Mobility Data Anonymization – A Literature Review and an Industry-Driven Survey

Maximilian Walter, Wasilij Beskorovajnov, Fridtjof Lieberwirth, Jan Sürmeli, Pascal Zwick, Robert Heinrich
Please note:
This Report has been published on the Internet under the following Creative Commons License:
http://creativecommons.org/licenses/by-nc-nd/4.0/de.
Mobility Data Anonymization – A Literature Review and an Industry-Driven Survey

Maximilian Walter\textsuperscript{KIT}, maximilian.walter@kit.edu
Wasilij Beskorovajnov\textsuperscript{FZI}, beskorovajnov@fzi.de
Fridtjof Lieberwirth\textsuperscript{DFE}, lieberwirth@dresearch-fe.de
Jan Sürmeli\textsuperscript{FZI}, suermeli@fzi.de
Pascal Zwick\textsuperscript{FZI}, zwick@fzi.de
Robert Heinrich\textsuperscript{KIT}, robert.heinrich@kit.edu

August 2023

\textbf{KIT} Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
\textbf{FZI} FZI Forschungszentrum Informatik, Karlsruhe, Germany
\textbf{DFE} DReserach Fahrzeugelektronik GmbH, Berlin, Germany

Findings from the Project “Competence Cluster Anonymization for Connected Mobility Systems” (ANYMOS)
Abstract

The transformation of mobility is on the cusp of a significant shift, driven by data-centric technologies in both individual and public transport. However, this data often contains sensitive private data, which can be used, for instance, for tracking a person. Hence, anonymization of this mobility data is important. In this report, we present a structured literature review about anonymization methods in the mobility domain. Based on our findings, we present different anonymization methods and discuss their application scenarios and characteristics for public transport and individual one. Additionally, an industry-driven survey on anonymization methods within public transport, particularly centered around video technologies, is presented. This industry-driven survey, conducted within a video surveillance solutions company, highlights current trends and underscores the necessity for continued research. This report was created within the ANYMOS project.

1 Introduction

The future of mobility is poised to undergo a profound transformation fueled by the pervasive influence of data-driven technologies in both individual and public transport. This paradigm shift is propelled by two major driving forces: the steady progression towards automated mobility systems and the strategic endeavor to scale up public transport networks. The advent of autonomous vehicles, spanning from passenger cars to shuttle buses and delivery robots, mandates a perpetual awareness of other road users' behaviors to ensure safe and efficient operations. Achieving the desired expansion of public transport necessitates leveraging existing infrastructure more efficiently, wherein meticulous data analysis enables accurate capacity forecasts and personalized user recommendations to optimize utilization.

To catalyze innovation and foster healthy competition, a variety of stakeholders must have access to mobility data. Simultaneously the privacy of individuals needs to be safeguarded, in line with the stringent requirements outlined in the EU Data Protection Regulation. In this report, mobility data is synonymous to data in the transportation section including public transportation, e.g., buses or trams, and private transportation, e.g., cars or bikes. Effective data anonymization or pseudonymization methods play a pivotal role in striking this balance. As the lines blur between individual and public transport, intelligent and adaptable mobility solutions emerge as the frontrunners, propelling traditional distinctions into the background.

This technical report is a result of the ANYMOS [1] research project and represents the findings of work package 5.1. The scope of this work package, and this report, encompasses:

1. A comprehensive literature review, following the guidelines of Kitchenham et al. [2]. The review aims to offer valuable insights and address
challenges in the pursuit of a data-powered future of mobility.

2. An industry-driven survey by DFE focusing on anonymization methods in public transport, particularly for camera-based data. The survey will list the current industrial approaches and identifies key challenges from an industry perspective.

The paper reports on both parts, and is structured as follows. We describe the configuration of our literature review in section 3 and report the results of our review in section 4. We discuss our findings and interpret them in section 5 and examine limitations in section 6. The survey performed by DFE is presented and discussed in section 7. We conclude our report in section 8.

2 Work Package Goals and Description

In the later section, we introduce our configuration for the literature review. The configuration is built upon our work package goals and requirements as written in our project proposal. In this section, we summarize the research goals of our working package 5.2 in the ANYMOS project.

