beliefs in article previews. In particular, more distinct ratings (i.e. very high or low) impact the perceived credibility and increase perceived certainty in these decisions. This effect takes the intended form: news articles with information quality rated as poor are viewed more critically, while articles with good ratings are perceived more positively. More ambiguous ratings (medium information quality), on the other hand, tend to unsettle the reader. However, this is a desirable effect, because it potentially leads to more intense confrontation with an article preview before a clear assessment is made about its content.

On the other hand, too frequent confrontation in daily use could also overwhelm the user and thus avoid the use of the rating. This needs to be investigated in future work. Interestingly, the observed effects are similar for both simple and difficult decisions. Considering the difficulty of decision, it shows that the Nunti-Score can not eliminate the reader's uncertainty. For difficult decisions, the assessment of the credibility of information is more difficult and vice versa, independent of the rating. Overall, this shows that the Nunti-Score serves as a support for the reader's decision, but does not cancel out the reader's own intention. The fact that our participants did not follow blindly the rating is a positive effect, especially when possible errors in the rating or potential for abuse are taken into account. As one participant also noted, however, this effect could diminish over longer periods of using such a rating system, which calls for further investigation.

As expected, these effects are only present if the rating is accepted and the reader engages with it. The results show overall a potential acceptance to use such a contextualization solution, which is, however, tied to conditions. Comprehensibility and verifiability are important for participants in order to be able to trust such a solution and thus weave it into the assessment process. This shows that the idea of a transparent, hierarchical preparation of the automatically extracted meta data and its rating process is a reasonable approach, which, however, should be investigated as a whole in further studies. The desire for transparency is an important fact for all contextualization approaches and is also supported by the results of Kirchner et al. [248], whose warning signs in combination with an explanation, achieved the best effect. However, some of the participants generally reject a rating or labeling of information, whether transparent or not, and perceive it as manipulative rather than supportive.

6

Discussion and Conclusion

The ongoing digitization of all areas of life has revolutionized the way information is exchanged. Nowadays, almost anyone can easily receive, create, modify, and distribute information with almost unlimited reach. The resulting democratized flow of information and the elimination of spatial separation and temporal boundaries opens up new opportunities for individuals and society, but also presents new challenges. On one hand, the unlimited amount of digital information available and the simultaneous creation and dissemination of misleading, false, influencing, and malicious content, aggravates the assessment of received information. This results in an increased need for approaches to verify and evaluate the credibility of information. On the other hand, the interactive nature of digital information exchange makes information flows observable and exposes a lot of sensitive information. Thus, an increased need for data-protection mechanisms arises. However, technical and legal solutions cannot solve these issues completely. In fact, the technical solutions to both problems are even opposed to each other. The more data is disclosed, the more possible surveillance; the less, the less control over shared information. Hence, in practice, new and old technological developments and their systemic measures are always subject to negotiation processes between safety, freedom, and utility. Transparency and education concerning the consumption and unconscious disclosure of digital information and thus increased awareness of end users is, therefore, an important contribution. On the one hand, to fill the gaps of systemic measures and, on the other hand, to empower users and societies to (co-)determine the negotiation processes themselves - and thus to counteract new power asymmetries as well as to become part of the solution.

In this work, we investigated possible gaps for different use cases, developed transparency solutions and concepts for identified gaps, and partly evaluated the desire for, as well as the impact of, transparency. First, we had a look at *traditional information exchange of printed documents*. We analyzed how technologies for verifying analog documents have changed as a result of digitization. Overall, we found, in particular, one practical used method which strongly reflects the gap between security and privacy. With so-called machine identification codes or *yellow dots*, information about the printing process and the exact source device are embedded into almost every color laser printout while printing, invisible to the human eye. What is a valuable tool from a verification point of view, e.g. to protect the reliability of printed

documents or to track criminals, the integration of the exact source device is a severe limitation to the privacy of users. Overall, however, this approach is non-transparent and partly unknown. Thus we created a data set of 1515 printouts, investigated embedded yellow dots, and developed an extraction algorithm. Overall, we found 6 different tracking dot patterns and decoded the pattern structure as well as partly the integrated information. Based on our investigations we developed a transparency tool called *deda*. Transparency, on one hand, for verification of information and, on the other hand, for privacy to make transparent, which information the own printer discloses. In addition, we developed an anonymization approach to defeat arbitrary tracking and embed it within the *deda* toolkit. Finally, with a brief user survey, we found that participants, in general, are unaware of these techniques, see an intrusion into their privacy by their existence, and overall mentioned a high desire for transparency.

Second, we investigated the field of new digital developments, with a focus on the IoT. Especially the integration of small sensors into any physical objects is increasingly blurring the boundary between the digital and analog worlds and letting the information exchange disappear more and more unconsciously in the background.

On one hand, we looked in detail at the change in mobility, esp. *connected driving*. Future vehicles will regularly broadcast data such as position, speed, or direction to all receivers in the vicinity to enable coordinated vehicle movements, and overall ensure safer and more efficient road traffic. Pseudonymization of this communication should thereby enable all positive opportunities, without enabling vehicle tracking, i.e. via privacy by design. Thus, we analyzed a pseudonym change strategy that has a good chance to be included in a future European standard, regarding how it solves the security/utility vs privacy gap. By simulating a realistic urban traffic scenario within Luxembourg and applying, and attacking the pseudonym change strategy we could evaluate its effectiveness. Our results suggest that the introduction of connected driving, even with the European pseudonym scheme, enables the tracking of vehicles, and thus will decrease location privacy in the future. Overall, the gap between security and privacy is weakened but not solved, which should at least be communicated transparently to the user.

On the other hand, we investigated the issue of *bystanders* who unconsciously disclose data through surrounding recording sensors, like cameras in smart vehicles, without even being an active part of the system. Currently, it is nearly impossible for such an uninvolved user to identify surrounding smart devices and corresponding data handling, which leads to foreign control of the bystander's recorded data. Thus, we developed and prototypically implemented a transparency concept for bystanders, to enable insights into surrounding, audiovisual smart devices in everyday life and their privacy implications. Further, we conducted a qualitative semi-structured interview to analyze the desire for transparency as well as the acceptance and usability of our solution. We verified that there is a high demand for transparency, which, however, is very individual and context-dependent. The in-depth interviews uncovered that this desire related primarily to a feared loss of control, misuse of recorded data, and abstract negative consequences. Moreover, our solution helped to satisfy the bystanders' desire for transparency.

Third, we investigated changes in information consumption due to digitization. In particular, we focused on news consumption within social media and the simultaneous spreading of misinformation, campaigning, and populism (freedom vs safety gap).

First, we investigated the situation within a comprehensive population with its political system and media

market, the *German-speaking Twitter community*. Collecting the entirety of German tweets, we analyze what types of content are disseminated and what types of actors are involved. Our data covers ~ 77 million tweets from ~ 7 million users, collected between the 2nd of April and the 2nd of June 2019, involving the European Parliament election. It contains observable artifacts corresponding to major political events and exposes clear, artificial dissemination structures around news outlets and populist parties from the political right wing. The results also indicate that not only political actors but content providers, too, have established highly influential profiles that are heavily engaged in political discourse.

Finally, we developed a transparency concept for the assessment of information, which enriches news article previews with context information. It visualizes the information quality of linked news articles with a rating, named *Nunti-Score*, which is based on automatically extracted background information. We investigate the utility and acceptance of this approach. Based on results from a quantitative online experiment with 455 participants, we obtained indications that users were better able to judge the credibility of news articles in OSNs with higher certainty. Additional feedback confirmed that transparency and comprehensibility of the rating were fundamental for its acceptance.

Overall, in all use cases examined, neither technical solutions nor legal foundations can in practice provide complete protection against malicious information or the disclosure of personal information, but are subject to negotiation processes. Negotiation processes between safety (safeguarding digital systems, law enforcement, verification of information, prevention of malicious information), freedom (privacy, freedom of speech), and utility (ensuring that services function sensibly and effectively). Due to the gaps in systemic measures, we discovered for all three use cases a high desire and interest in transparency solutions and our concepts. Nevertheless, transparency was also tied to conditions. Important indicators for the use of a transparency system were, in particular, trust and verifiability of transparency information. This is overall a very important property of transparency solutions, as they have also the potential to be used manipulatively. It was also important that the transparency solution interfered as little as possible with their daily routine (with regard to realtime TETs). Both points reinforce the decision of a hierarchical structure of a transparency solution. On the first level, a quick insight into the situation can be provided through abstraction without being too distracting, while at the same time offering transparency concerning the decision criteria and its information origin. In addition, the desire for transparency was very individual and context-specific, which needs to be evaluated and incorporated more precisely in future developments. Overall, it appears that users in both problem areas want to make the decisions themselves within these gaps and desire transparency to support those decisions, but only in certain situations. In all three use cases, there was also a part of participants who tend to reject transparency in general. Either because they perceive that they are not affected by the gap, or, especially in the area of information consumption, even perceived transparency as paternalism. Nevertheless, as mentioned above, overall there was a high level of interest and desire for transparency in all of the use cases examined. We received positive feedback regarding the developed solutions and measured positive impacts of the transparency solution regarding the consumption of news information.

In the overall context, however, any kind of digital information exchange in principle leaves these gaps. It can be strongly assumed that with future developments, such as tactile internet, virtual¹ and augmented

¹e.g. <https://about.meta.com/what-is-the-metaverse>

environments, and smartening everything, the abundance of such gaps will increase. In addition, in an increasingly complex world, where technical and regulatory principles can not fully solve emerging problems, there are a lot of additional gaps in which transparency is also propagated as a solution. For example, transparency and labels regarding the consumption of food, energy consumption, climate change, animal welfare, or trade chains to assess how and under which circumstances goods were produced. This gives users abstracted, transparent insights to support their decisions. But they have to decide for themselves, based on their world views and beliefs, what they consume and use and what they do not. Even though in our use cases a high desire for transparency and partly also its positive impacts were found, in the overall complex and in symbiosis with all other areas the question arises: how much transparency can users tolerate and process? How to prevent fatigue towards warnings and transparency information in the abundance of decisions as well as a general loss of trust towards e.g. online content and thus resignation towards the actual problem? I.e., what are the impacts of transparency solutions in the long term? Is the abundance of information useful? Or, shouldn't we create legal and technical preconditions and design systems in such a way that users are protected from decisions and challenges?

However, the author questions whether the development and design of such systems is possible at all, especially concerning privacy and information consumption. If at all, such systemic development in democratic systems should only emerge from social discourse. However, social debates and discourses on the design and development of such systems can only take place if the problems are known and understood. From the author's point of view, transparency and awareness solutions are important tools to enable these discourses in the first place, to fill gaps that have arisen, and to empower users and societies to (co-)determine the negotiation processes themselves. However, these solutions should not only be considered on a small scale, but also in their entirety and impact - in an interdisciplinary network, e.g., with computer science, sociology, technology assessment, psychology, or the law, as well as integrated into the social education system².

On a small scale, the author would hope that transparency and awareness will strengthen the mental models and literacy of users regarding digital information exchange, thus can concretely assess between risks and benefits, meaning making real informed decisions, and thus becoming more robust regarding the issues.

²<https://www.social-web-macht-schule.de/>

7

Thanks

This work also marks the end of a wonderful and eventful period of my life. Now I am facing new tasks and challenges, for which I am prepared thanks to my studies and my doctorate. The successful mastering of this time, with its ups and downs, I owe not only own achievements, but above all also many wonderful companions. Therefore, I would like to take this opportunity to thank Thorsten Strufe and Stefan Köpsell in particular, who made this time possible in the first place and always supported me with valuable expertise, advice, and affirmation.

My thanks also go to the many collaborations, of students, staff and professors, who accompanied me during this time and also participated in making this work possible. In addition to the names mentioned in the introduction, Jan Reubold deserves special mention. My special thanks goes to my family, my wife, for the constant company and support and beyond that, my parents, grandparents and my brother, who made this path possible and also always stood (and stand) by me in all problems and last but not least, a big thank you also to my children, who give me strength and joy in life and always remind me of the important things in life.

- Thank You -

8

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. This paper was not previously presented at another examination board and has not been published in this form. I am aware that a false statement entails legal consequences.

Name

City, Date, Signature

APPENDIX I

Appendix to Analog Information Exchange

A Benchmarking Toolkits for Perceptual Image Hashing Algorithms

The progressive digitalization of all areas of life and the associated accumulation of large amounts of data require tools that make such amounts of data efficiently and quickly searchable, comparable and analyzable. In the multimedia area, such tools extract certain features from a multimedia object that can be used to describe the object or its content. This feature extraction is used to reduce the sheer size of multimedia objects, to prevent redundancy and noise, as well as to be stable against changes such an object undergoes during its lifetime. Working with extracted features instead of the multimedia object itself allows a faster and more efficient processing of large amounts of data. However, by reducing the information to features, the decision whether a multimedia object resembles another or if it is the content to be identified is no longer unambiguous but depends on probabilities or thresholds. Thus, the quality of these tools is measured on the one hand by the robustness of the extracted features to changes of the multimedia object and on the other hand by their sensitivity to different multimedia objects.

An example to illustrate is *Perceptual image hashing* (PIH). PIH, even called image signature, robust image hashing, soft hash or robust image fingerprinting, is an umbrella term for hash functions which produces a short and fix-sized fingerprint out of an image file of arbitrary size. Contrary to well-known cryptographic hashes like SHA-1 or MD5, PIH generate the fingerprint not on the basis of the binary representation of the file but on the basis of the perceptible content of the image, e.g. the structure of the scene. This is necessary because the digital representation of an image undergoes a number of modifications during its lifetime. Many of these changes, named *perceptual preserving transformations* (PPT), leave the perceptual content of the image untouched. These include geometric transformations, like rotation, cropping or scaling, pixel value transformations, like compression, brightness- or contrast adjustments as well as slight alterations such as adding text (e.g. memes) or watermarks. While cryptographic hashes produce completely different fingerprints when only one pixel of an image changes, PIH methods try to ignore such transformations. In addition, the hashing of the perceptual content makes it possible to compare the

generated fingerprints. The more similar the perceptible content of images, the more similar the generated fingerprints. With distance analysis of the hashes, duplicates or similar images can be found. Such PIH algorithms should be both robust against PPTs as well as sensitive, i.e. avoiding mismatches.

The range of applications for such algorithms is large. From reverse image search¹, to image authentication [435], object tracking [153] and digital forensics, e.g. detecting copyright infringement or illegal image material like extremist imagery, child [419] or revenge pornography² up to and including phishing website detection [232]; i.e. to verify information.

So far, there are many different PIH approaches. Each of these algorithms allows different PPTs, but often solve only one aspect such as an affine transformation with high performance. The evaluation of these algorithms vary widely, usually considers only a partial area of robustness, and are only compared with algorithms of similar design. In addition, combinations of transformations such as those occurring in a print scan scenario and the sensitivity characteristics are rarely taken into account.

Existing Benchmarking solutions which try to solve these issues, like PHabs [503] or Rihamark [500], are mainly outdated. For this reason we propose our benchmarking toolkit *MIHBS* (Modular Image Hashing Benchmarking System) with which existing and new PIH algorithms can be evaluated and compared. The main focus is on the evaluation of robustness regarding PPTs and sensitivity of the algorithm as well as in terms of application datasets. Example evaluations with common hashing algorithms on both benchmarking tests and also on the basis of a print-scan dataset will be presented. Furthermore we improve the benchmarking to a second toolkit, named Twizzle, to enable the evaluation of algorithms for semantic comparisons of multimedia object pairs, independent of their purpose, feature extraction and decision making.

A.1 Perceptual Image Hashing

Hash functions are functions that map an input of any size to an output of a short and fixed size. Compared to conventional cryptographic hashing methods, PIH algorithms produce a fingerprint of the perceptible content of an image. Such perceptual hash function H must fulfill three characteristics in particular [320]:

1) *Robustness*: $P(H(I_1) = (H(\tilde{I}_1))) \approx 1$

where \tilde{I}_1 is a perceptual similar image to I_1 and P is the probability

2) *Sensitivity*: $P(H(I_1) = (H(I_2))) \approx 0$

where I_2 is a perceptual distinct image to I_1

3) *Equal distribution of hash values*: $P(H(I) = h) \approx \frac{1}{2^l}, \forall h \in \{0, 1\}^l$

where h is a l -bit binary hash value

The general workflow of a PIH approach could be seen in Fig. I.1. To reach robustness against different transformations the image would be firstly preprocessed. The preprocessing steps used are, for example, scaling to a fixed size to ignore scaling transformations and reduce complexity [494], grayscaling [104], apply smoothing filter to ignore noise [431], log polar- [438], ring partition [439] or radon [486] transformation for rotation robustness or segmentation for cropping resistance [419]. Subsequently, perceptual features are extracted from the preprocessed image. Examples are methods based on Wavelet

¹<https://tineye.com>

²<https://www.theguardian.com/technology/2017/nov/07/facebook-revenge-porn-nude-photos>

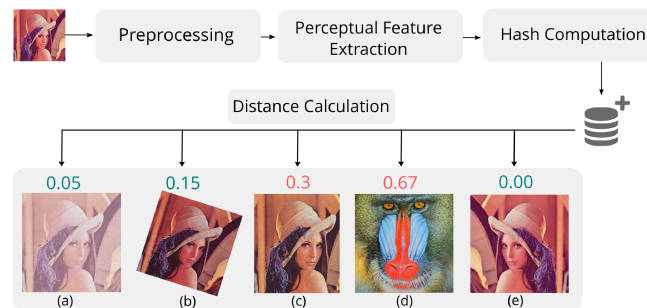


Figure I.1: General workflow for perceptual image hashing. (a-c) print-scan images (a) toner-save mode (b) rotation (c) non-uniform crop (d) different image (e) flip

coefficients [462, 266], Fourier-Mellin coefficients [431, 428], matrix invariants [105, 318] or Feature Patterns [104, 320, 319]. Finally the extracted features are quantized [320, 438] and represented mainly as binary string of fixed length. In order to examine two images for perceptible similarity, their generated hashes are subjected to a distance comparison. In most cases this will be done by hamming distance calculation. The threshold up to which degree an image is recognized as literally equal is rarely given and varies between the different methods.

A.2 MIHBS - Benchmarking PIH Algorithms

Since none of these algorithms currently fullfills the requirements of characteristic 1) and 2) perfectly and all use their own evaluation, we propose our benchmarking system *MIHBS* to compare and evaluate such algorithms regarding *robustness*, *sensitivity*, equal distribution of hash values as well as specific *application* datasets. Currently seven hashing approaches are integrated. These are four simple but efficient and often mentioned algorithms named ahash, dhash, phash and whash³ [153, 152] and blockhash [494]. Further we implemented an approach based on radon, wavelet and FFT transformations, emphasizing the strong stability against a print-scan application (radonhash) [486] as well as a histogram-based approach (hishash) [491]. The toolkit is written in python3 with multithreading support and uses the opencv, scimage and numpy libraries⁴. To integrate another algorithm it must be provided in python, with an image as input and a binary hash as output. Overall each testrun is freely configurable (parameters, algorithms, datasets). After generating all hashes and distance calculations within a configured test, results are saved in a SQLite database and are accessible via a provided analyser using pandas⁴. Thereby results could be simply plotted and evaluated. A complete documentation, the source code and evaluations are available at dfd.inf.tu-dresden.de. In the following we describe the three test fields with sample evaluations in detail.

A.2.1 Robustness

The first benchmark is used for evaluating the concrete robustness of an PIH algorithm against a specific PPT. Input for this is a freely selectable image folder. In our test case we used ten images with different sizes including well-known benchmark images like lena or baboon⁵. Transformations which are currently evaluable can be found in Table I.1.

³<https://github.com/JohannesBuchner/imagehash>

⁴pandas.pydata.org / opencv.org / numpy.org / scikit-image.org

⁵sipi.usc.edu/database

Table I.1: Supported Transformations

Class	Transformations
Geometric	<i>Scaling</i> (uniform/non-uniform)
	<i>Cropping, Shearing</i> (uniform/non-uniform)
	<i>Rotation</i> (with/without rescaling)
	<i>Flipping, Shifting</i> (vertical/horizontal)
Pixel Value	<i>Contrast, Brightness, Gamma</i>
	<i>Noise</i> (Gaussian, Speckle, Salt& Pepper)
	<i>Filtering</i> (Gaussian, Median)
	<i>Compression</i> (JPEG)
Alterations	<i>Watermarking</i>
	<i>Paper-Blending</i>

Each transformation could be de- or enabled as well as combined with another and evaluated as cascade. Further the parameters for each of these transformations are individually adjustable and own transformations can easily be integrated. The general workflow of this benchmark could be seen in Fig. I.2. After generating the hash of a test image, the benchmarking system would apply each enabled transformation step by step onto the image, create the hashes and calculate the hamming distance between the transformed and the original image.

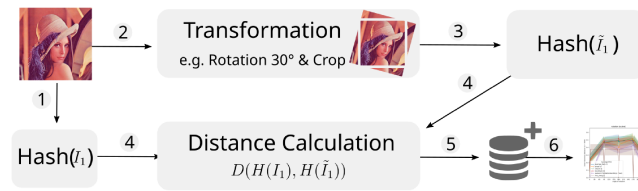


Figure I.2: Workflow of the robustness benchmark, step 2-5 are repeated for each transformation and its parameters (e.g. 360 times for rotation in 1° increments)

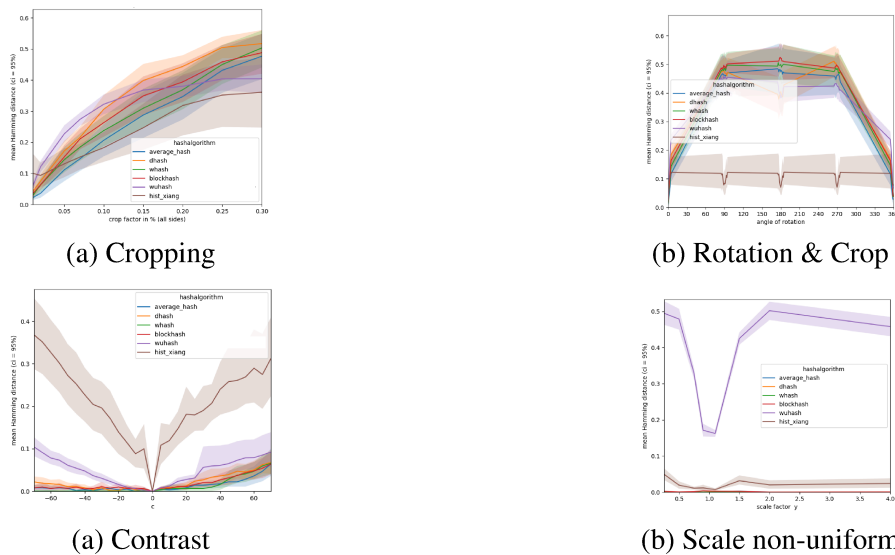


Figure I.3: Sample evaluations for the Robustness Benchmark

To evaluate the general robustness, the benchmark calculates the average stability out of the results from a set of images. A perfect hash algorithm would generate a hamming distance of zero for all transformations.

Fig.I.3 shows example evaluations regarding uniform crop (1 to 30%, 9 steps), contrast alteration (-70 to +70 in increments of 5), a non-uniform resize ($height * a, a \in [0.25, 4.0]$, 10 steps) and a transformation cascade of rotating ($0, 90, 180, 270 \pm 5^\circ$ in increments of 1) and subsequent cropping (see Fig. I.2). The colored areas are the confidence interval of .95.

The results show that with regard to the transformation cascade only the histogram approach is approximately stable with a distance threshold of ≈ 0.1 (fig. I.3 (b)). However, this is the only algorithm that shows weakening in contrast adjustments. With regard to cropping, all algorithms tested exhibit very low stability. For non-uniform scaling, the algorithms, with the exception of radon hash, are close to resistance. One reason for the susceptibility of radon hash regarding scaling is the lack of integration of a resize step in pre-processing.

A.2.2 Sensitivity

The second benchmark is used for evaluating the sensitivity of a PIH algorithm. Therefore a larger dataset with similar but distinct images is needed. Here we used one with $a=20,580$ dog photos [242]. Generally our algorithm would choose a random seed of 1% out of the dataset and compare these with all other images. Here we choose one image set of one dog breed ($b=170$ images) to prevent real duplicates. Overall this results in $b * (a - b) \approx 3.5m$ hash comparisons.

With this setup one can evaluate the best threshold for sensitivity, i.e. the highest threshold at which no image is assigned to another one (min in Tab. I.2). Furthermore with this benchmark one can analyse the equal distribution of hash values. From a certain dataset size on, the normalized mean hash distance should level at 0.5. Results could be seen in Tab. I.2 as well as example plots in Fig. I.4.

Table I.2: Sensitivity Results (p - percentile)

PIH	min	max	mean	p25	p75
ahash	0.016	0.984	0.498	0.422	0.578
block	0.055	0.945	0.499	0.438	0.563
dhash	0.078	0.891	0.497	0.453	0.547
hist	0.018	0.959	0.474	0.349	0.598
phash	0.000	0.688	0.303	0.250	0.359
radon	0.205	0.735	0.475	0.435	0.515
whash	0.000	1.000	0.499	0.406	0.563



Figure I.4: Sensitivity plot for blockhash (left) and radon hash

The results show that the algorithms selected are not suitable for a larger data set with images of similar structure and content. An exception is the radonhash which provides acceptable results. The distribution of hash values is balanced at all but phash. This is either due to the algorithm itself or its implementation.

A.2.3 Calculate the Threshold

In most cases, the threshold, which indicates up to which hash distance images are declared to be similar, is not specified. Out of both benchmarking tests one can now calculate the best threshold t for an application suited algorithm so that characteristic 1) and 2) are fulfilled as best as possible.

$$T(d) = \begin{cases} I_x \text{ ident } I_y & \text{if } d < t \\ I_x \text{ diff } I_y & \text{if } d \geq t \end{cases}$$

with Images I_x, I_y and Distancefunction $D(H(I_x), H(I_y)) = d$. If, for example, one needs rotation stability up to 5 degree, scale resistance and a high sensitivity, he can benchmark these attacks, find the most suitable algorithm and calculate the distance threshold for the best balance between sensitivity and stability dependent on his application.

A.2.4 Application Test

The third bechmark enables the complete evaluation of robustness and sensitivity regarding a labeled application dataset. In our example evaluation we focused on a print-scan scenario, a special transformation chain that is rarely considered. In this scenario, a variety of PPT are applied to the image, such as pre-processing steps (scaling, rotation through landscape print, etc.), halftoning, color model changes or alterations through different print and scan settings (Fig. I.1) Additionally there are potential PPT through

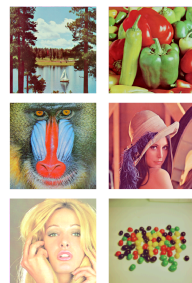
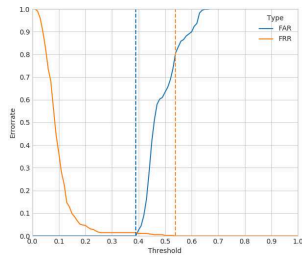


Figure I.5: Test pattern for PIH application analysis.

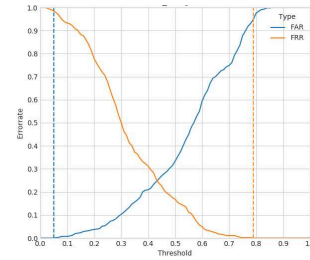
imperfections of each device in the chain, e.g. warped paper feed, dust in scanner or blocked inkjet nozzles. To evaluate this scenario we produced a test pattern with six different well-known benchmarking images (Fig. I.5). This pattern was printed by five printer devices (two inkjet, three EP) with different print settings (various resolutions, toner-save mode, etc.). The prints were scanned by two feed scanners with different resolutions. Overall after deskewing and cropping we got 564 labeled images. Each original thus has 94 perceptual transformed duplicates. To evaluate stability, each original is compared with the hashes of its transformed images, resulting in 564 hash comparisons. For the sensitivity, each original is compared with a random seed of 94 images which are not assigned to it. Averaged results can be seen in Tab. I.3 and example plots in Fig. I.6. Concerning the equal error rate (ERR) blockhash, dhash and whash obtain the best results. However, even if an error rate of 0.7% sounds low, applied to a larger dataset, e.g. with 1m images, an average of 7000 images are wrongly or not allocated. It is interesting to note that radonhash, which is said to be particularly resistant to this kind of transformation, does not stand out.

Table I.3: Application Results (t-threshold at EER)

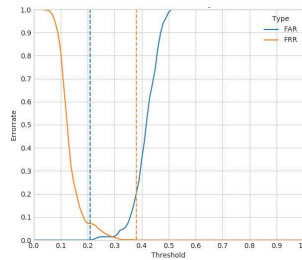
PIH	FAR=0	FRR=0	EER / t
ahash	0.328	0.531	0.013 / 0.35
blockhash	0.391	0.531	0.007 / 0.39
dhash	0.359	0.484	0.007 / 0.37
hist	0.057	0.786	0.250 / 0.43
radon	0.220	0.378	0.017 / 0.28
whash	0.375	0.531	0.007 / 0.38



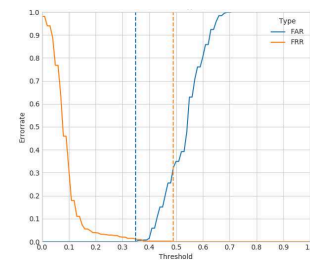
(a) Blockhash



(b) Histogram Hash



(c) Radon Hash



(d) dhash

Figure I.6: Print-Scan Application results (FRR, FAR)

Lets assume a practical print-scan scenario, like an insurance company which carry out the damage report on the basis of redigitized images (print-scan) and may want to use specific duplicate hits to uncover insurance fraud, since the finding of incorrect allocations means additional effort. For such an scenario blockhash would be the best choice cause failed acceptance starts late and failed rejection is relatively low with FRR=0.007 from an threshold of 0.26 (Fig. I.6).

A.3 Twizzle - a Multi-Purpose Benchmarking Framework

MIHBS works well for PIH algorithms using a binary hash representation. However, algorithms with different hash representations or decision calculation, e.g. by involving segmentation to enhance robustness to cropping attacks [419], could not really be benchmarked with MIHBs. Thus we abstract the task to the following Question: Are two objects (in this case images) the same or not? Therefore we have designed a modular benchmarking pipeline in which the algorithms to be compared have to solve this task independently of their feature extraction and decision making. The resulting framework is written in python 3 and freely available⁶ under GPLv3. The whole workflow of the benchmarking pipeline could be seen in Figure I.7.

Overall, Twizzle consists of the “Challenge creator“ (A.3.1), the algorithm tests and the “Analyser“ (A.3.4),

⁶github.com/dfd-tud/twizzle

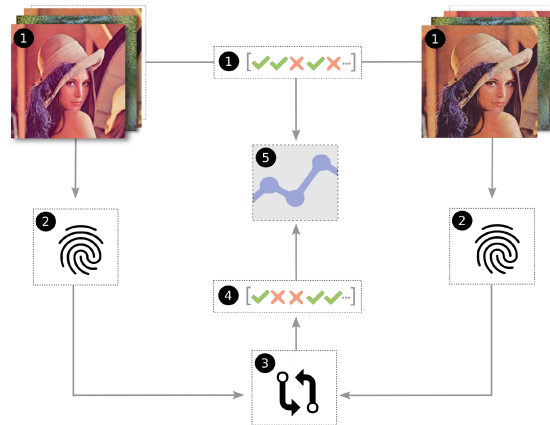


Figure I.7: General workflow of Twizzle. (1) Input: Challenge - objects to compare and ground truth (2) feature extraction (3) decision making (4) Output: list of decisions (5) Analysis: decisions vs. ground truth

where the algorithm tests can be further separated into “wrapping” the underlying algorithms to evaluate (A.3.2) and the “Test runner” ((A.3.3). Each of these parts can be independently reused in several different pipelines for different algorithms and different problem cases.

A.3.1 Challenge Creation

The first step of the benchmarking pipeline represents the creation of a specific challenge an algorithm has to solve (Fig. I.7 step 1). Twizzle originally was designed for challenges which have the form of a pairwise comparison of multimedia objects. I.e. the algorithm has to decide whether the objects, their characteristics or the content within the objects are the same or not. Thus a data set with comparison pairs, each consisting of an original object paired with a comparative object, has to be created. For each pair the expected decision (ground truth) needs to be specified. This leads to an expected decision for each media pair. I.e. if a pair consists of the same objects the decision should be “True” (“These are the same objects.”) otherwise “False“. For PIH challenges, object pairs could for example consist of similar but not same objects for a sensitivity challenge, of same but modified objects for a robustness challenge (Listing I.1) or of both as a practical use case challenge.

Listing I.1: Example Challenge Creation

```
DBPath = "test.db"
tw = Twizzle(DBPath)
sChallengeName = "image_hashing_challenge_print_scan"
aOriginals = ["c1.png", "c2.png", "c3.png"]
aComparatives = ["c1.png", "c5.png", "c6.png"]
aTargetDecisions = [True, False, False]
dicMetadata = {
    "printer": "DC783",
    "paper": "recycled_paper",
    "print_dpi": 300
}
tw.add_challenge(sChallengeName, aOriginals, aComparatives, aTargetDecisions, dicMetadata)
```

From a practical point of view, the first step is to initiate a new instance of Twizzle where a name for the database, in which challenges and results will be saved, needs to be specified. Furthermore, a list of paths to original objects, a list of corresponding comparative objects and a Boolean list of ground truth decisions for each original-comparative pair has to be created. Dimensions of the lists of original objects, corresponding objects and ground truth decisions need to be the same. Optionally, additional metadata

can be passed as python dictionary, for example to describe the specific challenge. Finally the created challenge is added to the database. Thus the challenge is prepared for any number of tests, such as the evaluation of different algorithms or the testing of different parameters of an algorithm.

A.3.2 Wrapping an Algorithm

Next a wrapper for each algorithm to be evaluated has to be created. Wrappers need to have at least two input parameters for the original and comparative objects, which are specified in the created challenge. Additional named arguments can be passed, like different parameters for the algorithm to be evaluated. The first step of such a wrapper function is to load the objects linked to in the two lists. Then, for each object pair, it has to evaluate whether the two objects are the same or not, which is done via the user-defined algorithm.

An exemplary structure of a wrapper for a user-defined passive forensic algorithm could consist of iterating over each image pair, extracting the intrinsic features and generating the fingerprint for each print out (Fig. I.7 step 2). With the user-defined decision making algorithm, like a specific threshold for hamming distance of two binary hashes, it is decided if both hashes represent the same image or not (Fig. I.7 step 3). Finally, the wrapper must return the list of algorithm decisions for all object pairs (Fig. I.7 step 4) and if desired additional arbitrary metadata.

Listing I.2: Example Wrapper

```
def wrapper(aOriginalImages, aComparativeImages, param1, ...)
    for i, sOriginalImage in enumerate(aOriginalImages):
        sComparativeImage = aComparativeImages[i]
        hashOriginal = algorithm(sOriginalImage, param1)
        hashComparative = algorithm(sComparativeImage, param1)
        deviation = distance(hashOriginal, hashComparative)
        if(deviation <= threshold): bDecision = true
        aDecisions.append(bDecision)
    return(aDecisions, dictMetadata)
```

A.3.3 Test Runs

Tests for Twizzle are black box tests. This means, that the internal workings of an algorithm are not known to Twizzle. All Twizzle expects from the user-defined algorithm is that it can handle a set of original objects and corresponding comparative objects and return some kind of decision values, which is done through the wrapper.

“Test runs“ specifies which algorithm has to solve which challenge and provide any additional parameters for “Test wrappers“. All decisions made during a test execution are returned to the Twizzle framework, where it compares the algorithm decisions with the ground truth decisions specified during “Challenge creation“ and calculates the TPR, TNR, FPR, FNR, accuracy, precision and F1 score (Fig. I.7 step 5). Additional also user-defined metadata can be returned by each test, for example the used algorithm parameters. Based on this outputs an algorithm can be easily compared to others. Tests defined in Twizzle are executed in parallel with a user-defined number of threads and can therefore also be set up on a cluster.

Listing I.3: Example Test Run

```
oRunner.run_test_async("pjh_challenge", wrapper, {"param1":param1})
```

A.3.4 Analyse Results

Twizzle also provides an Analysis component, which will collect and merge all tests and the corresponding challenges and returns a pandas dataframe [305]. This dataframe contains the test results, evaluation metrics per test and all metadata added during “Challenge creation” and running the actual test. Comparing tested algorithms can be easily done due to Twizzle abstracting the binary classification task and generating typical classification evaluation metrics. The evaluation metrics provided by Twizzle include custom metrics, challenge name as well as the metrics mentioned above.

A.3.5 Twizzle Features

Overall, Twizzle accepts any user-defined feature extraction, and decision making algorithms, independent of their type and functionality and enables the comparison of them. Further, Twizzle enables the creation of user-defined Challenges, consisting of user-defined data sets independent of their data type, with which the algorithms can be evaluated. Although Twizzle was developed for image comparisons, it can therefore be used for any pairwise comparison task, e.g. of text or audio files. Thereby Tests and Challenges are independent of each other, i.e. each created Challenge can be used for other Tests, as well as each created algorithm wrapper can be tested for any challenge. Twizzle represents the test data and evaluation metrics for analysis of each test and simplifies the comparison of the results. Further tests and analysis can be extended with user-defined meta data. Finally test runs could be executed easily in parallel.

Due to the modular design and the independence of feature extraction, decision making and the data set, Twizzle is not limited to PIH approaches, but can be used for a variety of similar applications, like face recognition or biometric authentication.

It could, for example, also be used for evaluating different printer forensic algorithms mentioned in section 3.2. For the challenge of comparing two print-outs based on their signature similarity, the benchmarking pipeline looks exactly like for a PIH approach. Therefore the data set could consists of scanned print-outs from same and different printer devices with different robustness parameters (like different fonts, print settings or paper types).

original	["printer1_font2.png", "printer2.jpg", "printer3_model1.png"]
comparative	["printer1_font1.png", "printer1.png", "printer3_model2.png"]
ground truth	[True, False, False]

After extracting the fingerprint of original and comparative print-out, the decision algorithm (classification, euclidean distance, ..) used decides whether they are from the same source device or not. After the test runs, decisions and ground truth are compared whereby tested algorithms can be compared and evaluated, independent of used signatures, extraction algorithm or decision making.

B Re-Investigation of Banding as Intrinsic EP-Printer Signature

As described in Section 3.1.1, in laser printers the image is transferred to the paper via a constantly rotating OPC drum. Fluctuations in the rotational speed of the OPC drum and errors in the gear train can cause the distances between the exposed lines on the drum to decrease or increase. These irregularities are then reflected in the printout as geometric distortions, called *banding*. Since the errors in the gear train

repeat at regular intervals, banding artifacts appear periodically as light or dark lines perpendicular to the print direction in the printout [79] (see Fig. I.8).

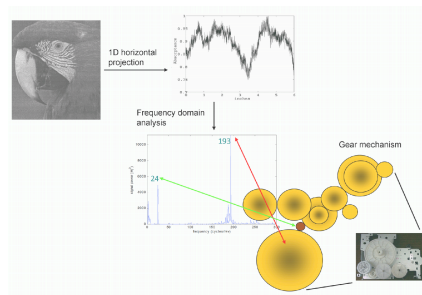


Figure I.8: Visualization of Banding and Extraction Workflow by Chiang et al. [79]

Ali et al. [14] described, in one of the first contributions on digital printer forensics, that these specific banding properties are suitable to determine the used printer model from a printed document; i.e. it could be used as an intrinsic signature. Frequency analysis was used to determine the specific banding frequencies for a printer model, allowing a clear identification of a printout, scanned at 600 dpi, from four laser printers a two manufacturers.

B.0.1 Experiment Setup

As the authors did not investigate the stability of the signature, we re-investigated it with regard to the robustness towards different test patterns (proportion of print area required), print resolutions, drivers, toner cartridges as well as the influence of the scanner. Therefore, an experimental setup with four commercially available laser printers was created (see Tab. I.4). Since Chiang et al. [79] mentioned that identification by banding is based on 'large midtone regions of a document, typically occurring in printed images', grayscale images were used as test patterns. Additional, we created line test patterns, which were used by Ali et al. [14]. Four different patterns were created in Postscript format, which differ in the percentage of the printed area by lines. For the 12.5% pattern (lp12), a 1px black line is followed by a 7px white space, for the 25% pattern (lp25) by a 3px space, and for the 50% pattern (lp50) a space of 1px. The 50% pattern was additionally created with three different line widths. Furthermore, an area completely filled with black was used (r6). The different test pattern are shown in Figure I.9.

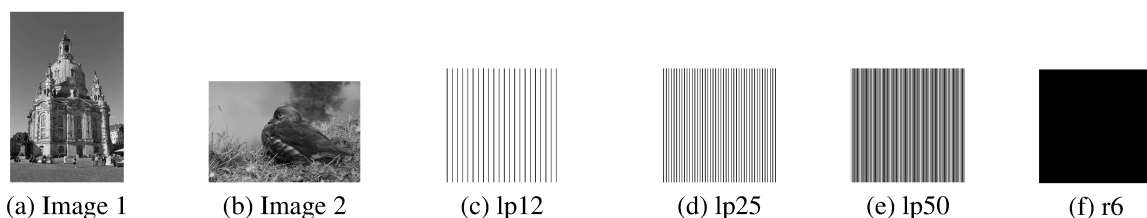


Figure I.9: Testpattern: grayscale images (a,b) and line pattern (c-f).

Each printout was scanned with two scanners (Epson Perfection V30; Canon CanoScan LiDE 25) with 600dpi resolution. Overall, 328 line patterns were created from 164 printouts (see Tab. I.4).

For the automated analysis of the banding frequencies, an extraction tool was created, which was implemented in Python using the OpenCV library. The workflow of the tool followed the description of Ali et al. [14] (Crop, Rotation, Horizontal 1-D Projection, fast Fourier transform).

B.0.2 Results

First results with grayscale images were promising. Clear peaks could be identified in the frequency spectrum. However, it was noticed that different frequencies occurred when analyzing different print resolutions. After in-depth analysis it was determined that the frequencies were not from banding, but reflected the halftoning of the prints (Fig. I.10). Thus, the proposed extraction is not suitable for the analysis of banding frequencies from grayscale image printouts.

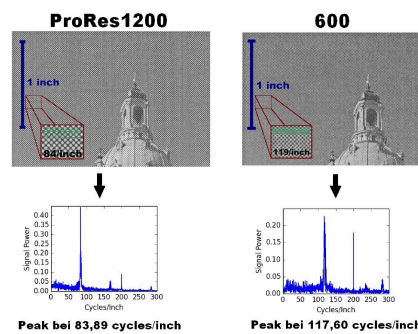


Figure I.10: Comparison of the spectrum at two different resolutions with grayscale images (HP LJ4350dtn).

In order to eliminate the halftoning influence, the following part of investigation was performed with the line patterns (which were also used in the experiments of Ali et al. [14]). For the evaluation of the results, frequency diagrams were considered. Some of them differ considerably in their appearance (Fig. I.11). For some scans there are (a) clear peaks, (b) weaker peaks and (c) no peaks at all. Ali et al. does not provide any information on this. There are several possibilities to explain the worse extraction of banding frequencies in some printouts compared to Ali et al. The extraction could be hindered by additional scan artifacts, e.g. wavy paper, which could be additional resulting in slightly inaccurate rotation. In addition, as shown by Lin et al. [277], banding artifacts are generally not exactly perpendicular to the print direction. This is due to the fact that the OPC drum continues to rotate during exposure. This results in a minimal offset between the left and right ends of a line.

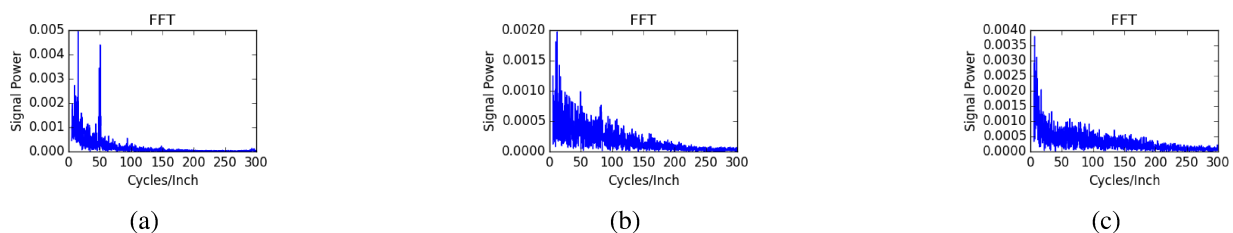


Figure I.11: Diagrams of the same printer with different peak visibility.