This research project revolves around anonymization and aims to establish the foundation for a toolkit of methods for the ANYMOS project. The primary objective is to systematically investigate the current state of the art in the mobility domain.

To achieve this, the project entails a systematic literature review. The formulation of specific search queries for anonymization methods will be a key step. Concurrently, criteria suitable for effectively comparing the identified methods will be defined. Subsequently, these methods will be compared based on the established criteria, highlighting their respective advantages, disadvantages, and application domains.

One of the significant contributions of this work package is the development of competence in the application of anonymization methods in the mobility domain. The research findings will be summarized in a technical report, offering insights into the current state of the art. Through this analysis, the work package seeks to provide a foundation for other work packages, address limitations, and enable further advancements in this domain. The protocols used in the literature review, including the configurations and results, will be documented. In addition, the work package will assess and reflect upon the literature review results.

A general overview of our research project can be found in Kneis et al. [1].
3 Configuration

We performed a structured literature review based on the guidelines from Kitchenham et al. [2] to answer our research questions. Based on the guidelines and our project description, the first step is to develop a search configuration. As recommended by Kitchenham et al. [2], we first derived a configuration and then performed a prestudy, where we tested our configuration. Based on the prestudy results, we refined the initial configuration. In the following, we will only report on the final configuration. However, the detailed changes in the configuration and results of the prestudy can be found in our data set [3]. In addition, the data set also contains the detailed final configuration.

3.1 Research Questions

The first step in the configuration is to explicitly define the research questions to be answered by the literature review. We derived the research question based on our initial task description of our research project ANYMOS[1] In the following, we first list the research question and then describe the reason for the question in more detail. Our research questions are:

**RQ1** What are data anonymization methods used in the mobility domain?

**RQ2** What are typical application scenarios for data anonymization methods in the mobility domain?

**RQ3** What type of data is anonymized?

**RQ4** What are the advantages of certain data anonymization methods in the mobility domain?

**RQ5** What are the disadvantages of certain data anonymization methods in the mobility domain?

The first research question **RQ1** investigates anonymization methods in the mobility domain. We concentrate on anonymization methods because our work packages should provide an overview of the used anonymization approaches. The application domain is the mobility domain. As previously stated, mobility refers to public transport and private cars. The mobility focus is also based on our task description because it explicitly names approaches in the mobility domain.

The second research question **RQ2** investigates the scenarios in which data anonymization methods are used. This is important because the different anonymization approaches can differ between different scenarios. This means, we cannot investigate an anonymization method without considering

[1] ANYMOS project page: [https://www.anymos.de/](https://www.anymos.de/)

3
the application scenario for it. Additionally, the same aspect is found in our
work package description.

The third research question RQ3 investigates which data is anonymized.
This question is important because different mobility data might require
different anonymization methods. For example, location data through GPS
position may be anonymized differently than the name of a person. Conse-
quently, we need to identify for which data the anonymization methods can
be used.

The fourth research question RQ4 and the fifth research question RQ5
investigate the advantages and disadvantages of a certain anonymization
method. They are important because they enable us to compare the different
anonymization methods in a certain scenario. The previous questions only
give us insights regarding the potential application areas but do not discuss
how good a method is in these areas. These two research questions are
there to enable this discussion. In addition, these questions are also directly
derived from our initial task description.

3.2 Search Terms, Inclusion & Exclusion Criteria

We decided to perform a database search to identify the relevant literature
for our review. The database search is beneficial, in our case, because we did
not have prior knowledge on anonymization methods in the mobility domain.

For the database search, we need to identify keywords and then build search
terms. These search terms are then used to query a database with the help
of a search engine. We derived the search terms based on our task definition
in our project proposal and the previously presented research questions.

We first identified the important keywords. In our project, we want to
investigate the anonymization in mobility systems. We look at mobility
systems that consist of two aspects, a car related part with automotive and
a public transport part. Based on these aspects, we then identified synonyms
through related work or prior knowledge of the authors. After finding the
keywords and the synonyms, we derived the search terms. In our case, we
split the search terms into two terms. The first search term investigates more
general anonymization methods in mobility systems. The second search term
focuses on tracking entities within a mobility system. Our two search terms
are:

1.\((\text{Anonymization}|\text{Privacy})&(\text{"mobility system"}|\text{automotive}|\text{in-vehicle}|\text{"public transport"}|\text{"transportation system"})\)

2.\((\text{privacy}|\text{anonym}|\text{Pseudo})&(\text{"driver identification"}|\text{trajectory}|\text{"passenger tracking"})&(\text{mobility}|\text{"transportation system"}|\text{automotive}|\text{in-vehicle})\)

The ampersand (&) stands for a conjunction ("and") and the pipe | for a
disjunction ("or") of keywords. In case of a keyword consisting of multiple
words, such as *public transport*, we use quotation marks to indicate the complete keyword.

While our search terms would yield a set of publications, this set might still include irrelevant publications. Therefore, we defined the following inclusion and exclusion criteria, as recommended by Kitchenham *et al.* [2].

- **Inclusion Criteria**
  - any data anonymization method within the mobility domain,
  - methods as foundations for the method kit.

- **Exclusion Criteria**
  - non-English publications,
  - non peer-reviewed publications,
  - sole description of attacks on anonymization/privacy,
  - non-privacy or non-anonymization approaches (e.g. Integrity ...),
  - no full text of the publication available,
  - surveys,
  - challenge papers,
  - vision papers.

Based on our research questions, we derived the criteria as suggested by Kitchenham *et al.* [2]. In addition, we used our task description. Our inclusion criteria selects publications with anonymization or approaches for the method kit. We excluded non-English publications because English results increase the reproducibility, and we might not be able to understand other languages. We also only wanted to include peer-reviewed publications to increase the scientific quality and only to identify approaches with a certain maturity. We also excluded publications that only discuss attacks because they do not provide methods for our kit. In addition, another work package in ANYMOS investigates attacks. We also excluded publications that do not investigate anonymization or privacy because other security properties are, in our case, not relevant. We also excluded publications where we could not access the full text. In addition, we excluded survey, challenge, and vision papers because they do not provide methods, which was the goal for our survey.

### 3.3 Search Engines

In the next step, we selected search engines. Our selection criteria for the search engines were (1) support for our search terms and (2) indexation of commonly used publication venues in computer science. Based on these criteria, we chose the following four search engines:
1. Google Scholar
2. IEEE Xplore
3. ACM Digital Library
4. Semantic Scholar

In addition to the mentioned engines, we also considered the search engines of Springer and Elsevier, respectively. However, they did not comply to our first criterion: We could not enter our search terms since they do not support sophisticated search queries. However, Google Scholar includes the results from Springer and Elsevier.

For each selected search engine, we slightly adapted the search terms syntactically, so that the term corresponded to the required syntax of the search engine. Originally, we planned to apply the search terms to both abstract and title where supported by the search engine. However, due to a high number of results (e.g., 637 publications only for IEEE) and a high number of false positives we found in the results, we decided to apply the search term only to the titles. Afterward, we extracted the found papers, removed duplicates and applied our inclusion and exclusion criteria. We used the title and abstract to decide the conformance to exclusion and inclusion criteria. Where title and abstract were insufficient, we used the conclusion of the respective paper. We performed our search at the start of July 2023, and as such only papers published at this point time could be considered.

3.4 Data Extraction Form

Table 1 illustrates the data we extracted for every valid publication. A publication is valid if it was identified in the previous steps taking into account our inclusion and exclusion criteria. The left column contains our considered data fields, the right column describes possible values. While some data fields are enumerations of predefined values, some fields allow free text and are marked accordingly with free text.