For further investigation of the specific banding frequencies of a printer, the frequency peaks of all printouts, detected by peak detection⁷, were determined. Table I.4 shows the detected characteristic frequencies of the printer models.

The high similarity of the histograms of HP LJ 4350dtn and HP LJ 4100 is striking, which both reach high values in the frequency ranges of 51 cycles/inch as well as around 29 and 10 cycles/inch. This can be explained by the fact that both devices belong to the same printer series. The printer structure and

⁷<https://pypi.python.org/pypi/PeakUtils>

Table I.4: List of printers used in our experiments; count of scanned line patterns; specific banding frequencies

Printer	#lp12	#lp25	#lp50	#r6	Frequencies (cycles/inch)
HP LaserJet 4350dtn	22	22	36	54	51 , 10, 31, 95
HP LaserJet 4100	12	12	44	18	51 , 29, 10, 98
HP LaserJet 1200	6	6	18	12	29
HP Color LaserJet M553	6	6	18	36	41 , 75, 23, 56

characteristics are thus very similar. It should be noted that Ali et al. tested an HP LaserJet 4050, which also belongs to this series. For this printer, 51 cycles/inch was specified as the main banding frequency, what coincides with our results. The frequency of 51 cycles/inch is obviously characteristic for devices of this series. For the HP Color LaserJet M553 the main banding frequency is at 41 cycles/inch.

Ali et al. [14] examined the HP LaserJet 1200, and found an identifying banding frequency at 69 cycles/inch. Testing the same model we found only peaks at 29 cycles/inch. The most likely explanation here is the toner cartridge. Instead of using a original toner of HP, we used a toner cartridge from a third-party supplier (Q2613X - remanufactured). There exist different constructions for OPC drums. Samsung or HP use mainly a combination of toner, developer unit and OPC drum. This is called one-way OPC drums. I.e. if the toner is exhausted also the OPC drum will be changed. In contrast, e.g., Brother or Epson use mainly a separated OPC drum. This drums have thicker organic photoconductive surface and wear slower. For the HP LJ1200 the OPC drum is built directly into the toner cartridge. The change of the signature thus shows that the influence of 3rd party components plays a major role for the intrinsic signature of banding artifacts.

The analysis of the influence of the different test patterns showed that the extraction of the banding frequencies is worse the smaller the proportion of the printed area. For the lp12 pattern in particular, there was a large scatter of values, which makes it unsuitable for determining clear characteristic frequencies. The use of different scanners and the change of the print resolution had no influence on the extraction of banding frequencies. Additional, a original toner change was performed for the HP LaserJet 4350dtn during the tests. Although the print image changed a lot (wear of the old toner produced much worse printout), the main banding frequency remained stable at 51 cycles per inch.

B.0.3 Concluding Remarks regarding the Banding Signature

The robustness in relation to print resolution, and (original) toner changes, as well as the matching banding frequency for the devices of the same series (HP LJ 4350dtn, LJ 4100 and LJ 4050), showed that banding frequency is generally suitable as a stable intrinsic printer signature. However, the latter also shows the limitation that banding is not suitable for clearly assigning a document to an individual device, but at most to a model type or even only to a model series. Moreover, 3rd party OPC drums could possibly change the signature. In addition, in our analysis of individual documents, sometimes no meaningful frequencies could be extracted, so a reliable identification of the printer for this documents was not possible - the reason for this should be further analyzed. The restriction of extraction to line patterns or larger printed areas calls the practical benefits into question. With printed images, additional influences such as halftoning or horizontal structures in the content are added, which can hinder the extraction and thus the identification rate.

C Manufacturer Statements concerning Tracking Dots

Konica Minolta DE: vielen Dank für Ihre Serviceanfrage. Über die gelben Punkte auf Ihren Ausdrucken kann ich Ihnen folgendes mitteilen. Diese Punkte befinden sich aus rein technischer Sicht auf dem Papier. Beim Laserdruckprinzip muss das Papier mit Hilfe von elektrischer Ladung von der Trommel getrennt werden. Dies erreicht man mit dem Einsatz von gelben Toner (da dieser am schwersten für das menschliche Auge sichtbar ist). So kann gewährleistet werden, dass auch ein leeres Blatt sich von der Trommel löst.

Xerox Europe: Xerox can't disclose any information on Xerox Anti-Counterfeit Technology.

Epson DE: Der Maschinenidentifizierungscode ist eine Technologie für Laserdrucker, die von den Herstellern zur Bekämpfung von kriminellen Aktivitäten wie Nachahmung oder Fälschung und zur Zusammenarbeit mit der Justiz eingesetzt wird. Grundsätzlich ist die Technologie in der Lage, das ("gefälschte") Dokument, das mit den "gelben Punkten" versehen ist, über seine Seriennummer mit einem bestimmten Gerät zu verknüpfen. Allerdings kann die Technologie den Eigentümer des Geräts nicht direkt identifizieren, es sei denn, der Laserdrucker und die Seriennummer sind registriert. Dies kann der Fall sein, wenn der Endbenutzer seinen eigenen Laserdrucker zum Zweck der Epson-Garantie registriert hat. In diesem Fall werden die registrierten und mit einem bestimmten Laserdrucker verknüpften personenbezogenen Daten auf der Grundlage der Maschinenidentifikationscodes nur im Zusammenhang mit gesetzlichen Verpflichtungen zur Verhinderung oder Verfolgung von Straftaten oder zur Zusammenarbeit mit der Justiz verarbeitet.

Ricoh: Das digitale Farbdrucksystem ist entsprechend den Forderungen zahlreicher Regierungen mit einem fälschungssicheren Kennzeichnungs- und Banknotenerkennungssystem ausgerüstet. Jede Kopie wird mit einer Kennzeichnung versehen, die nötigenfalls die Identifizierung des Drucksystems ermöglicht, mit dem sie erstellt wurde. Dieser Code ist unter normalen Bedingungen nicht sichtbar. Siehe auch: https://de.wikipedia.org/wiki/Machine_Identification_Code.

HP: At the request of law enforcement authorities, yellow dot technology is used by color laser printer and copier manufacturers to assist with investigations into certain illegal activity, such as counterfeiting currency. As a responsible market leader, HP supports the voluntary cooperation between industry and law enforcement agencies of many nations. HP has no ability to decode the yellow dot patterns and only law enforcement has direct access to the encoded information. Privacy is an important issue to HP that we take very seriously. We are committed to respecting your wishes for privacy to the best of our ability.

Brother: Auf Ausdrucken von Brother Geräten befinden sich die Seriennummer des Druckers und das Datum des Ausdrucks. Deren Dekodierung kann allerdings allein durch unsere Konzernmutter in Nagoya (Japan) auf Anfrage der dortigen Behörden erfolgen. Und diese wiederum reagieren ausschließlich auf Anfragen von Interpol. Über die Identifikation des Druckertyps und der Seriennummer lässt sich bestenfalls der Vertriebsweg innerhalb der Handelskette nachvollziehen. Es kann dadurch eventuell, sprich nur bei sauberer Dokumentation über den gesamten Vertriebsweg, der letzte Verkaufsort an einen möglichen Verbraucher / eine Privatperson sicher festgestellt werden. Die DSGVO kommt hierbei nicht zum Tragen, weil bis zum letzten Verkaufsort keine Privatperson involviert ist und weil eine Seriennummer und ein Kalenderdatum keine personenbezogenen Daten darstellen.

D User Perceptions regarding Tracking Dots

To gain an insight into whether users are aware of the integration of yellow dots, whether this integration is perceived as a problem, and also to raise awareness of it, we presented our findings at the Output⁸ and the Dresden Science Night 2022⁹. For this purpose, a booth was set up, with USB microscope, scanner, deda tool, pattern visualizations, and sample printouts. Both the advantages from a forensic point of view and the disadvantages regarding the loss of privacy were discussed. For the former, visitors were able to analyze a ransom note and compare it with three suspect patterns in order to convict the culprit. For the latter, the practical example of the impact for Reality Winner and their published documents, as well as the easy tracking of the serial number throughout the entire supply chain was used. Finally, the users were asked to fill out a small survey on paper, which a total of 58 users have complied (see Tab. I.5).

Table I.5: Overview of Participants

Age	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Count	9	31	7	5	5	0	1

Despite the fact that the integration of yellow dots in color prints has been known since 2004 and has been picked up in the media several times, their existence was surprising for most visitors (Fig. I.12

⁸<https://output-dd.de>

⁹<https://www.wissenschaftsnacht-dresden.de>

(a)). In the survey itself, 70% had never heard of them, only 25% were aware of their existence. In the assessment, the discrepancy between privacy and security is also evident in the users' opinions. 45% see the benefits of Yellow Dots for document verification and analysis, 26% see no advantages. Overall, 70% of the participants see an intrusion into their privacy by the existence of tracking dots, however, when comparing advantages and disadvantages, the opinion is again split. 28% perceive privacy impacts to be less important than benefits, 46% weigh privacy constraints more highly, and 26% perceive the pros and cons to be balanced. Overall, however, 93% of participants would like to see more transparency.

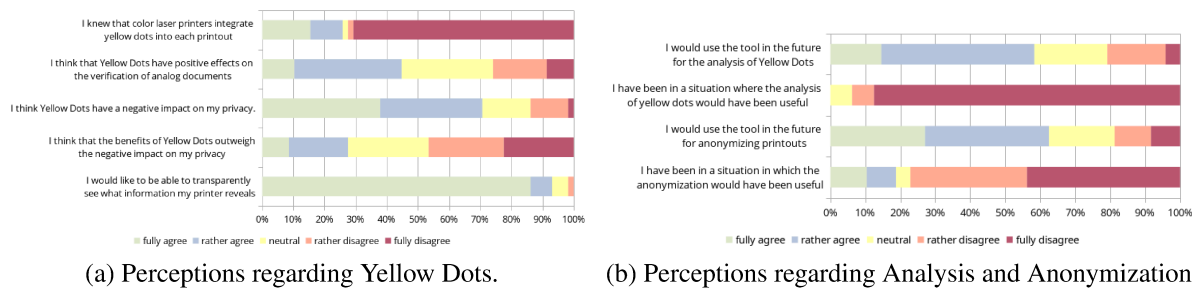


Figure I.12: Results with 58 participants.

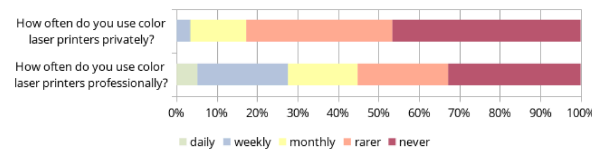


Figure I.13: Color Laser Printer Usage.

48 of the participants (83%) also tried out the deda tool in practice (see Fig. I.12 (b)). In this context, 16% stated that they had already been in a situation where anonymization of analog documents would have been important for them. None of the participants experienced an incident in which the forensic analysis of analog documents would have been necessary. For the deda tool itself, 52% stated that they would use the tool in the future for anonymization and 48% for analysis. Even if these figures are certainly due to the novelty of the information, they show, also with the reactions of the visitors, a high interest for the topic. For most visitors, however, yellow dots are of little practical significance (Fig. I.13). Only 5% of the participants use color laser printers on a daily basis, and only for business purposes. At least 28% use color lasers weekly for work and 4% for private use. Overall, however, color laser printers are used professionally by 67% of the participants and privately by 53%, even if not very regularly.

E A new Type of Tracking Dots, a unified Data Set and other Open Challenges

In our tracking dot investigations we focused mainly on unprinted areas of analog documents, as yellow dots are clearly visible and extractable at these areas. As the analysis progressed, we also considered colored areas. Here, we noticed for some printouts that the yellow dots adjust to the color area in the image. A similar behavior was mentioned by Embar et al. [129]. However, a deeper examination showed that this is not the case, but that in addition to Yellow Dots, **Inverse Yellow Dots**, which remove the

yellow part completely, are integrated. The dots are therefore not printed e.g. magenta for red areas, or cyan for green areas, as the first glance suggested, but the yellow value of this color disappears. Green is simulated by the halftoning of the colors yellow and cyan - if the yellow value disappears, cyan is shown. Instead of one yellow dot, in this setup, three dots are 'printed': inverse dot - yellow dot - inverse dot.

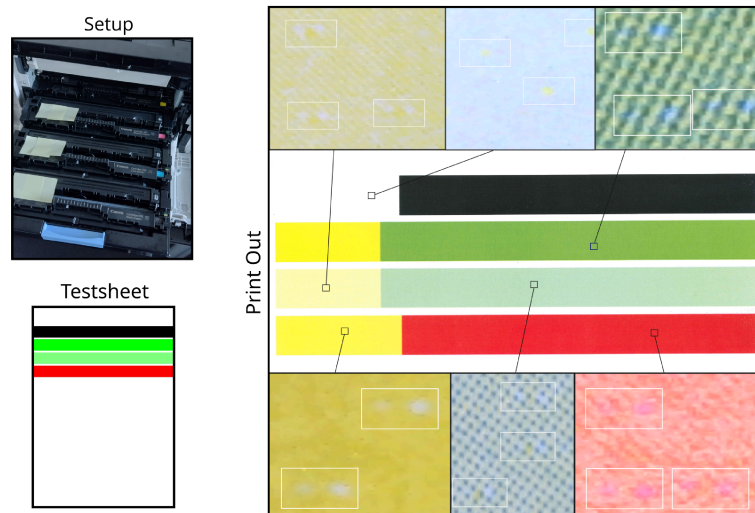


Figure I.14: Test setup and investigation results of inverse yellow dots with a Canon i-SENSYS LBP663.

Thereby, the integration of the dots is independent of the image content. This was exemplified on a Canon i-SENSYS LBP663Cdw. This is a single-pass printer, which has four individual OPC drums for the CMYK colors used. By masking the light beam, so that the laser could no longer reach the OPC drum, it was possible to ensure that only the yellow content of color areas was printed. The setup and result can be seen in Figure I.14. We found this new, unknown type of Yellow Dot pattern for Canon and HP printers in our data set. Overall, inverse yellow dots complement existing patterns and can thus be detected more easily in color areas. The encoded structure of the patterns remain the same as investigated in Section 3.3.2.

Even if not all printers are affected, our anonymization approach is invalid for the type of inverse yellow dots. In particular, when overprinting with enlarged dots (Fig. I.15), the resulting overlay with inverse dots directly identifies the anonymization dots. In addition, within color areas, the recognition of the pattern by the inverse dots is easily possible. These are visible in almost any color area, except for areas with nearly pure cyan, magenta or black. In future work, the anonymization must therefore be revised. Here, inverse dots must be added to the anonymization mask. With this, the calibration step is enormously important, because the dot sizes must now correspond exactly to those of the pattern. On the other hand, the inverse dots could be utilized for automated extraction from color areas.

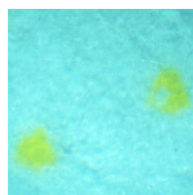


Figure I.15: Anonymized print of a Canon LBP663 printer with enlarged anonymization dots

In general, the **extraction of dots from color areas** is considered an open challenge for all pattern types,

Table I.6: Composition of the Unified Dataset.

Manufacturers	Printers	Models	Printouts
Brother	3	3	4
Canon	36	19	70
Dell	5	3	9
Epson	8	6	15
Hewlett Packard	48	21	92
IBM	1	1	2
Konica Minolta	42	20	81
Kyocera	7	6	13
Lanier	1	1	2
Lexmark	7	6	9
NRG	1	1	2
Okidata	10	7	19
Panasonic	1	1	1
Ricoh	9	7	16
Samsung	10	8	16
Savin	1	1	2
Tektronix	3	3	6
Xerox	20	18	37
total	213	132	400

even without inverse Dots. As the analysis of inverse dots showed, dots are integrated regardless of the image content. Thus, in each color area, the actual CMYK value is overlaid with the full yellow value. An approach to recognize dots within color areas could be, e.g., the iterative analysis of crops in the size of the yellow dots (depending on the pattern). The specific color difference between the surrounding color and the dot candidate (crop) could serve as a parameter for the extraction. Since yellow dots are printed independently of the document content – even behind black areas – it should be investigated if there are methods which could make them visible, e.g. due to different toner characteristics. Support for analysis and thus transparency via smartphones is also conceivable, but also poses some new open challenges.

For future work, the **dataset has to be improved** and enlarged. During our analyses, we gradually discovered that parts of our dataset were faulty or defective due to the crowd-based approach. Very poor scan quality (impossible to detect dots), different scan formats/resolutions, scans of inkjet printers, faulty meta data, missing serial numbers or model designations, can thus complicate and falsify the results of extraction and analysis. As a first step, we therefore started to unify and clean up our dataset. A maximum of two scans from each printer were included. The selection was based on dot visibility and scan quality. The meta data were collected in tabular form, which includes the following: manufacturer, model, serial number, date, pattern type, and the presence of inverse yellow dots. Manufacturer names and model designations have been unified, which facilitates filtering by properties. Printouts which were incorrectly labeled (missing/wrong meta data), duplicate entries, inkjet printouts, etc. were not adopted. In addition, we have enriched the crowd-based data with own printouts. The focus was on the manufacturers Canon, Konica-Minolta and HP to further investigate Pattern 2, 3, and 5. As test sheet T1 was used (Fig. 3.6 in Sec. 3.2.2.1). In total, 54 printers could be added from nine manufacturers a 26 models. The unified dataset is publicly available¹⁰ and an overview is given in Tab. I.6.

The combined dataset provides a basis for the analyses of tracking dots. In the future, it should be further unified, e.g., by converting different file/scan formats or dpi values by suitable conversions. Printouts in the dataset very often contain the same test images. While this is advantageous for comparative analyses, it could be an obstacle when using the data for machine learning algorithms. Thus the set could be extended/modified in order to be able to use machine learning approaches e.g. for the extraction of yellow

¹⁰<https://dfd.inf.tu-dresden.de/dataset>

dots. One could also think about generating synthetic data for training extraction algorithms, for known patterns and by adding synthetic print/scan distortions and noise [276]. The respective patterns of the prints were determined by automatic analysis and in unclear cases by hand. The information on inverse yellow dots was also determined by hand. Errors can therefore not completely ruled out. In addition, the dataset should be further expanded, esp. with patterns that have not yet been decoded.

Reverse engineering the **firmware** of printers to find the actual embedding and the concrete yellow dot functionality should be considered as an approach to analyzing the dots¹¹.

The **anonymization** could be improved, e.g., by integrating the mask into the specific printer drivers (e.g. cups). Although this would have to be done separately for each printer model, it would be much easier for the end user to use than to perform the calibration step themselves. The anonymization of scans has to be improved by considering also printed areas.

Overall, all color laser prints should be investigate further, regarding **other tracking methods**, as it could be possible that there are other hidden information embedded than tracking dots¹².

In addition to the invasion of privacy of yellow dots: Tracking dots are excellent usable for reassembling **shredded documents**¹³. This is currently not prevented with our anonymization approach.

¹¹<https://events.ccc.de/congress/2011/Fahrplan/events/4780.en.html>

¹²<https://www.eff.org/pages/list-printers-which-do-or-do-not-display-tracking-dots>

¹³<https://www.bloomberg.com/news/articles/2011-12-15/tip-for-bad-guys-burn-dont-shred>

APPENDIX II

Appendix to Information Exchange in the IoT

A Theoretical Considerations regarding the V2X EU-Strategy

The minimum distance of 800m between two changes is intended to prevent the attacker from being able to observe several pseudonym changes from the same observation point. According to the C2C-CC, the average radio range in rural areas, with a clear view between transmitter and receiver, is about 300–500m and in urban areas it is sometimes less than 100m, due to the density of buildings, which is a radio obstacle [64]. Thus, the idea behind the minimum distance of 800m should work well at least in the area of a city. The distance interval in which a vehicle carries out its third pseudonym change depends additionally on the speed of the vehicle, due to added time parameters. A footnote in the strategy description in [139] explains that the conditions from the third change of pseudonym onwards provide probable protection against the interception of the same pseudonym by two listening stations of an attacker, if they have a minimum distance of 2.5-6km in urban environments and 5-14km in motorway environments. These numbers are obviously based on an assumed average speed of 50km/h in the city and 130km/h on the motorway: $800m + 2min * 50km/h \approx 2.5km$ and $800m + 6min * 50km/h \approx 6km$. It is noticeable that the average speeds used for the calculations appear to be too high, at least during the day. In Berlin, e.g. in 2008, the average speed was 24km/h and Munich had an avg. speed of 32km/h¹. In Dresden in 2005, an average inner-city speed of 28.9km/h was recorded². Thus, the third pseudonym change is more likely to result in a distance interval of 2 – 4km. Since a more frequent change of pseudonym leads to less linkability, this result is initially to be assessed positively. It should be noted, that the average speed could be higher at times of low traffic, e.g. at night.

However, the declared goal of the switching strategy to divide at least 95% of all trips into at least three unlinkable segments, still needs the ‘exemplary estimate’ [139] that 95% of all trips are longer than 10min or longer than 3km. According to [139], this estimate is based on DLR traffic statistics that are not specified in detail. Various publicly available data suggest that this estimate is worthy of discussion.

In 2012 the German Federal Ministry of Transport, Building and Urban Affairs conducted a study

¹<https://tinyurl.com/5n8ryfwe>

²<https://www.dresden.de/media/pdf/amtsblatt/ddamt-2006-kw31-32.pdf>

“Kraftfahrzeugverkehr in Deutschland 2010” (KiD 2010) (motor vehicle traffic in Germany 2010) [479] with focused on commercial transport (passenger and freight transport) with ‘small’ commercial vehicles (passenger cars of commercial owners and trucks 3.5t payload). A result of this study was that commercial passenger cars travelling a maximum of 3km per journey is 20% (records from Monday-Friday). In terms of the duration, 31.5% of these vehicles complete their journey in 10min or less. The “Mobilität in Deutschland” (MiD- Mobility in Germany) studies by the German Federal Ministry of Transport, Building and Urban Affairs is based on the questioning of randomly selected households on their everyday traffic behavior³. According to the MiD 2002, 10% of all distances travelled by car and motorized two-wheelers (share of cars nearly 100%) reached a maximum length of 1km [215]. According to the MiD 2008, 23.2% of the distances travelled were less than 2km. At the same time, 23.8% of these trips lasted less than 10min (excluding regular work trips) [216]. According to MiD 2017, 18% of journeys were less than 10min (excluding regular work trips), while 19% of all journeys were less than 2km [116].

Although the figures vary between the individual years, none of the studies examined confirm the assumption of the C-ITS strategy. The proportion of short journeys seems to be significantly underestimated in the C-ITS strategy. Despite this results, it should be discussed to what extend the usage of averages and probabilities in the area of data privacy is feasible. For example depending on time or day this statistical distribution changes, hence the 95% goal might be missed. However, even if the assumptions of the strategy turns out to be correct and the values addressed prove to be largely stable, 5% of journeys, i.e. those that are too short, are not protected.

B Bridging V2X Pseudonym Changes with additional Identifiers

Independent of the effectiveness of the established pseudonym change algorithm, such solution can only be beneficial if all wireless identifiers of a vehicle would change at the same time. Hence, ETSI states that “[...] all the IDs associated with the ITS-Station (ITS-S) across different layers of the communication stack shall be changed synchronously with the Authorization Ticket” [140]. This shall be done via the ‘ID Change Notification service’. Further, ETSI also states that “All the IDs associated with a node across different layers of the ITS stack shall be changed synchronously using the ID Change Notification service [...]” [140]. However, this focuses only on the ITS stack and does not consider IDs not directly related to ITS-S.

Modern vehicles contain multiple identifier, ranging from traditional number plates over different identifier for wireless devices to multiple types of sensors. Overall the several identifiers can be divided into 2 different classes: vehicle independent and vehicle related identifiers.

Vehicle unrelated identifiers are those, which are not directly linked to the vehicles such as user devices (e.g. smart phones and -watches) of vehicle drivers and passengers. These devices communicate via cellular network (e.g. 4G), Wi-Fi and Bluetooth, using unique identifiers such as IMSI or MAC addresses and could be used to link V2X-Pseudonyms or for tracking itself [179].

Next to V2X-Pseudonyms there are additional *vehicle related identifiers*, which are directly interconnected with the vehicle and are part of the vehicular ecosystem. These include, for example, wireless sensors of the direct Tire Pressure Monitoring System (TPMS), which analyse the tire pressure and is either retrofitted (on the valve) or integrated in the tire [461]. In addition to sensor data, a 32-bit ID is transmitted, which is

³<http://www.mobilitaet-in-deutschland.de>

sent at intervals of 30-90 seconds while driving and can be received at a range of 10-40m. Different work [377, 383, 255] analyzed tracking of vehicles using the TPMS. For the implementation of the infotainment system within the vehicle, additional connections such as *Bluetooth BR/EDR* or in-car *Wi-Fi access points* (APs) are added. The Bluetooth Special Interest Group (SIG) estimates by 2023 93% of new vehicles will be equipped with Bluetooth and 54% of all vehicles will support Bluetooth [47]. In the vehicular environment, a range of around 30m is mostly realistic. One of the main issues with Bluetooth in the context of tracking is the usage of public MAC addresses, which is a unique not changeable identifier of a device. The practice of tracking based on the public MAC is widely used and researched in traffic analyses [30, 1, 192]. Even if MACs are changed regularly, devices might still could be tracked [34]. For traditional Wi-Fi APs, the range can be estimated in vehicles between 15m and 60m. In order to find an AP, it usually continuously sends advertising information called beacon frames (time interval of 102.4 ms) [69], which contain identifying features such as BSSID or SSID. Due to the range and frequency of the send beacons as well as the cheap and commonly available hardware, Beacon frames from APs hold a high potential for tracking. Additional *cellular connections* (4G/5G; TMSI, GUTI, IMSI) e.g. via sim cards to access the internet, or for electronic toll collections (*ETC*) systems could be integrated. Other ETC Systems are for example RFID (TID) or DSRC based. Integrated identifiers could also potentially be used to trace vehicles and break pseudonymization [393, 274].

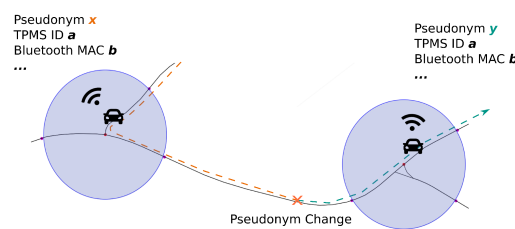


Figure II.1: The attacker can link pseudonyms of the vehicle, even if the car has changed his pseudonyms outside of the observation zones (blue), based on additional wireless identifiers.

Overall there are many IDs, which can be used to link ITS-Stack based pseudonyms and hence enable an attacker to track a vehicle even after an unobserved pseudonym change (Fig. II.1). To mitigate such attacks, IDs could be, if possible, hidden (e.g. via encryption [461]) or a system wide ID change for transmitted wireless IDs is necessary together with the pseudonym change. First, this would require the pseudonym system or a central unit to know all wireless IDs. Second, the change of all identifiable IDs would need to be initiated synchronously and retain functionality, safety and usability. This would require an interface there also third party equipment such as car radios receiver with Bluetooth can be connected to. These devices also need to be able to communicate on a common interface to change their IDs simultaneously. Some IDs (e.g. WLAN-APs MAC or SSID) are not intended to be changed instantly and could cause a longer interruption of the service. Additionally, WLAN-AP location privacy was, until the time of writing, not addressed and is still an open research question. Additionally, IDs on higher layers as well as (meta-)information contained in messages has to be researched further regarding their impact on location privacy. Consequently, there are still several open issues and research questions, which need to be addressed if V2X systems should be deployed in a privacy friendly manner.

C Open Challenges & Future Work regarding Transparency for Bystander

Enhanced Transparency Information Based on the interview results we enhanced transmitted information to: retention time, location of data storage and processing, third party access, usage of a trigger/wake word before recording, local or external processing of biometric data (e.g. text2speech on device and disclosure of transcribed text only), usage of anonymization techniques, linkage of recorded data with other data sources, and the compliance with minimum security requirements of the DP (Fig II.2, left⁴).

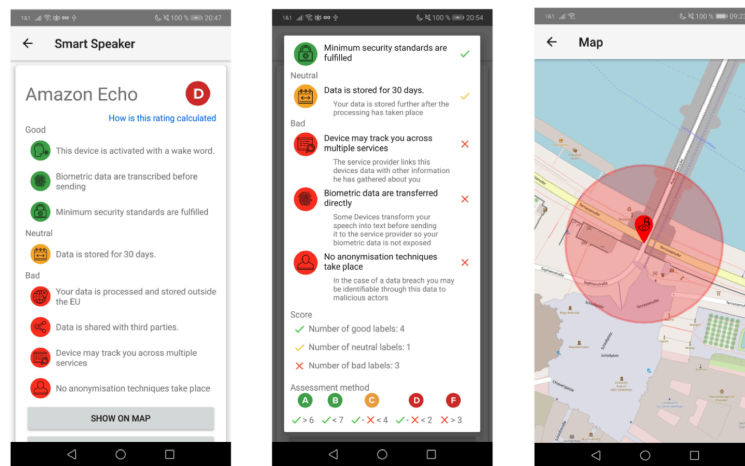


Figure II.2: Enhanced transparency information, simple metric and map visualization

The latter is inspired by the Mozilla Privacy Not Included Project⁵ and includes information such as encryption, security updates, and vulnerability management. For better insight/assessment, the storage/processing location could be compared with the country-specific regulations regarding privacy⁶. Additional, to make the general privacy score more transparent, we have created and integrated a first simple metric. For this, each mentioned characteristic gets a score consisting of three gradations (Fig II.2, center).

Since the exact location of the devices was particularly important to the participants in both evaluations and this was also found in other studies [504, 298], we integrated the distance to the device, which could be calculated based on the received signal strength.

Through the interesting suggestions of a participant to map the history of surrounding IoT devices, this approach is adopted and prototypically integrated. In addition, this approach also offers the possibility of a hybrid solution in which, user devices commit crowd-based surrounding stationary IoT devices into an external database, thus enabling a general map approach and preventive warnings. Here, layer 1 information (privacy grade, recording channel), the distance to the device as well as its recording radius are visualized on the map (Fig II.2, right). Currently, it remains open how the actuality of mapped devices can be guaranteed, e.g. when devices are removed after the absence of BLE signals.

Next to link the complete privacy policy of the device/DP, also a link to view and delete personal data should be available. Additionally, possible inferences of personal information from the recorded sensor

⁴Icons by <https://icons8.com>

⁵<https://foundation.mozilla.org/privacynotincluded>

⁶like <https://www.dlapiperdataprotection.com/index.html?t=world-map>

data could be visualized, to improve general privacy awareness, resulting in more confident privacy-decisions [273]. Further, it might be interesting to consider implicit/predicted data in addition to explicit information about data processing. Especially in the area of biometric data, from which further sensitive data, such as the user's health status, could be derived [253, 87, 234].

Support for DC Transparency Due to the lack of balance between DC and DP, we primarily focused on the visualization of transparency information of the DP. However transmitted content should at the best also entail information and privacy policies about the DC. Even though information about the DC was not felt to be really important to participants, the purpose of the recording was important to some, which can only be answered by the DC. One solution to enable this in our current transparency app could be to provide an interface for modifying parts of the beacon data, which are set by the DC during installation, and reflect the identity and contact details of the DC as well as the purpose of the recording. Dependent on the device, it is possible that also third party applications could have access to recorded data. Therefore, additional application-based information has to be shared about the data processing of this third party apps involving their additional privacy policies. This could be solved similar to DC information.

Privacy Issues for IoT Device Owner Whether direct or indirect solutions are used, the transmission of information about surrounding IoT devices enables transparency not only for bystanders but also for malicious listeners. These could be enabled for example to identify lucrative burglary locations, to find vulnerable devices, or to track mobile IoT devices and thus to create motion profiles of device owners. Additionally transparency could also have negative impacts on different applications, such as security systems. Thereby, the issue depends highly on the level of detail of transmitted transparency information. Thus, the transmitted information content must be weighed against bystander and owner privacy. For example, only the recording channel could be transmitted, the level of detail could depend on the application area and purpose, or information should only be transmitted if identifiability of the bystander is enabled. The concrete protection of device owners, however, remains an open question.

Towards Consent The possibility for interaction and intervention of recordings is important and is also required e.g. by the GDPR. However, it must be noted that intervention only makes sense in certain cases. A security system, for example, would lose all purpose if persons could decide whether or not to be recorded. Otherwise, this security system could also be used e.g. to measure visitor flows and customer interests, at the same time. With regard to the implementation of such an intervention, a general shutdown of sensors is not reasonable if the dynamic benefit of the IoT devices is to be maintained. Users with conflicting privacy settings can be within the recording range of a IoT device at the same time. While one generally rejects the recording, the other wants to use the associated functionality of it. The use of user-specific gadgets to prevent/block recordings [356] is also questionable as a solution suitable for everyday use. Accordingly, recordings must be filtered by privacy settings of the recorded persons. A "privacy by design" [267] implementation of such a dynamic interaction, where the data subject is not recorded / biometric properties are hidden, unless she explicitly consent to it, would be the most sensible approach. Whether opt-in or opt-out, in both cases a communication interface is required to transmit biometric profiles for filtering.

Filtering directly on IoT devices before sending the data to more powerful servers for further processing

would be most reasonable. Regarding our design, BLE offers the GATT (Generic Attribute Profile) protocol for 1-on-1 communication between BLE-supporting devices. This could allow the exchange of biometric profiles, which should be opted out/in of the recording/processing. The exchange and storage has to be implemented privacy friendly. The question here is whether the performance of IoT devices is sufficient to store such models and to filter and pre-process corresponding recordings. Additional, a direct interface would enable new potential security issues, especially with a lack of device updates. Another possibility would be to filter on DP side. Thus, BLE broadcasted transparency information could include a reference to the DP, which would provide an interface for transmitting biometric profiles and filtering them before actual processing. This could also be used to transfer privacy settings regarding multiple devices of the same DP. On the other hand, biometric data can be viewed both at the DC and on the way to the DP, similarly critical as e.g. to set cookies before consent [302].

Individual Privacy Settings Since the desire for transparency is highly individual and context-dependent, this should be taken into account for further developments. For example, the notice of surrounding devices should be filterable individually according to e.g. different DPs/DCs, purpose, user location, recording channel, or privacy properties of the recordings. A corresponding interaction solution should also take the minimal distraction principle into account, so that the user does not have to give his consent/rejection for every device. There is already some work on simplifying consent and/or automatic consent according to preset/detected privacy preferences, e.g. [185, 472, 272, 504]. Consideration should also be given to the results of Colnago et al. [89], which show that users "weigh the desire for control against the fear of cognitive overload", and thus the degree of automation should also be individually adjustable.

Technical Limitations of Bluetooth With the proposed way of recognizing smart devices via BLE, they must be equipped with appropriate beacons. Even while such micro controllers are relatively cheap to manufacture and have low upkeep costs, it could make IoT devices more expensive. Additional, if a large number of devices in the vicinity send out packets, collisions could occur, which in turn could increase power consumption of the user device, as well as the non-detection of devices. Since beacons usually do not perform authentication, devices can be spoofed without special effort. In addition, jamming attacks could affect the visibility of devices. The latter attacks could be detected and a warning could be displayed. In addition to Android implementation, a development for other platforms would be necessary. It must be noted that, e.g., iOS offers only limited capacities for BLE communicating apps while running in background.

Every Day Life Humans change behavioural patterns if they think they are being watched, even just through the presence of an eye symbol on a sign (also known as the Watching Eyes Effect) [107]. A similar effect could occur if a transparency solution is deployed. Furthermore, it is conceivable that our transparency solution could provoke a defensive reaction of people regarding the display of many recording devices, causing people to rather "close their eyes" than be confronted with reality. This as well as possible impacts of transparency on user behavior in different situations should be investigated in future work.

APPENDIX III

Survey Structures

A Analysis of the Privacy Paradox for Mobile and Web Applications

In October 2019 we conducted a small online questionnaire via LimeSurvey regarding web and mobile usage, tracking knowledge and used countermeasures. The survey was pretested three times and afterwards distributed via the student newsletter of the TU Dresden. Thus, the results mainly reflect the opinions of the user group of young German students. Overall 137 participants completed the survey, mainly aged 21-30 years (76.64%); followed by 10-20 years (18.25%); 31-40 (4.8%) and 41-50 (0.73%). In the following the questionnaire and results are shown.

Table III.1: (Q1,Q3) Which browser do you use mainly on your

	Q1 PC/Laptop	Q3 Mobile Device
None	0.00%	2.19%
Chrome	45.26%	41.61%
Firefox	39.42%	8.76%
Internet Explorer	1.46%	-
Edge	0.73%	0.00%
Opera	2.19%	0.00%
Safari	8.03%	28.47%
Tor Browser / Onion Browser	0.00%	0.73%
Brave Browser	0.00%	-
Samsung Internet Browser	-	5.11%
UC Browser	-	0.00%
DuckDuckGo Privacy Browser	-	6.57%
Firefox Klar / Focus	-	1.46%
Other	2.92%	5.11%

Table III.2: Q2,Q4: For which application areas do you use your browser for?

Q5: For which application areas do you use mobile apps instead?

	Q2 PC/Laptop	Q4 Mobile Device	Q5
Social media	56.20%	28.47%	75.91%
Search for information (Wikipedia, news, ...)	98.54%	95.62%	11.68%
Entertainment (video, music, ...)	89.78%	33.58%	78.83%
Games	13.87%	2.92%	45.26%
Shopping (travel booking, ...)	79.56%	40.88%	29.20%
Online banking	78.10%	8.76%	51.82%
Communication (e-mail, messenger, ...)	77.37%	27.74%	83.21%
Productivity (office application, calendar, ...)	37.23%	10.22%	56.20%
Other	9.49%	11.68%	10.22%
Question not shown	0.00%	2.19%	0.00%

Table III.3: Q6-Q9: Please specify to what extent you agree.

	I fully agree	I rather agree	Neutral	I rather disagree	I fully disagree
Q6 I believe that most visited websites display personalized content (e.g. search results, ads, product recommendations based on my interests).	35.77%	50.36%	5.11%	8.03%	0.73%
Q7 I believe that personalized content enhances the quality and usability of websites.	5.84%	21.17%	25.55%	37.23%	10.22%
Q8 I believe that personalized content has negative effects on my privacy.	43.80%	35.77%	10.22%	7.30%	2.92%
Q9 I believe that the advantages of personalization outweigh the negative effects on my privacy.	5.84%	11.68%	23.36%	35.04%	24.09%

Table III.4: Q10-Q13: Please specify how comfortable you feel towards the following statements.

	Extremely Comfortable	Somewhat Comfortable	Neutral	Somewhat Uncomfortable	Extremely Uncomfortable
Q10 While searching for a good restaurant on your smartphone you receive suggestions to restaurants nearby.	32.85%	37.23%	10.95%	14.60%	4.38%
Q11 While visiting a different website you receive advertisements for restaurants nearby.	0.73%	4.38%	16.06%	48.18%	30.66%
Q12 While visiting a website on your PC/laptop you receive advertisements for restaurants nearby as well.	2.19%	6.57%	17.52%	31.39%	42.34%
Q13 An acquaintance you met for dinner in that restaurant is suggested as a friend in your social network.	0.73%	7.30%	14.60%	21.90%	55.47%

Table III.5: Q14-Q19: Please specify to what extent you agree.

	I fully agree	I rather agree	Neutral	I rather disagree	I fully disagree
<i>Q14</i> I believe it is necessary to analyze my surfing behavior, my location and similar data for the personalization of web contents.	33.58%	35.77%	5.84%	10.22%	14.60%
<i>Q15</i> I believe that solely the visited website is able to analyze my surfing behavior.	3.65%	7.30%	8.76%	31.39%	48.91%
<i>Q16</i> I believe that the analysis of my surfing behavior works solely when I am logged in to the corresponding website.	1.46%	5.11%	5.11%	22.63%	65.69%
<i>Q17</i> I believe that collected data is stored anonymously.	6.57%	16.79%	15.33%	39.42%	21.90%
<i>Q18</i> I believe that collected data is used for commercial purposes.	60.58%	34.31%	2.92%	2.19%	0.00%
<i>Q19</i> I believe that I can't do anything against the analysis of my surfing behavior.	11.68%	29.93%	21.17%	25.55%	11.68%

Table III.6: Q20: Please check all entities that in your opinion analyze your browsing behavior and your online activity.

None	0.73%
The Website you are visiting	89.05%
Advertiser/Analytic services	94.16%
Governments	31.39%
Internet service providers (e.g., Vodafone, Telecom, ...)	56.20%
Browser creators (e.g., Google, Mozilla, ...)	78.10%
Others	8.03%

Others via free text field: social networks 6, search engines 4, everyone who has something to do with marketing and has enough money, third parties via cookies/fingerprinting, business enterprise, consulting firms, intelligence agencies 2, hackers

Table III.7: Q21: Which of the following techniques/terms and their functionalities are known to you?

Cookies	IP-Address	Browser Fingerprinting	JavaScript	None of these
99.27%	97.08%	38.69%	80.29%	0.00%

Table III.8: Q22-Q26: Please specify to what extent you agree.

	I fully agree	I rather agree	Neutral	I rather disagree	I fully disagree
<i>Q22</i> As soon as the a cookie banner appears I close it immediately.	32.85%	36.50%	10.22%	17.52%	2.92%
<i>Q23</i> I consider the cookie banner useful.	5.84%	20.44%	15.33%	29.20%	29.20%
<i>Q24</i> I feel adequately informed on the usage of my personal data by cookie banners.	2.19%	11.68%	24.09%	38.69%	23.36%
<i>Q25</i> Since the implementation of cookie banners, I have been engaged with the analysis of my browsing behavior.	2.19%	16.79%	16.06%	30.66%	34.31%
<i>Q26</i> If cookie banners offer the opportunity to manage my settings, I change the settings regarding the processing of my data.	28.47%	29.93%	9.49%	14.60%	17.52%

Table III.9: Q27-Q29: Please specify to what extent you agree.

	Yes	No
Q27 I have inquired at an internet service before which personal information they have stored relating to me.	19.71%	80.29%
Q28 I have used the opt-out method of a web analytics service before to prevent the tracking of my browsing behavior by this service.	30.66%	69.34%
Q29 I have requested an internet service before to delete my collected data.	20.44%	79.56%

Table III.10: Q30,31: Which of the following settings have you changed in your Browser?

	Q30 PC/Laptop	Q31 Mobile Device
None	21.17%	37.96%
Block all cookies - activated	9.49%	10.95%
Block third party cookies - activated	44.53%	27.01%
Keep local data only until you quit your browser - activated	27.01%	17.52%
Content blocking / Tracking protection - activated	43.80%	25.55%
Withdrawn permissions (camera, microphone, location, ...)	67.88%	43.07%
Send "DoNotTrack" request - activated	39.42%	24.82%
Flash / Java / Silverlight - deactivated	27.74%	
JavaScript - deactivated	18.25%	6.57%
WebRTC - deactivated	5.11%	2.19%
Send referrer - deactivated	6.57%	2.92%
Data collection by browser - deactivated	32.85%	18.98%
Question not shown	0.00%	2.19%

Table III.11: Q31-: Which of the following techniques/terms and their functionalities are known to you?

Cookies	IP-Address	Browser Fingerprinting	JavaScript	None of these
99.27%	97.08%	38.69%	80.29%	0.00%

Table III.12: Q32: How often do you delete your private data on your desktop browser?

Q33: How often do you delete your private data on your mobile browser?

Q34: How often do you use the incognito/private mode on your desktop Browser?

Q35: How often do you use the incognito/private mode on your mobile browser?