The first four data fields are metadata regarding the reviewer of the paper or the paper itself. In our tooling, the fields Authors, Title and Identifier are automatically extracted based on the BibTeX file containing the found literature. Afterward, we collect the name of the approach. This is important, especially if approaches are repeatedly mentioned in different publications. The field domain relates to our RQ2, and as such provides three possible answers: Public Transport, In-Vehicle or a free text field represented by Other. The next field Data Types is used to answer our

\[\text{https://scholar.google.de/}\]
\[\text{https://ieeexplore.ieee.org/Xplore/home.jsp}\]
\[\text{https://dl.acm.org/}\]
\[\text{https://www.semanticscholar.org/}\]

6
research question [RQ3] The values are defined based on our values found in the prestudy. However, we also allow free text answers in case the predefined classes do not match the approach indicated by Other. Afterward, we have the fields for advantages and disadvantages, based on our research questions [RQ4] and [RQ5] respectively. We extracted these based on advantages and disadvantages of the approach as named in the paper by the authors. The next field is used to describe the validity of the results, aiming at describing how well the approach is evaluated and how reliable the results are. In other words, what evaluation goals, such as accuracy or performance, were targeted. This also includes whether a case study or only a structured discussion or other methods are used. The Data set field describes whether the paper provides a data set or uses a publicly available data set. Yes describes that a data set including the cases, the developed approaches, and evaluation data is available. In the case of Partial, only some but not all of these parts are available. No describes that no part of the data set is available. The next field describes the theoretical concept behind the approach. This is important because it lets us group different approaches. The last field is the Publication Type. It indicates in which venue the paper is published. The options are Journal, Conference, Workshop, and Other. This category is intended to show the maturity of the approach.

3.5 Data Extraction Process

For the data extraction, we used the SLR Toolkit [4] and combined it with a Git-Repository as data storage. We created for each paper an issue in our Git-Repository. The reviewers assigned themselves the issues and then extracted the data from the publications. Due to the high number of publications and the reviewer numbers, each paper was only classified by one reviewer. However, as Kitchenham et al. [2] described for the case of only single reviewers, we performed afterward quality methods to assure consistent data extraction. In our case, we randomly selected results from other reviewers and checked whether the results were correct.

4 Results

We first performed a prestudy for the keywords and the data extraction. We used the prestudy to adjust our search terms and the data extraction form. After the prestudy, we added, for example, the fields regarding the Publication Type to emphasize the maturity of an approach.

After the prestudy, we started the regular survey. Table 2 lists the number of publications found for each search term, and its breakdown into excluded and included publications based on our criteria in section 3. Overall, we found 235 publications for both terms. We excluded 125 publications based on our exclusion criteria and included 110. In detail, for the
Table 1: Data Extraction Form

<table>
<thead>
<tr>
<th>Object</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer’s Name</td>
<td>free text</td>
</tr>
<tr>
<td>Authors</td>
<td>free text</td>
</tr>
<tr>
<td>Title</td>
<td>free text</td>
</tr>
<tr>
<td>Identifier</td>
<td>free text</td>
</tr>
<tr>
<td>Name of Approach</td>
<td>free text</td>
</tr>
<tr>
<td>Domain</td>
<td>”Public Transport”, ”In-Vehicle” or ”Other: free text”</td>
</tr>
<tr>
<td>Data Types</td>
<td>”Text-Based”, ”Numerical”, ”Image”, ”Video” or ”Other: free text”</td>
</tr>
<tr>
<td>Advantages</td>
<td>free text</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>free text</td>
</tr>
<tr>
<td>Validity</td>
<td>free text</td>
</tr>
<tr>
<td>Data set</td>
<td>”Yes”, ”Partial” or ”No”</td>
</tr>
<tr>
<td>Theoretical Concept</td>
<td>free text</td>
</tr>
<tr>
<td>Publication Type</td>
<td>”Journal”, ”Conference”, ”Workshop” or ”Other: free text”</td>
</tr>
</tbody>
</table>
first query, we got 217 publications. There, we excluded 114 publications. The most common exclusion criteria are duplicates, surveys, or non-peer-reviewed publications. However, some of the non-peer-reviewed publications are arXiv preprints and, therefore, a duplicate for the later published version. This results in 103 included publications for the first search term.

For the second search term, we found 18 publications and excluded 11. This leaves us with seven included papers. Here the main exclusion criteria were papers only describing attacks, duplicates, or non-peer-reviewed articles. A complete list with all found and excluded and included publications can be found in our data set [3].

In the following, we describe our findings. We start by answering RQ1. We answer this research question by giving an overview of the theoretical concepts field (c.f. section 3) from our data extractions. These methods used are always dependent on a specific application or data type. However, by enumerating the approaches first, we provide a first impression of the variety of methods to choose from. Many approaches use some form of differential privacy [5] such as Qiu et al. [6] or Zhao et al. [7]. Overall 15 publications use a notion of differential privacy to anonymize data. Other approaches use $k$-anonymity [8] such as Stegelmann et al. [9] or Damjanovic-Behrendt [10]. Overall nine approaches use $k$-anonymity. The two concepts are often used for anonymization of trajectory data sets.

Another popular approach, especially in combination with exchanging data, is blockchain-based approaches such as Yang et al. [11] or Li et al. [12]. These methods are often used in the communication with road side units or exchanging status information with a backend. Other approaches make use of selective disclosure, meaning the selection of the transferred data by the user, for instance, through a firewall as in Klement et al. [13] or access control as in Plappert et al. [14]. They are often combined with access to bus systems, such as OBD. Also, random pseudonyms are created to hide the identity as in Wan et al. [15] or random noise is used to conceal a location [16]. For machine learning approaches, some methods use generated data to learn, such as Abdelwahed et al. [17]. Others use federated learning such as Ruan et al. [18] or combine it with homomorphic encryption such as Wang et al. [19] or Han et al. [20]. Some approaches use multi-party

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Found</th>
<th>Excluded</th>
<th>Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>217</td>
<td>114</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td><strong>∑</strong></td>
<td>235</td>
<td>125</td>
<td>110</td>
</tr>
</tbody>
</table>

Table 2: Overview of selected and rejected papers
computing (MPC) to operate within an anonymous system, such as Wang et al. [19] and Ying et al. [21]. A complete list of concepts used is described in our data set [3].

Next, we present the results for the application domains. This answers our research question RQ2. Figure 1 illustrates the different application domains. It is important to note that multiple selections are possible as an answer. The majority of publications are in the In-Vehicle domain, with 83 publications focusing on it. Other domains occur less frequently. The second-largest group are Public Transport and Other with each 18 publications. Only nine publications cover Pedestrian. Looking into more detail in the other application domains, we also see many car-specific applications, such as Singh et al. [22] or Fan et al. [23] with VANET. Other domains are transportation systems in general or smart grids together with charging protocols such as in Agilandeeswari et al. [24]. However, these are also more car-related topics. Hence, we can say that many publications are focused on car-related topics. One possible reason might be that also car manufacturers invest in research in this domain, for instance, publications such as Plappert et al. [14], Li et al. [25] or Asaj et al. [26] are co-authored by researchers from car manufacturers.

![Figure 1: Application Domains](image)

The next research question RQ3 is answered by the data type field in our extraction form. Figure 2 illustrates the data types found. Again, multiple
selections are possible for each publication. Overall, many publications could not be mapped to our predefined data types. This is the case because the approaches are often very specific or consist of complex data types. Examples of very specific anonymization data types are, for instance, the heart rate in Ruan et al. [18] or the identifiers of a car as in Hong et al. [27]. In these very specific cases, we selected the more general data type Numerical and additionally added the more specific information in the Other text field. Examples of complex data are Ticket-Punch-Data [28] or electric car data for the smart grid [29]. If we look more closely into the Other field, we can identify many approaches that focus on the anonymization of car or driver identifiers, such as Salem et al. [30] or Ahmed et al. [31]. Overall these are 13 publications. Two other larger groups in the Other field are location data or trajectory data, such as in Fan et al. [23] or Ruan et al. [18]. Overall 24 approaches cover trajectory or location data.

The following two research questions, RQ4 and RQ5, discuss the advantages and disadvantages. In this regard, we only extracted the advantages and disadvantages mentioned in the publications by the respective authors. In the next section (c.f. section 5), we will interpret these results and give a comparison. Some of the advantages of the approaches include confidential pseudonyms for authentication, unlinkability, resistance against common attacks, elimination of vehicle IDs, secure key exchange, location and usage privacy [6], [11], [17], [24], [32], support of billing, minimizing changes to ex-
isting standards [33], hiding trajectory movements [34], anonymization of location and passenger data [28, 36, 37], secure communication [38–40], protection against traffic flow analysis [41], controlled data sharing [10, 42], privacy-preserving authentication and communication [43, 44], task assignment based on privacy-preserving geo data [45], protection of vehicle location [46, 47], secure and private data exchange [48], cryptographic guarantees [7, 9, 25, 49–59].