	Several times daily	.. a week	.. a month	More rarely	Never	Q not shown
Q32 Cookies	6.57%	8.76%	10.95%	40.88%	32.85%	0.00%
Q32 Web Cache	5.11%	10.95%	13.14%	39.42%	31.39%	0.00%
Q32 Browser / Download history	6.57%	8.03%	12.41%	44.53%	28.47%	0.00%
Q32 Autofill form data	5.11%	3.65%	7.30%	40.15%	43.80%	0.00%
Q33 Cookies	1.46%	7.30%	12.41%	30.66%	45.99%	2.19%
Q33 Web Cache	1.46%	10.22%	13.14%	31.39%	41.61%	2.19%
Q33 Browser / Download history	2.92%	8.76%	13.87%	34.31%	37.96%	2.19%
Q33 Autofill form data	1.46%	5.11%	6.57%	32.85%	51.82%	2.19%
Q34	13.14%	27.01%	16.79%	18.25%	24.82%	0.00%
Q35	13.14%	18.98%	17.52%	20.44%	27.74%	2.19%

Table III.13: Q36: What is the main reason for using the incognito/private mode?

To impede the tracking of my surfing behavior	19.71%
To surf the Internet anonymously	8.03%
To leave no trace of my surfing behavior on the used device	45.26%
I was recommended to use it	4.38%
Others	5.84%
Question not shown	16.79%

Table III.14: Q37: Which add-ons/extensions have you installed on your desktop browser to impede the tracking of your surfing behavior or block ads?

None	21.17%
Adblock or Adblock Plus	57.66%
uBlock or uBlock Origin	24.82%
Ghostery	9.49%
Disconnect	2.92%
NoScript or ScriptSafe	11.68%
Blur	1.46%
Privacy Badger	5.11%
DuckDuckGo Privacy Essentials	5.11%
Other	8.03%

Table III.15: Q38-41: How often do you use these services?

	Several times daily	Several times a week	Several times a month	More rarely	Never
Q38: Tor (Desktop)	3.65%	2.92%	2.19%	13.87%	77.37%
Q39: Tor (Mobile)	2.19%	1.46%	0.73%	8.03%	87.59%
Q40: VPN (Desktop)	10.22%	11.68%	13.87%	32.12%	32.12%
Q41: VPN (Mobile)	3.65%	9.49%	5.84%	20.44%	60.58%

Table III.16: Q42: Do you take any action to impede the tracking of your surfing behavior on your mobile device?
Q44: Do you use any alternative web services which promise to protect your privacy?

	Yes	No
Q42	24.09%	75.91%
Q44	40.15%	59.85%

Q43 Please specify which actions you take on your mobile device (free text field):

Restricted permission management 12; browser addons (uBlock, NoScript) 8; browser settings 4; privacy friendly apps (see Q44) adblock security apps (Kaspersky) 12; VPN usage 5; Tor usage; fake loginsdata; avoiding extensive searches with the mobile device

Q45 Please specify which alternative web services you use to protect your privacy (free text field):

Duckduckgo (Search, Browser) 22; Telegram 10; Ecosia 8; Posteo 4; Jabber/XMPP 2; Threema 3; Signal; Wire; GMX 2; Tor 3; Vivaldi; Yahoo Mail; Kaspersky 2; Uni-Email; Aloha Browser; Startpage 4; Openstreetmap 3; Selfhosting (PW Manager, VPN, Mail, Matrix, Filesharing) 3; ixquick 2; Qwant.com; Google G Suite; ProtonMail; Nextcloud; mailbox.org; Cliqz browser;

B Nunti-Score - News Items and Distribution

Table III.17 shows the list of items used for the evaluation of the Nunti-Score. The part of the item text that was used as a heading is printed in bold. The last three columns indicate how a given item was contextualized in the respective group of participants (G1 - group 1, G2 - group 2, G3 - group 3). Here U stands for unambiguous, and A for an ambiguous rating by the contextualization. N means that the item was displayed without a rating.

Table III.17: News Items and Distribution

Difficulty	Truth Value	Content	G1	G2	G3
Simple	False	Ei im Beutel: Die völlig verrückte Welt der Kängurus! Glückliche Beobachter können zu dieser Jahreszeit erspähen, wie sich die nistenden Kängurupaare am Uluru beim Ausbrüten ihrer Eier in ihren Beuteln abwechseln.	U	N	A
Simple	False	Vorbild für Trump? Nachdem er sein Amt 2016 niedergelegt, hat sich Barack Obama vollständig dem Züchten von Edelzwiebeln verschrieben. Aufgrund seiner Leistungen dabei, wurde ihm letzte Woche der begehrte Züchterpreis „La Larme d’Or“ verliehen.	U	N	A
Simple	False	Neuausrichtung in Stuttgart Aufgrund absehbarer Probleme im Automobilsektor, will sich Mercedes-Benz in Zukunft auf die Herstellung luxuriöser Kaffeemaschinen konzentrieren und den Bau von Fahrzeugen größtenteils einstellen. Diese Entscheidung wurde	A	U	N
Simple	False	Dammbruch bei der Mobilität Wegen des hohen Lastverkehrs auf deutschen Autobahnen sollen diese, laut Plänen der Bundesregierung, teilweise durch Kanäle ersetzt werden. „Der Transport mittels Schiffen erlaubt ganz andere Größenordnungen im Warenverkehr“, so ein Sprecher	A	U	N
Simple	False	Geschäftszahlen von Bahlsen zeigen erneut: Spekulation ist das beliebteste Gebäck zu Ostern! Wie der Konzern bei seiner Aktionärsversammlung zum Ende des Geschäftsjahres verkündete	N	A	U
Simple	False	Uni Münster stiftet Preis zu Ehren Albert Einsteins, dem Entdecker der Schwerkraft Mit seiner Entdeckung legte Albert Einstein den Grundstein für die Naturwissenschaft der Physik. Im Rahmen der geplanten feierlichen Preisverleihung sollen daher	N	A	U
Simple	True	Massensterben vor 65 Millionen Jahren löschte die Dinosaurier aus auch wenn sie vor langer Zeit die vorherrschende Spezies auf der Erde waren, sind die Dinosaurier heute ausgestorben. Wie viele andere damals lebende Tierarten	U	N	A
Simple	True	Bester Freund des Menschen – Hunde können bei Depressionen helfen Die Gegenwart eines vierbeinigen Freundes hebt bei vielen depressiven Personen die Laune und vermindert Stress und Anspannung. Dies wurde durch eine in Augsburg mit Therapiehunden durchgeführte Studie nun erneut bestätigt	U	N	A
Simple	True	Die Niederlande – nicht nur flach, sondern tief! Rund ein Viertel der Fläche der Niederlande liegen aktuell unterhalb des Meeresspiegels. Dies birgt im Hinblick auf den Klimawandel enorme Herausforderungen	A	U	N
Simple	True	Ein Jahr hat im Schnitt 365 Tage, aber warum? Die Dauer eines Jahres richtet sich nach dem Zeitraum, den die Erde benötigt, um die Sonne einmal zu umrunden. Diese Umrundung dauert ca. 365 Tage, woraus sich	A	U	N
Simple	True	Rekordverdächtig – Blauwale sind die schwersten auf der Erde lebenden Säugetiere. Die größten Exemplare ihrer Art werden bis zu 200 Tonnen schwer. Sie leben	N	A	U
Simple	True	Zeitgeschichte: Im Jahr 1969 landete Neil Armstrong als erster Mensch auf dem Mond. Er beendet damit das „Rennen zum Mond“, dass sich die USA damals mit der Sowjetunion lieferten	N	A	U

Difficulty	Truth Value	Content	G1	G2	G3
Difficult	False	Kunstform mit martialischem Ursprung: Steptanz. Auch wenn der moderne Steptanz nichts von seinem Erbe ahnen lässt, kann sein Ursprung zu einer Übung der Schweizer Garde aus der späten Renaissance zurückverfolgt werden. Die korrekte Ausführung der komplexen Schrittfolgen sollte	U	N	A
Difficult	False	Menschen sehen scheinbar doch nur 24 Bilder pro Sekunde. In den letzten Jahren gab es - vor allem unter Videospielern - immer wieder erbitterte Kontroversen: Können Menschen nur 24 Bilder pro Sekunde sehen? Wie Forscher an der Universität Bochum nun herausgefunden haben, scheint dies zuzutreffen.	U	N	A
Difficult	False	8. Mai 1945: Fall des Faschismus Mit der bedingungslosen Kapitulation der Wehrmacht am 8. Mai 1945 fand die letzte autoritäre Diktatur in Westeuropa ihr Ende.	A	U	N
Difficult	False	Nicht nur eine gute Spürnase! Eine kürzlich durchgeführte Studie legt nahe, dass Hunde Ultraschall nutzen, um sich bei schlechter Sicht zu orientieren. Eine diesbezügliche Studie wurde an der Akademie für tiermedizinische Wissenschaften in Bruchsal durchgeführt.	A	U	N
Difficult	False	Ernährungsberater unterstreichen die Bedeutung von Kohlenhydraten in der Ernährung Kohlenhydrate haben auf das Gramm gerechnet mehr Kalorien als Eiweiß oder Fett.	N	A	U
Difficult	False	Held der Sowjetunion und Pionier der Raumfahrt. Juri Gagarin schaffte es als erster Bürger der Sowjetunion auf den Mond. Wegen dieser außergewöhnlichen Leistung erlangte er 1970 weltweite Bekanntheit.	N	A	U
Difficult	True	Super, Mario! Spieleindustrie wächst weiter. Die Videospieleindustrie ist nach wie vor der Umsatzstärkste Zweig der Unterhaltungsbranche, noch vor Filmen und Musik. Experten erwarten, dass dieser Trend sich ungebrochen fortsetzt und die kommenden Jahre sogar noch weiteres Wachstum	U	N	A
Difficult	True	WWF warnt: 60% der Tierarten, die 1970 noch auf diesem Planeten lebten sind mittlerweile ausgestorben! Die Organisation erwartet, dass der Klimawandel die verbleibenden Populationen vor enorme Herausforderungen stellen wird und so diesen Prozess weiter beschleunigt.	U	N	A
Difficult	True	Umweltschutzverbände alarmiert: Forscher finden bei Tauchgang im Marianengraben Plastikmüll. Es ist wahrscheinlich keine Übertreibung zu behaupten der Marianengraben mit seinen 12 km Tiefe sei einer der abgelegensten Orte der Welt. Dennoch scheint er nicht abgelegen genug zu sein, um den menschlichen Einfluss zu entgehen.	A	U	N
Difficult	True	Wasser kocht nicht immer bei 100 °C. Leute, die in der Schule gut aufgepasst haben, erinnern sich vielleicht daran, dass bei Experimenten oft „Normalbedingungen“ erwähnt wurden.	A	U	N
Difficult	True	Auch Rabenvögel besitzen scheinbar einen wichtigen Baustein des Bewusstseins. Krähen sind in der Lage Spiegel zu benutzen, um für sie ansonsten nicht sichtbare Informationen zu erhalten. In einer Studie der RuhrUniversität Bochum wurde	N	A	U
Difficult	True	Halsbrecherische Tradition am Cooper's Hill. Jahr für Jahr findet hier in England der gefährlichste Wettlauf der Welt statt, bei dem die Teilnehmer versuchen einen Käselaib einzuholen, der einen steilen Abhang mit bis zu 110 km/h hinabrollt. Die Teilnehmer nehmen schwere Verletzungen in Kauf,	N	A	U

Bibliography

- [1] Montasir Abbas, Lakshmi Rajasekhar, Asmita Gharat, and John Paul Dunning. Microscopic modeling of control delay at signalized intersections based on bluetooth data. *Journal of Intelligent Transportation Systems*, 2013.
- [2] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, and William Bainbridge. Encyclopedia of human-computer interaction. *Thousand Oaks: Sage Publications*, 2004.
- [3] Ruba Abu-Salma, M Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the adoption of secure communication tools. In *Security & Privacy*. IEEE, 2017.
- [4] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *Computing Surveys (Csur)*, 2018.
- [5] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 2015.
- [6] Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *Security & Privacy*, 2005.
- [7] Aashish Agarwal and Torsten Zesch. German end-to-end speech recognition based on deepspeech. In *KONVENS*, 2019.
- [8] Amol Agrawal. Clickbait detection using deep learning. In *International Conference on next generation Computing Technologies (NGCT)*. IEEE, 2016.
- [9] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J Lee. Tangible privacy: Towards user-centric sensor designs for bystander privacy. *CHI*, 2020.
- [10] Yoshinori Akao, Kazuhiko Kobayashi, Shigeru Sugawara, and Yoko Seki. Discrimination of inkjet-printed counterfeits by spur marks and feature extraction by spatial frequency analysis. In *Optical Security and Counterfeit Deterrence Techniques IV*. International Society for Optics and Photonics, 2002.
- [11] Yoshinori Akao, Atsushi Yamamoto, and Yoshiyasu Higashikawa. Improvement of inkjet printer spur gear teeth number estimation by fixing the order in maximum entropy spectral analysis. In *International Workshop on Computational Forensics*. Springer, 2010.
- [12] Muhammad Al-Qurishi, Mabrook Al-Rakhani, Atif Alamri, Majed Alrubaian, Sk Md Mizanur Rahman, and M Shamim Hossain. Sybil defense techniques in online social networks: a survey. *IEEE Access*, 2017.
- [13] Hassan Alhuzali and Sophia Ananiadou. Spanemo: Casting multi-label emotion classification as span-prediction. *arXiv*, 2021.

- [14] Gazi N Ali, Aravind K Mikkilineni, Jan P Allebach, Edward J Delp, Pei-Ju Chiang, and George T Chiu. Intrinsic and extrinsic signatures for information hiding and secure printing with electrophotographic devices. In *NIP & Digital Fabrication Conference*. Society for Imaging Science and Technology, 2003.
- [15] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 2017.
- [16] Hazim Almuhiemedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location has been Shared 5,398 Times!: A Field Study on Mobile App Privacy Nudging. In *Conference on Human Factors in Computing Systems*. ACM, 2015.
- [17] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 2015.
- [18] AltBeacon. Protocol specification v1.0. <https://github.com/AltBeacon/spec>, 2020.
- [19] Julio Angulo, Simone Fischer-Hübner, Tobias Pulls, and Erik Wästlund. Usable transparency with the data track: a tool for visualizing data disclosures. In *Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2015.
- [20] Johanna Ansohn McDougall, Christian Burkert, Daniel Demmler, Monina Schwarz, Vincent Hubbe, and Hannes Federrath. Probing for passwords—privacy implications of ssids in probe requests. In *International Conference on Applied Cryptography and Network Security*. Springer, 2022.
- [21] Nicolas Anspach, Jay Jennings, and Kevin Arceneaux. A little bit of knowledge: Facebook’s News Feed and self-perceptions of knowledge. *Research & Politics*, 2019.
- [22] Nalin Asanka Gamagedara Arachchilage and Steve Love. A game design framework for avoiding phishing attacks. *Computers in Human Behavior*, 2013.
- [23] Alberto Ardèvol-Abreu, Trevor Diehl, and Homero Gil de Zúñiga. Antecedents of Internal Political Efficacy Incidental News Exposure Online and the Mediating Role of Political Discussion. *Politics*, 2019.
- [24] Patricia Arias-Cabarcos, Saina Khalili, and Thorsten Strufe. ‘surprised, shocked, worried’: User reactions to facebook data collection from third parties. *Proceedings on Privacy Enhancing Technologies*, 2023.
- [25] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 2015.
- [26] Patricia Aufderheide. Media literacy: From a report of the national leadership conference on media literacy. In *Media literacy in the information age*. Routledge, 1997.
- [27] Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center*, 2019.
- [28] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. Exposure to social engagement metrics increases vulnerability to misinformation. *arXiv*, 2020.
- [29] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychol. Sci.*, 2015.
- [30] Jaume Barceló, Lidin Montero, Laura Marqués, and Carlos Carmona. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record*, 2010.
- [31] Melisa Basol, Jon Roozenbeek, and Sander Van der Linden. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition*, 2020.

- [32] Jon Bateman. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace, 2020.
- [33] Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. Testing and comparing computational approaches for identifying the language of framing in political news. In *Conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, 2015.
- [34] Johannes K Becker, David Li, and David Starobinski. Tracking anonymized bluetooth devices. *Proc. Priv. Enhancing Technol.*, 2019.
- [35] Zinaida Benenson, Freya Gassmann, and Robert Landwirth. Unpacking spear phishing susceptibility. In *International conference on financial cryptography and data security*. Springer, 2017.
- [36] James Benhardus and Jugal Kalita. Streaming trend detection in twitter. *International Journal of Web Based Communities*, 2013.
- [37] Alastair R Beresford, Dorothea Kübler, and Sören Preibusch. Unwillingness to pay for privacy: A field experiment. *Economics letters*, 2012.
- [38] Alastair R. Beresford, Andrew Rice, Nicholas Skehin, and Ripduman Sohan. MockDroid: trading privacy for application functionality on smartphones. In *Workshop on Mobile Computing Systems and Applications - HotMobile*. ACM, 2011.
- [39] Adam J Berinsky. Rumors and health care reform: Experiments in political misinformation. *British journal of political science*, 2017.
- [40] Julia Bernd, Ruba Abu-Salma, and Alisa Frik. Bystanders’ privacy: The perspectives of nannies on smart home surveillance. In *FOCI*, 2020.
- [41] Jan Hendrik Betzing, Matthias Tietz, Jan vom Brocke, and Jörg Becker. The impact of transparency on mobile privacy decision making. *Electronic Markets*, 2020.
- [42] Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael Horning, and Tanushree Mitra. FeedReflect: A Tool for Nudging Users to Assess News Credibility on Twitter. In *Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2018.
- [43] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. Inferring the demographics of search users: Social data meets search queries. In *WWW*, 2013.
- [44] Christoph Bier, Kay Kühne, and Jürgen Beyerer. Privacyinsight: the next generation privacy dashboard. In *Annual Privacy Forum*. Springer, 2016.
- [45] Paul Bleakley. Panic, pizza and mainstreaming the alt-right: A social media analysis of pizzagate and the rise of the qanon conspiracy. *Current Sociology*, 2021.
- [46] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O’Dell, Kevin Butler, and Patrick Traynor. Who are you (i really wanna know)? detecting audio DeepFakes through vocal tract reconstruction. In *USENIX Security*, 2022.
- [47] BluetoothSIGProprietary. Bluetooth market update (v5.1), 2019.
- [48] Mark Blythe, Helen Petrie, and John A Clark. F for fake: four studies on how we fall for phish. In *SIGCHI*, 2011.
- [49] Pablo Boczkowski, Eugenia Mitchelstein, and Mora Matassi. News Comes Across When I’m in a Moment of Leisure: Understanding the Practices of Incidental News Consumption on Social Media. *New Media Soc.*, 2018.

- [50] Alexander Bor and Michael Bang Petersen. The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American political science review*, 2022.
- [51] Abdelwahab Boualouache, Sidi-Mohammed Senouci, and Samira Moussaoui. A survey on pseudonym changing strategies for vehicular ad-hoc networks. *IEEE Communications Surveys & Tutorials*, 2017.
- [52] Antoine Boutet, Hyoungshick Kim, and Eiko Yoneki. What’s in Twitter, I Know What Parties are Popular and Who You are Supporting Now! *Social Network Analysis and Mining*, 2013.
- [53] Andrei Boutyline and Robb Willer. The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks. *Political Psychology*, 2017.
- [54] William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 2017.
- [55] Michael D Briscoe. The paperless office twenty years later: Still a myth? *Sustainability: Science, Practice and Policy*, 2022.
- [56] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.
- [57] Ceren Budak, Anitha Kannan, Rakesh Agrawal, and Jan Pedersen. Inferring User Interests From Microblogs. In *AAAI*, 2014.
- [58] Tomasz Bujlow, Valentín Carela-Español, Josep Sole-Pareta, and Pere Barlet-Ros. A survey on web tracking: Mechanisms, implications, and defenses. *Proceedings of the IEEE*, 2017.
- [59] Orhan Bulan, Junwen Mao, and Gaurav Sharma. Geometric distortion signatures for printer identification. In *ICASSP*. IEEE, 2009.
- [60] Bundeskriminalamt. Cybercrime | bundeslagebild 2018. *BKA*, 2019.
- [61] Bundeskriminalamt. Cybercrime | bundeslagebild 2019. *BKA*, 2020.
- [62] Bundeskriminalamt. Cybercrime | bundeslagebild 2020. *BKA*, 2021.
- [63] Levente Buttyán, Tamás Holczer, and István Vajda. On the effectiveness of changing pseudonyms to provide location privacy in vanets. In *ESAS*, 2007.
- [64] C2C-CC. FAQ regarding Data Protection in C-ITS, 2018.
- [65] Jan Camenisch and Anna Lysyanskaya. An efficient system for non-transferable anonymous credentials with optional anonymity revocation. In *EUROCRYPT*. Springer, 2001.
- [66] Jan Camenisch and Els Van Herreweghen. Design and implementation of the idemix anonymous credential system. In *ACM CCS*, 2002.
- [67] Gamze Canova, Melanie Volkamer, Clemens Bergmann, Roland Borza, Benjamin Reinheimer, Simon Stockhardt, and Ralf Tenberg. Learn to spot phishing urls with the android nophish app. In *IFIP*. Springer, 2015.
- [68] Deanna D Caputo, Shari Lawrence Pfleeger, Jesse D Freeman, and M Eric Johnson. Going spear phishing: Exploring embedded training and awareness. *Security & Privacy*, 2013.
- [69] T. Carpenter and L. Badman. Cwap certified wireless analysis professional official study guide: Cwap-402, 2016.
- [70] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv*, 2020.