Some of the disadvantages of the mentioned privacy methods in the mobility domain include communication overhead [11], no consideration of other attackers [6], high resource demands [17, 34], focus only on exchange of vehicle ID [24, 32, 24], no general anonymization of data [33].

The next fields of our extraction form are later used in our comparison (c.f. section 5) to differentiate between different approaches. We start by presenting the results for the validity. This is established through a variety of evaluation techniques for the presented anonymization methods. For instance, the approach by Yang et al. [11] is validated primarily through performance evaluation. Qiu et al. [6] conduct theoretical analysis on privacy properties and complexity, complemented by experimental evaluations on data sets. Abdelwahed et al. [17] validate their approach via testing on custom videos. Agilandeeswari et al. [24] combine simulation for identifying leaky data, performance evaluation, and energy assessment. While some approaches provide structured discussions or theoretical proofs, others rely on formal analysis, such as Ruan et al. [18] and Singh et al. [22]. Overall, validity is a difficult field because of the wide variety of used methods. However, many approaches do not prove that they actually satisfy the intended security property. They either only present their solution, e.g., Abuman-soor et al. [32], Duri et al. [50] or only discuss why their approach should provide these aspects, such as ShangGuo [60] or Rabadi et al. [61]. The most frequent evaluation method is a performance-based analysis (around 40 papers). Here, it is preferred to measure the runtime of a feature in the approach.

However, reproducing these results is often very hard because most approaches do not have a publicly available data set. Figure 3 illustrates this aspect. It shows the number of publications with a full data set, only a partial one or no data set at all. Only two publications [19, 62] had a full data set published. This includes all aspects, like the source code and data used. 22 publications have a partial data set published. This was often the case when the authors referenced a case study or data set used for the evaluation. For instance, Löbner et al. [63] used a data set containing data about the trajectory and energy consumption. Other approaches like Liu et al. [34] use a generator [65] to create trajectory data. However, not only trajectory data is used, for instance Ruan et al. [18] uses the heart rate
from the HARMONY and PPG-DaLiA data set. These data sets also contain other additional data, such as motion data of the device measuring the heart rate. Nevertheless, the publications marked with Partial often did not publish the modification on the data sets, concrete selection of the data, meaning which data fields are considered, or the source code. Most publications have no published data set. Despite that, many approaches claim to use a data set in their validation.

![Figure 3: Data sets included](https://osf.io/zextd/)

The last field that we present is the venue type of the publication. The venue type should show the maturity of the approach. Figure 4 illustrates the distribution of venue types. Most publications are published at conferences. These are 50 publications. The next big group are journal publications, with 45 ones. Only 11 publications are workshop publications. Two publications are neither of the previous ones. In our case, these were only poster submissions. Based on our findings, we can see that the research field is already established and contains mature approaches because many publications are found in journals or conferences.

### 5 Discussion of the Results

Based on our previously presented results, we discuss our different findings. An overall comparison of all the different methods is difficult to accomplish,
because of the various application domains and approaches. Also, the lack of the published data or evaluation makes it harder. In addition, many approaches do not discuss their disadvantages or only in a very limited setting, such as Singh [66]. These aspects are especially relevant for future work, which should build up on identified approaches.

However, with these work packages in mind, we can identify relationships to other work packages (WP). For instance, Righini et al. [67] or Rupp et al. [68] describe a novel ticketing system for public transportation. In our WP 1.3.4, we look at a similar case. Hence, this approach might be a starting point for further discussions. Other relevant publications are, for instance, Abdelwahed et al. [17] for WP 1.3.5 as generating anonymous video training data or Sciancalepore et al. [69] for privacy-aware collision avoidance in autonomous vehicles, which is related to our WP 1.2 about autonomous cars.

In addition, there is a relation to our WP 5.2, which investigates machine learning and anonymity. Here, especially relevant are the following approaches. Abdelwahed et al. [17] is a privacy-aware passenger observing approach. Other interesting methods are using federated learning or combining it with homomorphic encryption [18–20, 70].

Another relation can be found with the WP 6, which investigates de-anonymization approaches. Here, our publications can help to identify attacks because some approaches such as Da Silva et al. [71], Joy et al. [72] or Lu et al. [16] discuss in addition to their anonymization also attacks on
anonymized data.

Besides relating the identified publications to our other WPs in ANYMOS, we want to give an overview of different prerequisites for applying these approaches. Many approaches do not anonymize the data but rather encrypt them, such as Malina et al. [73], Salem et al. [39], or Mankar et al. [74]. This then requires a trusted entity to manage the data. Overall these approaches tend to use simpler encryption algorithms. Hence, they usually are more efficient and can be used on low-powered hardware. Other approaches like Liu et al. [34] use far more powerful hardware. In this case, it is an Intel Core i7. However, the approach also gives far more security guarantees about the reidentification than more straightforward approaches.

Regarding future research directions, we see a wide gap in the research between public transport and car-related research. Many approaches focus on car-related aspects or use their methods in car-related topics. In the future, it might be interesting to bring these two aspects closer together and close the gap between the different research areas. For instance, trajectory anonymization such as Liu et al. [34] can also be used to anonymize public transport data or person data. Also, our research project aims to unite both perspectives in our research.

Furthermore, it would be beneficial if the different approaches would provide data sets. This could help other researchers to compare different methods better. However, this is an ongoing problem, which also affects other areas like software engineering [75].

6 Limitations of the Review

In this section, we reflect on the limitations of our literature review. We start by discussing the limitations based on our decision about the configuration.

The first limitation is about the chosen search engines. As previously stated, we performed a database search. In this regard, the used databases are essential to identify the relevant publications. We can only find the publications if the database contains the relevant publication. In our case, we chose the most common and relevant databases in our field. In addition, we used Google Scholar, which indexes a wide variety of different publishers. Hence, we assume the limitation to be small.

Another point limiting our review are the used keywords. As discussed in the initial configuration (c.f. section 3), the keywords are essential for identifying relevant publications. We first created a list of keywords to avoid mistakes and discussed these in project workshops. Afterwards, we performed a prestudy to refine the keyword selection. Then, we again discussed the keywords in our project, and only then finally performed the actual literature review. This provides confidence to have identified the relevant keywords for our research questions. We note, however, that our
review focuses on anonymization approaches within the mobility domain. We did not investigate anonymization methods in general and how these could be applied in the mobility domain. Therefore, the review can only make statements about approaches already used in the mobility domain, and it is possible that additional approaches exist that could be applied in the mobility domain, but either have not been tried or the results published.

Next, having the extraction process performed by only one reviewer could limit the review quality. To reduce this threat, we performed a quality assurance step after the extraction. With this step, we unified the extracted data and verified the results. Our data set is publicly available, and as such, further checks could be performed by third parties.

Finally, while the survey lists different approaches, our results are not suitable for deciding for a single dominant anonymization strategy. As previously mentioned, the application area for anonymization in the mobility domain is vast, so up to now, and so far, no single approach appears to be able to solve all the problems. Our review provides a starting point for identifying necessary approaches, and could give rise to the development of new or the application of existing techniques.

7 An Industry-driven Survey on Status Quo and Demand

This section focuses on the anonymization techniques used specifically for video data in public transport. It is important to note that the discussion is limited to video data, excluding for example passenger counting data. The scope encompasses video data viewed outside of the vehicle, both in live and non-live scenarios.

The use cases discussed here are provided by DResearch Fahrzeugelektronik GmbH (DFE). DFE is a development-oriented company that offers CCTV solutions for public transport globally and has over 800 customers in "D-A-CH" region alone. Since its establishment in 1994, DResearch Media GmbH, along with its subsidiary company DFE (established in 2011), has accumulated extensive expertise in developing maintenance-free embedded systems for long-term service. With more than 40,000 active systems, DFE has solidified its position as market leader.

The insights and results presented in this section are derived from an informal in-house survey conducted among DFE project engineers. The survey aimed to gather information on the anonymization requirements, needs and usage of DFE customers in the public transport sector.