- [71] Claude Castelluccia, Mathieu Cunche, Daniel Le Métayer, and Victor Morel. Enhancing transparency and consent in the iot. In *European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2018.
- [72] Abdelberi Chaabane, Gergely Acs, Mohamed Ali Kaafar, et al. You are what you like! information leakage through users' interests. In *NDSS*. Citeseer, 2012.
- [73] Jarrett Chambers, W Yan, A Garhwal, and M Kankanhalli. Currency security and forensics: a survey. *Multimedia Tools and Applications*, 2015.
- [74] Julia Chantal and Serge Hercberg. Development of a new front-of-pack nutrition label in france: the five-colour nutri-score. *Public Health Panorama*, 2017.
- [75] Pranob K Charles, V Harish, M Swathi, and CH Deepthi. A review on the various techniques used for optical character recognition. *International Journal of Engineering Research and Applications*, 2012.
- [76] Yen-Pin Chen, Yi-Ying Chen, Kai-Chou Yang, Feipei Lai, Chien-Hua Huang, Yun-Nung Chen, Yi-Chin Tu, et al. The prevalence and impact of fake news on covid-19 vaccination in taiwan: Retrospective study of digital media. *JMIR*, 2022.
- [77] Yushi Cheng, Xiaoyu Ji, Tianyang Lu, and Wenyan Xu. Dewicam: Detecting hidden wireless cameras via smartphones. In *ASIA-CCS*, 2018.
- [78] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 2019.
- [79] Pei-Ju Chiang, Nitin Khanna, Aravind K. Mikkilineni, Maria V.O. Segovia, Sungjoo Suh, Jan P. Allebach, George T.-C. Chiu, and Edward J. Delp. Printer and scanner forensics. *IEEE Signal Processing Magazine*, 2009.
- [80] Jung-Ho Choi, Hae-Yeoun Lee, and Heung-Kyu Lee. Color laser printer forensic based on noisy feature and support vector machine classifier. *Multimedia Tools and Applications*, 2013.
- [81] Jung-Ho Choi, Hae-Yeoun Lee, and Heung-Kyu Lee. Color laser printer forensic based on noisy feature and support vector machine classifier. *Multimedia Tools and Applications*, 2013.
- [82] Matteo Cinelli, Stefano Cresci, Alessandro Galeazzi, Walter Quattrociocchi, and Maurizio Tesconi. The limited reach of fake news on twitter during 2019 european elections. *PloS one*, 2020.
- [83] William Clarkson, Tim Weyrich, Adam Finkelstein, Nadia Heninger, J Alex Halderman, and Edward W Felten. Fingerprinting blank paper using commodity scanners. In *Security & Privacy*. IEEE, 2009.
- [84] Katherine Clayton, Spencer Blair, Jonathan Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 2020.
- [85] Lara Codeca, Raphaël Frank, and Thomas Engel. Luxembourg sumo traffic (lust) scenario: 24 hours of mobility for vehicular networking research. In *VNC*, 2015.
- [86] Raviv Cohen and Derek Ruths. Classifying Political Orientation on Twitter: It's Not Easy! In *AAAI*, 2013.
- [87] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009.
- [88] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of communication*, 2014.

- [89] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. Informing the design of a personalized privacy assistant for the internet of things. In *CHI*, 2020.
- [90] European Commission. Eurobarometer. attitudes on data protection and electronic identity in the european union, 2011.
- [91] European Commission. Eurobarometer. data protection, 2015.
- [92] European Commission. Eu code of practice on disinformation, 2018.
- [93] European Commission. Eurobarometer. charter of fundamental rights and general data protection regulation, 2019.
- [94] European Commission. Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/ec. *Digital Services Act*, 2022.
- [95] European Commission. The strengthened code of practice on disinformation, 2022.
- [96] European Commission. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828. *Artificial Intelligence Act*, 2024.
- [97] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political Polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.
- [98] Lorrie Faith Cranor. P3P: Making privacy policies more useful. *IEEE Security & Privacy*, 2003.
- [99] Stefano Cresci. A decade of social bot detection. *Communications of the ACM*, 2020.
- [100] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *WWW*, 2017.
- [101] Antitza Dantcheva, Petros Elia, and Arun Ross. What else does your biometric data reveal? a survey on soft biometrics. *Trans. Inf. Forensics Secur.*, 2015.
- [102] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. Personalized privacy assistants for the internet of things: Providing users with notice and choice. *Pervasive Computing*, 2018.
- [103] Anupam Das, Martin Degeling, Xiaoyou Wang, Junjue Wang, Norman Sadeh, and Mahadev Satyanarayanan. Assisting users in a world full of cameras: A privacy-aware infrastructure for computer vision applications. In *CVPRW*, 2017.
- [104] Reza Davarzani, Saeed Mozaffari, and Khashayar Yaghmaie. Perceptual image hashing using center-symmetric local binary patterns. *Multimedia Tools and Applications*, 2016.
- [105] Reza Davarzani, Saeed Mozaffari, and Khashayar Yaghmaie. Perceptual image hashing using center-symmetric local binary patterns. *Multimedia Tools and Applications*, 2016.
- [106] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *WWW*. ACM, 2016.
- [107] Keith Dear, Kevin Dutton, and Elaine Fox. Do ‘watching eyes’ influence antisocial behavior? A systematic review & meta-analysis. *Evolution and Human Behavior*, 2019.
- [108] Wei Deng, Qinghu Chen, Feng Yuan, and Yuchen Yan. Printer identification based on distance transform. In *First International Conference on Intelligent Networks and Intelligent Systems*. IEEE, 2008.

- [109] Clemens Deußer, Steffen Passmann, and Thorsten Strufe. Browsing unicity: On the limits of anonymizing web tracking data. In *Symposium on Security and Privacy (SP)*. IEEE, 2020.
- [110] M. Uma Devi, C. Raghvendra Rao, and M. Jayaram. Statistical Measures for Differentiation of Photocopy from Print technology Forensic Perspective. *International Journal of Computer Applications*, 2014.
- [111] Capitaine Marie Deviterne-Lapeyre. Interpol review of questioned documents 2016–2019. *Forensic Science International: Synergy*, 2020.
- [112] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *SIGCHI*, 2006.
- [113] Giorgia Di Tommaso, Stefano Faralli, Giovanni Stilo, and Paola Velardi. Wiki-MID: A Very Large Multi-Domain Interests Dataset of Twitter Users With Mappings to Wikipedia. In *International Semantic Web Conference*, 2018.
- [114] Nicholas Dias, Gordon Pennycook, and David G Rand. Emphasizing organizations does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, 2020.
- [115] Cong Ding, Yang Chen, and Xiaoming Fu. Crowd crawling: Towards collaborative data collection for large-scale online social networks. In *Proceedings of the first ACM conference on Online social networks*, 2013.
- [116] DLR, infas. “Mobilität in Deutschland – Tabellarische Grundausswertung – Verkehrsaufkommen – Struktur – Trends”, 2018.
- [117] Claire Dolin, Ben Weinshel, Shawn Shan, Chang Min Hahn, Euirim Choi, Michelle L Mazurek, and Blase Ur. Unpacking perceptions of data-driven inferences underlying online targeting and personalization. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018.
- [118] Julie S Downs, Mandy Holbrook, and Lorrie Faith Cranor. Behavioral response to phishing risk. In *APWG eCrime researchers summit*, 2007.
- [119] Arne Dreißigacker, Bennet Skarczynski, and Gina Wollinger. Cyberangriffe gegen unternehmen in deutschland, ergebnisse einer repräsentativen unternehmensbefragung 2018/2019, forschungsbericht nr. 152. *Kriminologisches Forschungsinstitut Niedersachsen e.V.*, 2020.
- [120] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.
- [121] Dieter Eberwein. *Nietzsches Schreibkugel: ein Blick auf Nietzsches Schreibmaschinenzeit durch die Restauration der Schreibkugel*. Typoskript-Verlag Dieter Eberwein, 2005.
- [122] Juan Echeverria, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Shi Zhou. LOBO: Evaluation of generalization deficiencies in twitter bot classifiers. In *Proceedings of the 34th Annual Computer Security Applications Conference*, 2018.
- [123] Lilian Edwards and Wiebke Abel. The use of privacy icons and standard contract terms for generating consumer trust and confidence in digital services. *CREATE Working Paper Series*, 2014.
- [124] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You’ve been warned: an empirical study of the effectiveness of web browser phishing warnings. In *SIGCHI*, 2008.
- [125] Elizabeth L. Eisenstein. *The Printing Revolution in Early Modern Europe*. Cambridge University Press, 2012.
- [126] Sara Elkasrawi and Faisal Shafait. Printer Identification Using Supervised Learning for Document Forgery Detection. In *International Workshop on Document Analysis Systems*. IEEE, 2014.

- [127] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the experts: What should be on an iot privacy and security label? In *Security & Privacy*. IEEE, 2020.
- [128] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring how privacy and security factor into iot device purchase behavior. In *CHI*, 2019.
- [129] Maya Embar, Louis F. McHugh IV, and William R. Wesselman. Printer watermark obfuscation. In Becky Rutherford, Lei Li, Susan Van de Ven, Amber Settle, and Terry Steinbach, editors, *RIIT*. ACM, 2014.
- [130] Arielle Emmett. Networking news: Traditional news outlets turn to social networking web sites in an effort to build their online audiences. *American Journalism Review*, 2008.
- [131] Sven Engesser and Edda Humprecht. Frequency or Skillfulness: How Professional News Media Use Twitter in Five Western Countries. *Journalism studies*, 2015.
- [132] Nicole Ernst, Sven Engesser, Florin Büchel, Sina Blassnig, and Frank Esser. Extreme Parties and Populism: An Analysis of Facebook and Twitter Across Six Countries. *Inf. Commun. Soc.*, 2017.
- [133] Samson Esayas, Tobias Mahler, and Kevin McGillivray. Is a picture worth a thousand terms? visualising contract terms and data protection requirements for cloud computing users. In *International Conference on Web Engineering*. Springer, 2016.
- [134] Stephan Escher and Stefan Köpsell. Durchführung eines integrierten anti-phishing-trainings. In *Sicherheit in vernetzten Systemen. 23. DFN-Konferenz*, 2018.
- [135] Stephan Escher, Jan Reubold, Richard Kwasnicki, Joachim Scharloth, Lutz Hagen, and Thorsten Strufe. Towards automated contextualization of news articles. In *MIS2: Misinformation and Misbehavior Mining on the Web, WSDM Workshops*, 2018.
- [136] ETSI. TS 102 894-2 (1.3.1) - Intelligent Transport Systems (ITS); Users and applications requirements; Part 2: Applications and facilities layer common data dictionary, 2018.
- [137] ETSI, EN 302 637-2: Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service, Version 1.3.1, 2014.
- [138] ETSI, EN 302 637-3: Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 3: Specifications of Decentralized Environmental Notification Basic Service, Version 1.2.1, 2014.
- [139] ETSI, TR 103 415 (v1.1.1) - Intelligent Transport Systems (ITS); Security; Pre-standardization study on pseudonym change management, 2018.
- [140] ETSI, TS 102 940 (v1.3.1) - Intelligent Transport Systems (ITS); Security; ITS communications security architecture and security management, 2018.
- [141] Katrin Eitzrodt and Sven Engesser. Ubiquitous tools, connected things and intelligent agents: Disentangling the terminology and revealing underlying theoretical dimensions. *First Monday*, 2019.
- [142] EuropeanCommission. Certificate Policy for Deployment and Operation of European Cooperative Intelligent Transport Systems (C-ITS) – Release 1, 2017.
- [143] EuropeanCommission. Security Policy & Governance Framework for Deployment and Operation of European Cooperative Intelligent Transport Systems (C-ITS) – Release 1, 2017.
- [144] EuropeanParliament. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995.

- [145] EuropeanParliament. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016.
- [146] EuropeanUnion. Regulation (eu) 2018/858 of the european parliament and of the council, 2018.
- [147] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *ICARCV*. IEEE, 2018.
- [148] Paul F. Langer. Lessons from china-the formation of a social credit system: Profiling, reputation scoring, social engineering. In *DG.O*, 2020.
- [149] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale readability analysis of privacy policies. In *Conference on web intelligence*, 2017.
- [150] Stefano Faralli, Giovanni Stilo, and Paola Velardi. Large Scale Homophily Analysis in Twitter Using a Twixonomy. In *IJCAI*, 2015.
- [151] Lisa Fazio. Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, 2020.
- [152] Mengjuan Fei, Zhaojie Ju, Xiantong Zhen, and Jing Li. Real-time visual tracking based on improved perceptual hashing. *Multimedia Tools and Applications*, 2017.
- [153] Mengjuan Fei, Jing Li, and Honghai Liu. Visual tracking based on improved foreground detection and perceptual hashing. *Neurocomputing*, 2015.
- [154] Bjarke Felbo, Alan Mislove, Anders Søgaaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing*, 2017.
- [155] Albert Feller, Matthias Kuhnert, Timm O Sprenger, and Isabell M Welp. Divided They Tweet: The Network Structure of Political Microbloggers and Discussion Topics. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.
- [156] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android permissions demystified. *CCS*, 2011.
- [157] Emilio Ferrara. The history of digital spam. *Communications of the ACM*, 2019.
- [158] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 2016.
- [159] Anselmo Ferreira, Luiz C. Navarro, Giuliano Pinheiro, Jefersson A. dos Santos, and Anderson Rocha. Laser printer attribution: Exploring new features and beyond. *Forensic Science International*, 2015.
- [160] Richard Fletcher and Sora Park. The impact of trust in the news media on online news consumption and participation. *Digital journalism*, 2017.
- [161] International Organization for Standardization. Iso 9241-11:2018(en) ergonomics of human-system interaction — part 11: Usability: Definitions and concepts, 2018.
- [162] International Organization for Standardization. Iso 9241-110:2020(en) ergonomics of human-system interaction — part 110: Interaction principles, 2020.
- [163] David Förster, Frank Kargl, and Hans Löhr. A framework for evaluating pseudonym strategies in vehicular ad-hoc networks. In *WiSec*, 2015.

- [164] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *CSUR*, 2018.
- [165] Electronic Frontier Foundation. Docucolor tracking dot decoding guide. <https://w2.eff.org/Privacy/printers/docucolor/>, 2005. Accessed: 2023-01.
- [166] Agence France-Presse. AFP Updates Guidelines on Using Social media. *AFP Newsletter [Online]*, 2013.
- [167] Gerald Friedland and Robin Sommer. Cybercasing the joint: On the privacy implications of geo-tagging. In *5th USENIX Workshop on Hot Topics in Security (HotSec 10)*, 2010.
- [168] Batya Friedman, Peyina Lin, and Jessica K Miller. Informed consent by design. *Security and Usability*, 2005.
- [169] Bundesamt für Sicherheit in der Informationstechnik. Die lage der it-sicherheit in deutschland 2021. *BSI*, 2021.
- [170] Lukas Gal, Michaela Belovičová, Michal Ceppan, Michal Oravec, and Miroslava Palková. Analysis of laser and inkjet prints using spectroscopic methods for forensic identification of questioned documents. In *Symposium on Graphic Arts*, 2013.
- [171] Kevin Gallagher, Sameer Patil, and Nasir Memon. New me: Understanding expert and non-expert perceptions and usage of the tor anonymity network. In *SOUPS*, 2017.
- [172] Brett S Gardner. Responsive web design: Enriching the user experience. *Sigma Journal: Inside the Digital Ecosystem*, 2011.
- [173] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *WebConf*, 2018.
- [174] Matthew D. Gaubatz and Steven J. Simske. Printer-scanner identification via analysis of structured security deterrents. In *First IEEE International Workshop on Information Forensics and Security*. IEEE, 2009.
- [175] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *SSCORE*, 2013.
- [176] Johann Gebhardt, Markus Goldstein, Faisal Shafait, and Andreas Dengel. Document Authentication Using Printing Technique Features and Unsupervised Anomaly Detection. In *International conference on document analysis and recognition*. IEEE, 2013.
- [177] Yegin Genc, Yasuaki Sakamoto, and Jeffrey V Nickerson. Discovering Context: Classifying Tweets Through a Semantic Transform Based on Wikipedia. In *FAC*, 2011.
- [178] Homero Gil de Zúñiga, Brian Weeks, and Alberto Ardèvol-Abreu. Effects of the news-finds-me perception in communication: Social media use implications for news seeking and learning about politics. *Journal of computer-mediated communication*, 2017.
- [179] Hadi Givvehchian, Nishant Bhaskar, Eliana Rodriguez Herrera, Héctor Rodrigo López Soto, Christian Dameff, Dinesh Bharadia, and Aaron Schulman. Evaluating physical-layer ble location tracking attacks on mobile devices. In *SP*. IEEE, 2022.
- [180] Jochen Gläser and Grit Laudel. *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen*. Verlag für Sozialwissenschaften, 2009.
- [181] Maria Glenski, Corey Pennycuff, and Tim Weninger. Consumers and curators: Browsing and voting patterns on reddit. *CSS*, 2017.
- [182] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021.
- [183] Diane Golay. The bad news game: a defense against fake news. *XRDS: Crossroads, The ACM Magazine for Students*, 2020.

- [184] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC*, 2012.
- [185] Richard Gomer, MC Schraefel, and Enrico Gerding. Consenting agents: semi-autonomous interactions for ubiquitous consent. In *UbiComp Adjunct*, 2014.
- [186] Hongmei Gou, Ashwin Swaminathan, and Min Wu. Intrinsic sensor noise features for forensic analysis on scanners and scanned images. *Transactions on Information Forensics and Security*, 2009.
- [187] Anti-Phishing Working Group. Phishing activity trends report. *2nd Quarter*, 2022.
- [188] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 2019.
- [189] Andrew Guess, Brendan Nyhan, and Jason Reifler. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council*, 2018.
- [190] Hartmut Günther and Otto Ludwig. *Schrift und Schriftlichkeit / Writing and its Use*. Walter de Gruyter, 2008.
- [191] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. TweetCred: Real-time Credibility Assessment of Content on Twitter. In *International Conference on Social Informatics*. Springer, 2014.
- [192] Gaby Gurczik, Marek Junghans, and Sten Ruppe. Conceptual approach for determining penetration rates for dynamic indirect traffic detection. *ITS World Congress*, 2012.
- [193] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. Away from prying eyes: Analyzing usage and understanding of private browsing. In *SOUPS*, 2018.
- [194] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. Toggles, dollar signs, and triangles: How to (in) effectively convey privacy choices with icons and link texts. In *CHI*, 2021.
- [195] Lutz Hagen. Nachrichtenjournalismus in der vertrauenskrise. „lügenpresse“ wissenschaftlich betrachtet: Journalismus zwischen ressourcenkrise und entfesseltem publikum. *ComSoc Communicatio Socialis*, 2015.
- [196] Tzipora Halevi, Nasir Memon, and Oded Nov. Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks. *Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks*, 2015.
- [197] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. Automated identification of media bias by word choice and labeling in news articles. In *Joint Conference on Digital Libraries (JCDL)*. ACM/IEEE, 2019.
- [198] Michael Hameleers, Thomas E Powell, Toni GLA Van Der Meer, and Lieke Bos. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 2020.
- [199] Roozbeh Hamzehyan, Farbod Razzazi, and Alireza Behrad. Printer source identification by feature modeling in the total variable printer space. *Journal of Forensic Sciences*, 2021.
- [200] Simon Hanisch, Patricia Arias-Cabarcos, Javier Parra-Arnau, and Thorsten Strufe. Privacy-protecting techniques for behavioral data: A survey. *arXiv*, 2021.
- [201] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring price discrimination and steering on e-commerce web sites. In *IMC/ACM*, 2014.
- [202] Katrin Hartwig and Christian Reuter. TrustyTweet: An Indicator-based Browser-Plugin to Assist Users in Dealing with Fake News on Twitter. *International Conference on Wirtschaftsinformatik*, 2019.

- [203] ASM Touhidul Hasan, Qingshan Jiang, and Chengming Li. An effective grouping method for privacy-preserving bike sharing data publishing. *Future Internet*, 9, 2017.
- [204] Stephen Hawkins, Daniel Yudkin, Miriam Juan-Torres, and Tim Dixon. Hidden tribes: A study of america's polarized landscape, 2019.
- [205] Zhen He and Charles A Bouman. Am/fm halftoning: A method for digital halftoning through simultaneous modulation of dot size and dot placement. In *Color Imaging: Device-Independent Color, Color Hardcopy, and Applications VII*, volume 4663. International Society for Optics and Photonics, 2001.
- [206] Hans Hedbom. A survey on transparency tools for enhancing privacy. In *IFIP Summer School on the Future of Identity in the Information Society*. Springer, 2008.
- [207] Itai Himelboim, Marc Smith, and Ben Shneiderman. Tweeting Apart: Applying Network Analysis to Detect Selective Exposure Clusters in Twitter. *Communication methods and measures*, 2013.
- [208] Caspar Hirschi. Dreyfus, zola, and the graphologists. from the failure of experts to the victory of intellectuals? *HISTORISCHE ZEITSCHRIFT*, 2016.
- [209] Sascha Hölig, Uwe Hasebrink, and Julia Behre. *Reuters Institute digital news report 2020: Ergebnisse für Deutschland*. Leibniz-Institut für Medienforschung | Hans-Bredow-Institut, 2020.
- [210] Leif-Erik Holtz, Katharina Nocun, and Marit Hansen. Towards displaying privacy information with icons. In *IFIP primelife international summer school on privacy and identity management for life*. Springer, 2010.
- [211] Leif-Erik Holtz, Katharina Nocun, and Marit Hansen. Towards displaying privacy information with icons. In *Privacy and Identity Management for Life*. Springer, 2011.
- [212] Jeff Howe et al. The rise of crowdsourcing. *Wired magazine*, 2006.
- [213] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic prediction based on user's browsing behavior. In *WWW*, 2007.
- [214] Soheil Human, Max Schrems, Alan Toner, Ben Wagner, et al. Advanced Data Protection Control (ADPC). *WU Vienna University of Economics and Business*, 2021.
- [215] infas Institut für angewandte Sozialwissenschaft. Mobilität in Deutschland: Ergebnisbericht, 2004.
- [216] infas Institut für angewandte Sozialwissenschaft. Mobilität in Deutschland 2008: Tabellenband, 2010.
- [217] Interpol. Interpol global crime trend report, 2022.
- [218] Jim Isaak and Mina J Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 2018.
- [219] Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. Social phishing. *Communications of the ACM*, 2007.
- [220] Ankit Kumar Jain and BB Gupta. A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterprise Information Systems*, 2022.
- [221] Hardik Jain, Gaurav Gupta, Sharad Joshi, and Nitin Khanna. Passive classification of source printer using text-line-level geometric distortion signatures from scanned images of printed documents. *arXiv*, 2017.
- [222] Maurice Jakesch, Moran Koren, Anna Evtushenko, and Mor Naaman. The role of source and expressive responding in political news evaluation. In *Computation and Journalism Symposium*, 2019.
- [223] Ilka Jakobs, Tanjev Schultz, Christina Viehmann, Oliver Quiring, Nicklaus Jakob, Marc Ziegele, and Christian Schemer. Mainzer langezeitstudie medienvertrauen in deutschland 2020. *Uni Mainz*, 2021.

- [224] Markus Jakobsson and Steven Myers. *Phishing and countermeasures: understanding the increasing problem of electronic identity theft*. John Wiley & Sons, 2006.
- [225] Milena Janic, Jan Pieter Wijnbenga, and Thijs Veugen. Transparency enhancing tools (tets): an overview. In *Third Workshop on Socio-Technical Aspects in Security and Trust*. IEEE, 2013.
- [226] Weina Jiang, Anthony TS Ho, Helen Treharne, and Yun Q Shi. A novel multi-size block benford's law scheme for printer identification. In *Pacific-Rim Conference on Multimedia*. Springer, 2010.
- [227] Neil F Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. The online competition between pro-and anti-vaccination views. *Nature*, 2020.
- [228] Marco Schreyer Thomas M. Breuel Joost Van Beusekom. Automatic counterfeit protection system code classification. *Proc.SPIE*, 2010.
- [229] Kenneth Joseph, Peter M Landwehr, and Kathleen M Carley. Two 1% s don't make a whole: Comparing simultaneous samples from twitter's streaming api. In *SBP*, 2014.
- [230] Sharad Joshi and Nitin Khanna. Single classifier-based passive system for source printer classification using local texture features. *Transactions on Information Forensics and Security*, 2017.
- [231] Sharad Joshi and Nitin Khanna. Source printer classification using printer specific local texture descriptor. *Transactions on Information Forensics and Security*, 2019.
- [232] White Joshua, Matthews Jeanna, and Stacy John. A method for the automated detection phishing websites through both site characteristics and image analysis. In *Cyber Sensing*. SPIE, 2012.
- [233] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *IJCV*, 2022.
- [234] Alexander Kachur, Evgeny Osin, Denis Davydov, Konstantin Shutilov, and Alexey Novokshonov. Assessing the Big Five personality traits using real-life static facial images. *Sci. Rep.*, 2020.
- [235] Joseph Kahne and Benjamin Bowyer. Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. *American Educational Research Journal*, 2017.
- [236] Ben Kaiser, Jerry Wei, Elena Lucherini, Kevin Lee, Nathan Matias, and Jonathan Mayer. Adapting Security Warnings to Counter Online Disinformation. In *USENIX*, 2021.
- [237] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. My data just goes everywhere: User mental models of the internet and implications for privacy and security. In *SOUPS*, 2015.
- [238] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society*, 2021.
- [239] Asif Karim, Sami Azam, Bharanidharan Shanmugam, Krishnan Kannoorpatti, and Mamoun Alazab. A comprehensive survey for intelligent spam email detection. *Access*, 2019.
- [240] Eric Kee and Hany Farid. Printer profiling for forensics and ballistics. In *Workshop on Multimedia and Security*. ACM, 2008.
- [241] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A " nutrition label" for privacy. In *Symposium on Usable Privacy and Security*, 2009.
- [242] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

- [243] Antino Kim, Patricia Moravec, and Alan Dennis. Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, 2019.
- [244] Do-Guk Kim, Jong-Uk Hou, and Heung-Kyu Lee. Learning deep features for source color laser printer identification based on cascaded learning. *arXiv*, 2017.
- [245] Do-Guk Kim and Heung-Kyu Lee. Color laser printer identification using photographed halftone images. In *European Signal Processing Conference (EUSIPCO)*. IEEE, 2014.
- [246] Do-Guk Kim and Heung-Kyu Lee. Colour laser printer identification using halftone texture fingerprint. *Electronics Letters*, 2015.
- [247] Helmut Kipphan, editor. *Handbook of print media: technologies and production methods*. Springer, 2001.
- [248] Jan Kirchner and Christian Reuter. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction*, 2020.
- [249] Herbert Klimant, Rudi Piotraschke, and Dagmar Schönfeld. *Informations- und Kodierungstheorie*. Springer, 2012.
- [250] Elena Kochkina, Maria Liakata, and Isabelle Augenstein. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. In *International Workshop on Semantic Evaluation (SemEval)*, 2017.
- [251] Spyros Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security*, 2017.
- [252] Konrad Kollnig and Nigel Shadbolt. TrackerControl: Transparency and Choice around App Tracking. *Journal of Open Source Software*, 2022.
- [253] Michal Kosinski. Facial recognition technology can expose political orientation from naturalistic facial images. *Sci. Rep.*, 2021.
- [254] Jacob Leon Kröger, Philip Raschke, and Towhidur Rahman Bhuiyan. Privacy implications of accelerometer data: a review of possible inferences. In *ICCSP*, 2019.
- [255] Milos Krstic, Nemanja Savic, Rolf Kraemer, and Marek Junghans. Applying tire pressure monitoring devices for traffic management purposes. In *ISSSE*. IEEE, 2012.
- [256] Steve Krug. Don't make me think, revisited: A common sense approach to web usability. *New Riders*, 2014.
- [257] Markus G Kuhn. *The euriion constellation*, 2002.
- [258] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv*, 2018.
- [259] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: a real-world evaluation of anti-phishing training. In *SOUPS*, 2009.
- [260] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting people from phishing: the design and evaluation of an embedded training email system. In *SIGCHI*, 2007.
- [261] Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In *APWG eCrime Researchers Summit*, 2007.
- [262] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Lessons from a real world evaluation of anti-phishing training. In *APWG eCrime Researchers Summit*. IEEE, 2008.

- [263] Ponnuram Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *TOIT*, 2010.
- [264] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *WebConf*, 2010.
- [265] Jordi Laguarda, Ferran Hueto, and Brian Subirana. COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings. *OJEMB*, 2020.
- [266] Gerold Laimer and Andreas Uhl. Key-dependent jpeg2000-based robust hashing for secure image authentication. *EURASIP Journal on Information Security*, 2008.
- [267] Marc Langheinrich. Privacy by design—principles of privacy-aware ubiquitous systems. In *International conference on Ubiquitous Computing*. Springer, 2001.
- [268] Marc Langheinrich. A privacy awareness system for ubiquitous computing environments. In *UbiComp*. Springer, 2002.
- [269] Marc Langheinrich. *Personal privacy in ubiquitous computing: Tools and system support*. PhD thesis, ETH Zurich, 2005.
- [270] Dominic Lasorsa, Seth Lewis, and Avery Holton. Normalizing Twitter: Journalism practice in an emerging Communication Space. *Journalism studies*, 2012.
- [271] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational Social Science. *Science*, 2009.
- [272] Scott Lederer, Anind K Dey, and Jennifer Mankoff. Everyday privacy in ubiquitous computing environments. In *UbiComp Workshop*, 2002.
- [273] Hosub Lee and Alfred Kobsa. Confident privacy decision-making in iot environments. *TOCHI*, 2019.
- [274] Mikko Lehtonen, Antti Ruhanen, Florian Michahelles, and Elgar Fleisch. Serialized tid numbers-a headache or a blessing for rfid crackers? In *international conference on RFID*. IEEE, 2009.
- [275] Huaxin Li, Zheyu Xu, Haojin Zhu, Di Ma, Shuai Li, and Kai Xing. Demographics inference through wi-fi network traffic analysis. In *INFOCOM*. IEEE, 2016.
- [276] Ching-Yung Lin and Shih-Fu Chang. Distortion modeling and invariant extraction for digital image print-and-scan process. In *Int. Symp. Multimedia Information Processing*. Citeseer, 1999.
- [277] Guo-Yau Lin, Jimmy M Grice, Jan P Allebach, George T-C Chiu, Wayne Bradburn, and Jeff Weaver. Banding artifact reduction in electrophotographic printers by using pulse width modulation. *Journal of Imaging Science and Technology*, 2002.
- [278] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: understanding users’ mental models of mobile app privacy through crowdsourcing. In *Conference on Ubiquitous Computing*. ACM, 2012.
- [279] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The privacy policy landscape after the gdpr. *arXiv*, 2018.
- [280] Jenna Lindqvist. New challenges to personal data processing agreements: is the GDPR fit to deal with contract, accountability and liability in a world of the Internet of Things? *International Journal of Law and Information Technology*, 2018.
- [281] Jenna Lindqvist. New challenges to personal data processing agreements: is the GDPR fit to deal with contract, accountability and liability in a world of the Internet of Things? *IJLIT*, 2018.

- [282] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhiemedi, Shikun Zhang, Norman Sadeh, Alessandro Acquisti, and Yuvraj Agarwal. Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions. *SOUPS*, 2016.
- [283] Gang Liu, Guang Xiang, Bryan A Pendleton, Jason I Hong, and Wenyin Liu. Smartening the crowds: computational techniques for improving human verification to fight phishing scams. In *SOUPS*, 2011.
- [284] Haoyue Liu, Ishani Chatterjee, MengChu Zhou, Xiaoyu Sean Lu, and Abdullah Abusorrah. Aspect-based sentiment analysis: A survey of deep learning methods. *CSS*, 2020.
- [285] Andrew J Lohn. Downscaling attack and defense: Turning what you see back into what you get. *arXiv*, 2020.
- [286] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie WieBner. Microscopic traffic simulation using sumo. In *ITSC*, 2018.
- [287] Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. Digital media and democracy: a systematic review of causal and correlational evidence worldwide. *SocArXiv*, 2021.
- [288] Francisco Lupiáñez-Villanueva, Alba Boluda, Francesco Bogliacino, Giovanni Liva, Lucie Lechardoy, and T Rodríguez de las Heras Ballell. Behavioural study on unfair commercial practices in the digital environment: dark patterns and manipulative personalisation. *European Commission*, 2022.
- [289] Lauren Lutzke, Caitlin Drummond, Paul Slovic, and Joseph Árvai. Priming critical thinking: Simple interventions limit the influence of fake news about climate change on facebook. *Global environmental change*, 2019.
- [290] LeGouvernement Luxembourg. Parc automobile du luxembourg. <https://data.public.lu/fr/datasets/parc-automobile-du-luxembourg/>, 2019. Accessed: 2019-12.
- [291] John C Mace. Printer identification techniques and their privacy implications. *School of Computing Science Technical Report Series*, 2010.
- [292] Dominique Machuletz and Rainer Böhme. Multiple purposes, multiple problems: A user study of consent dialogs after gdpr. *arXiv*, 2019.
- [293] Natasha Mack, Cynthia Woodsong, Kathleen MacQueen, Greg Guest, and Emily Namey. Qualitative research methods: A data collector’s field guide. *FHI*, 2005.
- [294] Mary Madden. Public perceptions of privacy and security in the post-snowden era. *Pew Research Center*, 2014.
- [295] Silvia Majó-Vázquez, Mariluz Congosto, Tom Nicholls, and Rasmus Kleis Nielsen. The role of suspended accounts in political discussion on social media: Analysis of the 2017 french, uk and german elections. *Social Media+ Society*, 2021.
- [296] Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. Privacy awareness about information leakage: Who knows what about me? In *Workshop on privacy in the electronic society*. ACM, 2013.
- [297] Momin M Malik and Jürgen Pfeffer. A macroscopic analysis of news content in twitter. *Digital Journalism*, 2016.
- [298] Shrirang Mare, Franziska Roesner, and Tadayoshi Kohno. Smart Devices in Airbnbs: Considering Privacy and Security for both Guests and Hosts. *PoPETs*, 2020.
- [299] Karola Marky, Alexandra Voit, Alina Stöver, Kai Kunze, Svenja Schröder, and Max Mühlhäuser. I don’t know how to protect myself: Understanding Privacy Perceptions Resulting from the Presence of Bystanders in Smart Environments. In *NordiCHI*, 2020.

- [300] Philipp K Masur. How online privacy literacy supports self-data protection and self-determination in the age of information. *Media and Communication*, 2020.
- [301] J. Matas, C. Galambos, and J.V. Kittler. Robust detection of lines using the progressive probabilistic hough transform. *CVIU*, 2000.
- [302] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do cookie banners respect my choice?: Measuring legal compliance of banners from iab europe’s transparency and consent framework. In *Symposium on Security and Privacy (SP)*. IEEE, 2020.
- [303] Philipp Mayring. Qualitative content analysis. *A companion to qualitative research*, 2004.
- [304] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Isjlp*, 2008.
- [305] Wes McKinney. Data structures for statistical computing in python. In *Python in Science Conference*, 2010.
- [306] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual review of sociology*, 2001.
- [307] Paul Mena. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet*, 2020.
- [308] Panagiotis Metaxas and Samantha T Finn. The infamous# pizzagate conspiracy theory: Insight from a twittertrails investigation. *Computation and Journalism*, 2017.
- [309] Panagiotis Metaxas, Eni Mustafaraj, Kily Wong, Laura Zeng, Megan O’Keefe, and Samantha Finn. What Do Retweets Indicate? Results From User Survey and Meta-Review of Research. In *AAAI*, 2015.
- [310] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. Social and heuristic approaches to credibility evaluation online. *Journal of communication*, 2010.
- [311] Ulrike Meyer and Vincent Drury. Certified phishing: taking a look at public key certificates of phishing websites. In *SOUPS 2019*, 2019.
- [312] Nicholas Micallef, Mihai Avram, Filippo Menczer, and Sameer Patil. Fakey: A Game Intervention to Improve News Literacy on Social Media. *HCI*, 2021.
- [313] Aravind K Mikkilineni, Osman Arslan, Pei-Ju Chiang, Roy M Kumontoy, Jan P Allebach, George T-C Chiu, and Edward J Delp. Printer forensics using svm techniques. In *NIP & Digital Fabrication Conference*. Society for Imaging Science and Technology, 2005.
- [314] Aravind K. Mikkilineni, Pei-Ju Chiang, George TC Chiu, Jan P. Allebach, and Edward J. Delp. Channel model and operational capacity analysis of printed text documents. In *Electronic Imaging*. International Society for Optics and Photonics, 2007.
- [315] Aravind K Mikkilineni, Nitin Khanna, and Edward J Delp. Forensic printer detection using intrinsic signatures. In *Media Watermarking, Security, and Forensics III*. International Society for Optics and Photonics, 2011.
- [316] Mateusz Mikusz, Steven Houben, Nigel Davies, Klaus Moessner, and Marc Langheinrich. Raising awareness of IoT sensor deployments. *Living in the Internet of Things: Cybersecurity of the IoT*, 2018.
- [317] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *CSUR*, 2021.
- [318] Vishal Monga et al. Robust and secure image hashing via non-negative matrix factorizations. *Transactions on Information Forensics and Security*, 2007.
- [319] Vishal Monga and Brian L Evans. Robust perceptual image hashing using feature points. In *ICIP*. IEEE, 2004.

- [320] Vishal Monga and Brian L Evans. Perceptual image hashing via feature points: performance evaluation and tradeoffs. *Transactions on Image Processing*, 2006.
- [321] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. *arXiv*, 2019.
- [322] Barrington Moore. *Privacy: Studies in social and cultural history*. Routledge, 2017.
- [323] Patricia Moravec, Antino Kim, and Alan Dennis. Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media. *Information Systems Research*, 2020.
- [324] Victor Morel, Mathieu Cunche, and Daniel Le Métayer. A generic information and consent framework for the IoT. In *TrustCom/BigDataSE*, 2019.
- [325] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. Is the sample good enough? comparing data from twitter’s streaming api with twitter’s firehose. In *Proceedings of AAAI ICWSM*, 2013.
- [326] Fred Morstatter, Yunqiu Shao, Aram Galstyan, and Shanika Karunasekera. From alt-right to alt-rechts: Twitter analysis of the 2017 german federal election. In *Companion Proceedings of the The Web Conference*, 2018.
- [327] Jordan T Moss and Peter J O’Connor. Political correctness and the alt-right: The development of extreme political attitudes. *PloS one*, 2020.
- [328] Naveed Mufti and Syed Afaq Ali Shah. Automatic number plate recognition: A detailed survey of relevant algorithms. *Sensors*, 2021.
- [329] Patrick Murmann and Simone Fischer-Hübner. Tools for achieving usable ex post transparency: a survey. *IEEE Access*, 2017.
- [330] Patrick Murmann and Simone Fischer-Hübner. Usable transparency enhancing tools: A literature review. *Karlstads universitet*, 2017.
- [331] Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. In *Principles and practice of multi-agent systems*. Springer, 2015.
- [332] Nona Naderi and Graeme Hirst. Classifying frames at the sentence level in news articles. *Policy*, 2017.
- [333] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Privacy expectations and preferences in an IoT world. In *SOUPS*, 2017.
- [334] Efrat Neter and Gershon Ben-Shakhar. The predictive validity of graphological inferences: A meta-analytic approach. *Personality and Individual differences*, 1989.
- [335] L Neudert, Bence Kollanyi, and Philip N Howard. Junk news and bots during the german parliamentary election: What are german voters sharing over twitter? *The Computational Propaganda Project*, 2017.
- [336] Nic Newman, Richard Fletcher, Kirsten Eddy, Craig Robertson, and Rasmus Nielsen. Reuters institute digital news report 2023. Reuters Institute for the Study of Journalism, 2023.
- [337] Nic Newman, Richard Fletcher, Anna Schulz, Simge Andi, and Rasmus Nielsen. Reuters Institute Digital News Report. Reuters Institute for the Study of Journalism, 2020.
- [338] Thong Q. Nguyen, Yves Delignon, Lionel Chagas, and François Septier. Printer identification from micro-metric scale printing. In *ICASSP*. IEEE, 2014.
- [339] Patricia A Norberg, Daniel R Horne, and David A Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of consumer affairs*, 2007.
- [340] Donald A Norman. *The psychology of everyday things*. Basic books, 1988.

- [341] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *CHI*, 2020.
- [342] Jonathan A Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 2020.
- [343] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, Serge Egelman, et al. On the ridiculousness of notice and consent: Contradictions in app privacy policies. In *Workshop on Technology and Consumer Protection, in conjunction with the 39th IEEE Symposium on Security and Privacy*, 2019.
- [344] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *CHI*, 2017.
- [345] John Oliver and Joyce Chen. Use of signature analysis to discriminate digital printing technologies. In *NIP & Digital Fabrication Conference*, volume 2002. Society for Imaging Science and Technology, 2002.
- [346] Walter Ong. *Orality and Literacy*. Routledge, 202.
- [347] Claudia Orellana-Rodriguez, Derek Greene, and Mark Keane. Spreading One’s Tweets: How Can Journalists Gain Attention for their Tweeted News? *The Journal of Web Science*, 2017.
- [348] N Otsu. A threshold selection method from gray-scale histogram. *IEEE Transactions on Systems Man and Cybernetics*, 1979.
- [349] Aditya Pal and Scott Counts. Identifying Topical Authorities in Microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011.
- [350] Sungkyu Park, Jaimie Yejean Park, Jeong-han Kang, and Meeyoung Cha. The presence of unexpected biases in online fact-checking. *The Harvard Kennedy School Misinformation Review*, 2021.
- [351] Thomas Paul, Martin Stopczynski, Daniel Puscher, Melanie Volkamer, and Thorsten Strufe. C4ps-helping facebookers manage their privacy settings. In *International Conference on Social Informatics*. Springer, 2012.
- [352] Bogdan Pavliy and Jonathan Lewis. The Performance of Twitter’s Language Detection Algorithm and Google’s Compact Language Detector on Language Detection in Ukrainian and Russian Tweets. *Bulletin of Toyama University of International Studies*, 2016.
- [353] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 2020.
- [354] Gordon Pennycook and David G Rand. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 2020.
- [355] Iryna Pentina and Monideepa Tarafdar. From “information” to “knowing”: Exploring the role of social media in contemporary news consumption. *Computers in human behavior*, 2014.
- [356] Alfredo J Perez, Sherali Zeadally, Luis Y Matos Garcia, Jaouad A Mouloud, and Scott Griffith. Facepet: Enhancing bystanders’ facial privacy with smart wearables/internet of things. *Electronics*, 2018.
- [357] Jonathan Petit, Florian Schaub, Michael Feiri, and Frank Kargl. Pseudonym schemes in vehicular networks: A survey. *IEEE communications surveys & tutorials*, 2014.
- [358] Irene Pollach. What’s wrong with online privacy policies? *Communications of the ACM*, 2007.

- [359] Colin Porlezza. *Gefährdete journalistische Unabhängigkeit: zum wachsenden Einfluss von Werbung auf redaktionelle Inhalte*. Herbert von Halem Verlag, 2014.
- [360] Sarah Prange, Ahmed Shams, Robin Piening, Yomna Abdelrahman, and Florian Alt. PriView-Exploring Visualisations to Support Users' Privacy Awareness. In *CHI*, 2021.
- [361] Bernhard Preim and Raimund Dachseht. *Interaktive systeme: Band 1: Grundlagen, Graphical User Interfaces, Informationsvisualisierung*. Springer-Verlag, 2010.
- [362] Associated Press. Social Media Guidelines for AP Employees, 2013.
- [363] Robert W Proctor, M Athar Ali, and Kim-Phuong L Vu. Examining usability of web privacy policies. *Intl. Journal of Human-Computer Interaction*, 2008.
- [364] Swapan Purkait. Phishing counter measures and their effectiveness—literature review. *ICS*, 2012.
- [365] Philip Raschke, Axel Küpper, Olha Drozd, and Sabrina Kirrane. Designing a gdpr-compliant and usable privacy dashboard. In *IFIP international summer school on privacy and identity management*. Springer, 2017.
- [366] Steve Rathje, Jay J. Van Bavel, and Sander van der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 2021.
- [367] Adrian Rauchfleisch and Jonas Kaiser. The false positive problem of automatic bot detection in social science research. *Berkman Klein Center Research Publication*, 2020.
- [368] Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. Breaking down walls of text: How can nlp benefit consumer privacy? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.
- [369] Abbas Razaghpanah, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Christian Kreibich, Phillipa Gill, Mark Allman, and Vern Paxson. Haystack: In situ mobile traffic analysis in user space. *arXiv*, 2015.
- [370] Andrew G Reece and Christopher M Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 2017.
- [371] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. Why doesn't jane protect her privacy? In *PETs*. Springer, 2014.
- [372] Reuters. Reporting From the Internet and Using Social Media. Reuters, Thomson, 2013.
- [373] Ricoh. Going global, company history about ricoh, 2017.
- [374] Michael Rogers and Saleem Bhatti. How to disappear completely: A survey of private peer-to-peer networks. *networks*, 2007.
- [375] Ana-Cristina Rogoz, Mihaela Gaman, and Radu Tudor Ionescu. Saroco: Detecting satire in a novel romanian corpus of news articles. *arXiv*, 2021.
- [376] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. Prebunking interventions based on the psychological theory of “inoculation” can reduce susceptibility to misinformation across cultures. *The Harvard Kennedy School (HKS) Misinformation Review*, 2020.
- [377] Ishtiaq Rouf, Rob Miller, Hossen Mustafa, Travis Taylor, Sangho Oh, Wenyan Xu, Marco Gruteser, Wade Trappe, and Ivan Seskar. Security and privacy vulnerabilities of {In-Car} wireless networks: A tire pressure monitoring system case study. In *USENIX*, 2010.
- [378] Abhishek Roy and Sunil Karforma. A survey on digital signatures and its applications. *Journal of Computer and Information Technology*, 2012.

- [379] Scott Ruoti, Tyler Monson, Justin Wu, Daniel Zappala, and Kent Seamons. Weighing context and trade-offs: How suburban adults selected their online security posture. In *SOUPS*, 2017.
- [380] Seung-Jin Ryu, Hae-Yeoun Lee, Dong-Hyuck Im, Jung-Ho Choi, and Heung-Kyu Lee. Electrophotographic printer identification by halftone texture analysis. In *ICASSP*. IEEE, 2010.
- [381] Niloofar Safi Samghabadi, Parth Patwa, PYKL Srinivas, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and misogyny detection using bert: A multi-task approach. In *Workshop on Trolling, Aggression and Cyberbullying*, 2020.
- [382] Krishna Sampigethaya and Radha Poovendran. A survey on mix networks and their secure applications. *Proceedings of the IEEE*, 2006.
- [383] Nemanja Savić, Marek Junghans, and Miloš Krstić. Evaluating tire pressure monitoring system for traffic management purposes-simulation study. In *ITSC*. IEEE, 2014.
- [384] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Faith Cranor. Watching them watching me: Browser extensions’ impact on user privacy awareness and concern. In *NDSS workshop on usable security*, 2016.
- [385] Tatjana Scheffler. A German Twitter Snapshot. In *LREC*, 2014.
- [386] Marco Schreyer, Christian Schulze, Armin Stahl, and Wolfgang Effelsberg. Intelligent Printing Technique Recognition and Photocopy Detection for Forensic Document Examination. In *Informatiktage*, 2009.
- [387] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 2013.
- [388] Wolfgang Schweiger. Nachrichtenjournalismus, alternative und soziale medien. *Der (des) informierte Bürger im Netz: Wie soziale Medien die Meinungsbildung verändern*, 2017.
- [389] Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. Your installed apps reveal your gender and more! *SIGMOBILE*, 2015.
- [390] Ricky Sethi. Crowdsourcing the verification of fake news and alternative facts. In *Conference on Hypertext and Social Media*. ACM, 2017.
- [391] William Seymour, Martin J Kraemer, Reuben Binns, and Max Van Kleek. Informing the design of privacy-empowering tools for the connected home. In *CHI*, 2020.
- [392] Syed W Shah and Salil S Kanhere. Recent trends in user authentication—a survey. *IEEE access*, 2019.
- [393] Altaf Shaik, Ravishankar Borgaonkar, N Asokan, Valtteri Niemi, and Jean-Pierre Seifert. Practical attacks against privacy and availability in 4g/lte mobile communication systems. *arXiv*, 2015.
- [394] Shize Shang and Xiangwei Kong. Printer and scanner forensics. *Handbook of Digital Forensics of Multimedia Data and Devices*, 2015.
- [395] Shize Shang, Xiangwei Kong, and Xingang You. Document forgery detection using distortion mutation of geometric parameters in characters. *Journal of Electronic Imaging*, 2015.
- [396] Shize Shang, Nasir Memon, and Xiangwei Kong. Detecting documents forged by printing and copying. *EURASIP*, 2014.
- [397] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *Transactions on Intelligent Systems and Technology (TIST)*, 2019.

- [398] Rahul Anand Sharma, Elahe Soltanaghaei, Anthony Rowe, and Vyas Sekar. Lumos: Identifying and localizing diverse hidden IoT devices in an unfamiliar environment. In *USENIX*, 2022.
- [399] Rashmi Sharma, T. R. Baggi, Amit Chattree, Lav Kesharwani, and A. K. Gupta. Detection and identification of printer inks-a review report on laser and inkjet printer ink analysis. *International Journal of Current Research and Review*, 2013.
- [400] Peter Shaw, Mateusz Mikusz, Petteri Nurmi, and Nigel Davies. IoT Maps: Charting the Internet of Things. In *HotMobile*, 2019.
- [401] Elisa Shearer and Amy Mitchell. News use across social media platforms in 2020. *Pew Research Center*, 2021.
- [402] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *SIGCHI*, 2010.
- [403] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *SOUPS*, 2007.
- [404] Fatemeh Shirazi and Melanie Volkamer. What deters jane from preventing identification and tracking on the web? In *Workshop on Privacy in the Electronic Society*, 2014.
- [405] Ben Shneiderman and Catherine Plaisant. *Designing the user interface: Strategies for effective human-computer interaction*. Pearson Education India, 2010.
- [406] Ben Shneiderman, Catherine Plaisant, Maxine S Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. *Designing the user interface: strategies for effective human-computer interaction*. Pearson, 2016.
- [407] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD*, 2017.
- [408] Sandra Siby, Rajib Ranjan Maiti, and Nils Tippenhauer. Iotscanner: Detecting and classifying privacy threats in iot neighborhoods. *arXiv*, 2017.
- [409] Ana Lgia Silva de Lima, Luc JW Evers, Tim Hahn, Lauren Bataille, Jamie L Hamilton, Max A Little, Yasuyuki Okuma, Bastiaan R Bloem, and Marjan J Faber. Freezing of gait and fall detection in parkinson’s disease using wearable sensors: a systematic review. *Journal of Neurology*, 2017.
- [410] Marija Slavkovik, Clemens Stachl, Caroline Pitman, and Jonathan Askonas. Digital voodoo dolls. In *AAAI/ACM*, 2021.
- [411] Daniel J Solove. I’ve got nothing to hide and other misunderstandings of privacy. *San Diego L. Rev.*, 2007.
- [412] Yunpeng Song, Yun Huang, Zhongmin Cai, and Jason I Hong. I’m All Eyes and Ears: Exploring Effective Locators for Privacy Awareness in IoT Scenarios. In *CHI*, 2020.
- [413] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrcio Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In *Conference on fairness, accountability and transparency*, 2018.
- [414] Sargur N. Srihari and Venugopal Govindaraju. Analysis of textual images using the hough transform. *IEEE Transactions on Systems Man and Cybernetics*, 1989.
- [415] Sukamol Srikwan and Markus Jakobsson. Using cartoons to teach internet security. *Cryptologia*, 2008.

- [416] Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D Gosling, Gabriella M Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, et al. Predicting personality from patterns of behavior collected with smartphones. *PNAS*, 2020.
- [417] Charlotte Stanton, Thomas Kadri, David Danks, Jack Parker, and Megan Metzger. *The Legal, Ethical, and Efficacy Dimensions of Managing Synthetic and Manipulated Media*. Carnegie Endowment for International Peace, 2019.
- [418] Charlotte Stanton, Thomas E. Kadri, David Danks, Jack Parker, and Megan Metzger. The legal, ethical, and efficacy dimensions of managing synthetic and manipulated media. *Carnegie Endowment for International Peace*, 2019.
- [419] Martin Steinebach, Huajian Liu, and York Yannikos. Efficient cropping-resistant robust image hashing. In *Availability, Reliability and Security (ARES)*. IEEE, 2014.
- [420] Nili Steinfeld. “i agree to the terms and conditions”:(how) do users read privacy policies online? an eye-tracking experiment. *Computers in human behavior*, 2016.
- [421] David Sterrett, Dan Malato, Jennifer Benz, Liz Kantor, Trevor Thompson, Tom Rosenstiel, Jeff Sonderman, and Kevin Loker. Who shared it?: Deciding what news to trust on social media. *Digital journalism*, 2019.
- [422] Stefan Stieglitz, Florian Brachten, Björn Ross, and Anna-Katharina Jung. Do social bots dream of electric sheep? a categorisation of social media bot accounts. *arXiv*, 2017.
- [423] Sebastian Stier, Arnim Bleier, Haiko Lietz, and Markus Strohmaier. Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political communication*, 2018.
- [424] Simon Stockhardt, Benjamin Reinheimer, Melanie Volkamer, Peter Mayer, Alexandra Kunz, Philipp Rack, and Daniel Lehmann. Teaching phishing-security: which way is best? In *IFIP*. Springer, 2016.
- [425] Gianluca Stringhini, Oliver Hohlfeld, Christopher Kruegel, and Giovanni Vigna. The harvester, the botmaster, and the spammer: On the relations between the different actors in the spam landscape. In *ACM CCS*, 2014.
- [426] Sungjoo Suh, Jan P Allebach, George T-C Chiu, and Edward J Delp. Printer mechanism-level information embedding and extraction for halftone documents—new results. In *NIP & Digital Fabrication Conference*. Society for Imaging Science and Technology, 2007.
- [427] Mingshen Sun, Tao Wei, and John C.S. Lui. TaintART: A Practical Multi-level Information-Flow Tracking System for Android RunTime. In *SIGSAC*, Vienna Austria, 2016. ACM.
- [428] Rui Sun and Wenjun Zeng. Secure and robust image hashing via compressive sensing. *Multimedia tools and applications*, 2014.
- [429] Ali Sunyaev, Tobias Dehling, Patrick L Taylor, and Kenneth D Mandl. Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association*, 2015.
- [430] S. Suzuki and K. Abe. Topological Structural Analysis of Digitized Binary Images by Border Following. *CVGIP*, 1985.
- [431] Ashwin Swaminathan, Yinian Mao, and Min Wu. Robust and secure image hashing. *Transactions on Information Forensics and security*, 2006.
- [432] Adobe Systems, editor. *PostScript language reference*. Addison-Wesley, 1999.
- [433] Małgorzata Szafarska, Renata Wietecha-Posłuszny, Michał Woźniakiewicz, and Paweł Kościelniak. Application of capillary electrophoresis to examination of color inkjet printing inks for forensic purposes. *Forensic Science International*, 2011.

- [434] Alexander Sangerlaub, Miriam Meier, and Wolf-Dieter Ruhl. Fakten statt fakes - verursacher, verbreitungswege und wirkungen von fake news im bundestagswahlkampf 2017. *Stiftung Neue Verantwortung*, 2018.
- [435] Seyed Amir Hossein Tabatabaei, Obaid Ur-Rehman, Natasa Zivic, and Christoph Ruland. Secure and robust two-phase image authentication. *Transactions on Multimedia*, 2015.
- [436] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv*, 2017.
- [437] Edson Tandoc, Darren Lim, and Rich Ling. Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, 2020.
- [438] Zhenjun Tang, Ziqing Huang, Xianquan Zhang, and Huan Lao. Robust image hashing with multidimensional scaling. *Signal Processing*, 2017.
- [439] Zhenjun Tang, Xianquan Zhang, Xianxian Li, and Shichao Zhang. Robust image hashing with ring partition and invariant vector distance. *Transactions on Information Forensics and Security*, 2016.
- [440] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In *WWW*, 2018.
- [441] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deep-fakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 2020.
- [442] Catherine Tong, Gabriella M Harari, Angela Chieh, Otmane Bellahsen, Matthieu Vegreville, Eva Roitmann, and Nicholas D Lane. Inference of big-five personality using large-scale networked mobile and appliance data. In *MobiSys*, 2018.
- [443] Vuma Touchpoints. Den Markt im Blick - Basisinformationen fur fundierte Mediaentscheidungen., 2022.
- [444] Sabine Trepte, Doris Teutsch, Philipp K Masur, Carolin Eicher, Mona Fischer, Alisa Hennhofer, and Fabienne Lind. Do people know about privacy and data protection strategies? towards the "online privacy literacy scale" (oplis). In *Reforming European data protection law*. Springer, 2015.
- [445] Carmela Troncoso, Enrique Costa-Montenegro, Claudia Diaz, and Stefan Schiffner. On the difficulty of achieving anonymity for vehicle-2-x communication. *Computer Networks*, 2011.
- [446] Min-Jen Tsai, Chien-Lun Hsu, Jin-Sheng Yin, and Imam Yuadi. Digital forensics for printed character source identification. In *International Conference on Multimedia and Expo*. IEEE, 2016.
- [447] Min-Jen Tsai and Jung Liu. Digital forensics for printed source identification. In *International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2013.
- [448] Min-Jen Tsai, Jung Liu, Chen-Sheng Wang, and Ching-Hua Chuang. Source color laser printer identification using discrete wavelet transform and feature selection algorithms. In *International Symposium on Circuits and Systems*. IEEE, 2011.
- [449] Min-Jen Tsai and Imam Yuadi. Digital forensics of microscopic images for printed source identification. *Multimedia Tools and Applications*, 2018.
- [450] Min-Jen Tsai, Mam Yuadi, Yu-Han Tao, and Jin-Sheng Yin. Source identification for printed documents. In *International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2017.
- [451] Jason Tuohey. Government Uses Color Laser Printer Technology to Track Documents. *PCWorld*, 2004.
- [452] Ikhlaz ur Rehman. Facebook-cambridge analytica data harvesting: What you need to know. *Library Philosophy and Practice*, 2019.

- [453] André Calero Valdez and Martina Ziefle. The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies*, 2019.
- [454] Joost van Beusekom, Faisal Shafait, and Thomas M. Breuel. Combined orientation and skew detection using geometric text-line modeling. *IJDAR*, 2010.
- [455] Joost van Beusekom, Faisal Shafait, and Thomas M. Breuel. Automatic authentication of color laser print-outs using machine identification codes. *Pattern Analysis and Applications*, 2013.
- [456] Joost Van Beusekom, Faisal Shafait, and Thomas M Breuel. Text-line examination for document forgery detection. *International Journal on Document Analysis and Recognition (IJDAR)*, 2013.
- [457] GW Van Blarckom, John J Borking, and JG Eddy Olk. Handbook of privacy and privacy-enhancing technologies. *PISA*, 2003.
- [458] Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J. Weitzner, and Nigel Shadbolt. Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps. In *CHI*. ACM, 2017.
- [459] Tran Van Lanh, Kai-Sen Chong, Sabu Emmanuel, and Mohan S Kankanhalli. A survey on digital camera image forensic methods. In *International Conference on Multimedia and Expo*. IEEE, 2007.
- [460] Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. Dating documents using graph convolution networks. In *Annual Meeting of the Association for Computational Linguistics*, 2018.
- [461] Sankaranarayanan Velupillai and Levent Guvenc. Tire pressure monitoring [applications of control]. *IEEE Control systems magazine*, 2007.
- [462] Ramarathnam Venkatesan, S-M Koon, Mariusz H Jakubowski, and Pierre Moulin. Robust image hashing. In *Image Processing*. IEEE, 2000.
- [463] Luisa Vervier, Eva-Maria Zeissig, Chantal Lidynia, and Martina Ziefle. Perceptions of digital footprints and the value of privacy. In *IoTBDS*, 2017.
- [464] Nisha Vinayaga-Sureshkanth, Raveen Wijewickrama, Anindya Maiti, and Murtuza Jadliwala. Security and privacy challenges in upcoming intelligent urban micromobility transportation systems. In *Workshop on Automotive and Aerial Vehicle Security*. ACM, 2020.
- [465] Melanie Volkamer, Karen Renaud, Oksana Kulyk, and Sinem Emeröz. A socio-technical investigation into smartphone security. In *International Workshop on Security and Trust Management*. Springer, 2015.
- [466] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of torpedo: Tooltip-powered phishing email detection. *Computers & Security*, 2017.
- [467] Max von Grafenstein, Isabel Kiefaber, Julie Heumüller, Valentin Rupp, Paul Graßl, Otto Kolless, and Zsófia Puzst. Privacy icons as a component of effective transparency and controls under the gdpr: effective data protection by design based on art. 25 gdpr. *Computer Law & Security Review*, 2024.
- [468] Sandra Wachter. Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. *Computer Law & Security Review*, 2018.
- [469] Changyou Wang, Xiangwei Kong, Shize Shang, and Xin'gang You. Photocopier forensics based on arbitrary text characters. In *Media Watermarking, Security, and Forensics*, 2013.
- [470] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *J. Pers. Soc. Psychol.*, 2018.
- [471] Audrey Watters. How recent changes to twitter's terms of service might hurt academic research. *Read Write*, 2011.

- [472] Maarten Wegdam and Dirk-Jaap Plas. Empowering users to control their privacy in context-aware systems through interactive consent. *CTIT*, 2008.
- [473] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitingner, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L Mazurek, and Blase Ur. What twitter knows: Characterizing ad targeting practices, user perceptions, and ad explanations through users’ own twitter data. In *USENIX*, 2020.
- [474] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L Mazurek, and Blase Ur. Oh, the places you’ve been! user reactions to longitudinal transparency about third-party web tracking and inferencing. In *SIGSAC*, 2019.
- [475] Siegfried Weischenberg, Maja Malik, and Armin Scholl. Journalismus in Deutschland 2005. *Media Perspektiven*, 2006.
- [476] Mark Weiser. The Computer for the 21 st Century. *Sci. Am.*, 1991.
- [477] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What. hack: engaging anti-phishing training through a role-playing phishing simulation game. *CHI*, 2019.
- [478] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y Zhao. ’hello, it’s me’: Deep learning-based speech synthesis attacks in the real world. In *SIGSAC*. ACM, 2021.
- [479] M Wermuth, C Neef, R Wirth, I Hanitz, H Löhner, H Hautzinger, W Stock, M Pfeifer, M Fuchs, B Lenz, et al. Kraftfahrzeugverkehr in Deutschland 2010. *WVI, IVT, DLR, and KBA*, 2010.
- [480] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 1968.
- [481] Andrew Whitmore, Anurag Agarwal, and Li Da Xu. The internet of Things-A survey of topics and trends. *Inf Syst Front*, 2015.
- [482] Björn Wiedersheim, Zhendong Ma, Frank Kargl, and Panos Papadimitratos. Privacy in inter-vehicular networks: Why simple pseudonym change is not enough. In *WONS*, 2010.
- [483] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. Evaluating Information: The Cornerstone of Civic Online Reasoning. *Stanford Digital Repository*, 2016.
- [484] Chloe Wittenberg, Adam J Berinsky, Nathaniel Persily, and Joshua A Tucker. Misinformation and its correction. *Social media and democracy: The state of the field, prospects for reform*, 2020.
- [485] Lior Wolf, Liza Potikha, Nachum Dershowitz, Roni Shweka, and Yaacov Choueka. Computerized paleography: tools for historical manuscripts. In *International Conference on Image Processing*. IEEE, 2011.
- [486] Di Wu, Xuebing Zhou, and Xiamu Niu. A novel image hash algorithm resistant to print–scan. *Signal processing*, 2009.
- [487] Han Wu, Xiangwei Kong, and Shize Shang. A printer forensics method using halftone dot arrangement model. In *China Summit and International Conference on Signal and Information Processing*, 2015.
- [488] Wen Wu, Li Chen, Qingchang Yang, and You Li. Inferring students’ personality from their communication behavior in web-based learning systems. *IJAIED*, 2019.
- [489] Yubao Wu, Xiangwei Kong, Xin’gang You, and Yiping Guo. Printer forensics based on page document’s geometric distortion. In *ICIP*. IEEE, 2009.
- [490] Yuxi Wu, Panya Gupta, Miranda Wei, Yasemin Acar, Sascha Fahl, and Blase Ur. Your secrets are safe: How browsers’ explanations impact misconceptions about private browsing mode. In *WWW*, 2018.
- [491] Shijun Xiang, Hyoung-Joong Kim, and Jiwu Huang. Histogram-based image hashing scheme robust against geometric deformations. In *Proceedings of the 9th workshop on Multimedia & security*. ACM, 2007.

- [492] Luo Xiao, Qinghu Chen, and Yuchen Yan. Printed Characters Texture Identification Based on Two-factor Analysis. *Journal of Computational Information Systems*, 2015.
- [493] Sherif M Yacoub, Steven J Simske, Xiaofan Lin, and John Burns. Recognition of emotions in interactive voice response systems. In *Interspeech*, 2003.
- [494] Bian Yang, Fan Gu, and Xiamu Niu. Block mean value based image perceptual hashing. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*. IEEE, 2006.
- [495] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy perceptions and designs of bystanders in smart homes. *CHI*, 2019.
- [496] William S Yerazunis. The spam-filtering accuracy plateau at 99.9% accuracy and how to get past it. In *MIT Spam Conference*, 2004.
- [497] Razieh Nokhbeh Zaeem, Ahmad Ahbab, Josh Bestor, Hussam H Djadi, Sunny Kharel, Victor Lai, Nick Wang, and K Suzanne Barber. Privacycheck v3: Empowering users with higher-level understanding of privacy policies. In *Workshop on Privacy in the Electronic Society (WPES 21)*, 2021.
- [498] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. *IMC*, 2018.
- [499] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *J. Data and Information Quality*, 2019.
- [500] Christoph Zauner, Martin Steinebach, and Eckehard Hermann. Rihamark: perceptual image hash benchmarking. In *Media watermarking, security, and forensics III*. SPIE, 2011.
- [501] Bluma Zeigarnik. Das Behalten erledigter und unerledigter Handlungen [On finished and unfinished tasks]. *Universität Berlin*, 1927.
- [502] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.
- [503] Hui Zhang, Martin Schmucker, and Xiamu Niu. The Design and Application of PHABS: A Novel Benchmark Platform for Perceptual Hashing Algorithms. In *Multimedia and Expo*. IEEE, 2007.
- [504] Shikun Zhang, Yuanyuan Feng, Anupam Das, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Understanding people’s privacy attitudes towards video analytics technologies. *FTC PrivacyCon*, 2020.
- [505] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *CSUR*, 2020.
- [506] Sebastian Zimmeck and Steven M Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *USENIX*, 2014.
- [507] Christian Zimmermann. A categorization of transparency-enhancing technologies. *arXiv*, 2015.
- [508] Christian Zimmermann, Rafael Accorsi, and Günter Müller. Privacy dashboards: reconciling data-driven business models and privacy. In *International Conference on Availability, Reliability and Security*. IEEE, 2014.
- [509] Shoshana Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile books, 2019.