Since the focus increases on data privacy through regulations such as the General Data Protection Regulation (GDPR), the relevance of anonymization techniques, specifically used on video data will play a major role in future surveillance systems. The GDPR emphasizes the importance of pro-
tecting individuals’ personal data, including video data, and requires organizations to implement measures to ensure data privacy.

The survey revealed various use cases for video data in public transport, including an estimation on whether anonymization is required according to GDPR:

1. Surveillance: Video data plays a crucial role in surveillance systems for enhancing security measures. Law enforcement agencies rely on clear video data to identify and investigate security incidents, such as theft, vandalism or unauthorized access. In these cases, anonymization is not applicable, as the clarity and quality of the footage are essential for effective analysis.

2. Quality Management: Video data is utilized for quality management purposes, such as investigating passenger complaints, monitoring service quality or assessing operational efficiency. Under data protection regulations like the General Data Protection Regulation (GDPR), there is a need to anonymize video data used in quality management processes to protect passenger privacy.

3. External Verification of People Counting Systems: Video data is employed to verify and validate people counting systems deployed in public transport. These systems help gather information on passenger flow, capacity utilization, and planning. To ensure privacy compliance, video data used for external verification purposes also requires anonymization.

4. Ensuring Functionality of the Video System: Remote live-streaming of video data is used to assess the functionality and performance of the video systems installed in public transport vehicles. Anonymization techniques are necessary in this case to safeguard passenger privacy while allowing for real-time monitoring and evaluation.

5. Creation and Extension of AI Training Data: Video data can be utilized to create and extend training datasets for artificial intelligence (AI) algorithms. Anonymization is crucial when using video data to develop AI models to ensure that personally identifiable information is removed or obscured, protecting the privacy of individuals captured in the footage.

Furthermore, the survey gave insights into the anonymization techniques currently used today, which are limited to pixelization of static areas. This technique is used for example, to pixelate video data showing the driver’s seat in the vehicle, since German law prohibits the usage of surveillance at a workplace. However, to ensure ongoing GDPR compliance while maintaining usability, the implementation of more advanced anonymization techniques is necessary. For instance, incorporating dynamic object detection
and anonymization could prove to be beneficial. An area where this technique could be applied is real-time head detection and anonymization, which would effectively safeguard passenger privacy.

In conclusion, the current implementation of anonymization techniques for video data in public transport remains limited. However, with the increasing focus on data privacy and the enforcement of regulations such as the General Data Protection Regulation (GDPR), the relevance of anonymization techniques is expected to play a vital role in the future. Anonymization techniques will become crucial in complying with these regulations by removing personally identifiable information and preventing the risk of re-identification. By adopting effective anonymization methods, such as dynamic object detection and anonymization, public transport providers can strike a balance between privacy protection and data utility. This will enable them to continue benefiting from video data while safeguarding passenger privacy in accordance with GDPR rules.

8 Conclusion

This report presented our systematic literature review for the method building kit of the ANYMOS project. We presented our configuration with search terms and inclusion and exclusion criteria. We presented and discussed the results. In addition, we put the results in relation to other work packages in the research project. We also gave indicators on how to compare the different approaches and identified future research directions.

Besides the literature review, we also presented the result of an industry-driven survey by DFE regarding anonymization techniques used in public transport. This survey focuses on video technologies and was conducted within a company developing video surveillance technologies. It shows the current status quo and stresses the importance of further research in this area.

In the future, we plan to use our results in other working packages of the ANYMOS project.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) grant number 16KISA086 (ANYMOS) and the NextGenerationEU project by the European Union (EU).

References


[32] O. Abumansoor, A. Boukerche, B. Landfeldt, et al., “Privacy pre-
serving neighborhood awareness in vehicular ad hoc networks,” in Pro-
ceedings of the 7th ACM Symposium on QoS and Security for
Wireless and Mobile Networks, ser. Q2SWinet ’11, event-place: Mi-
ami, Florida, USA, New York, NY, USA: Association for Computing
Machinery, 2011, pp. 17–20, isbn: 978-1-4503-0899-1. DOI: 10.1145/
2069105.2069109.

[33] D. Angermeier, A. Kiening, and F. Stumpf, “PAL - privacy aug-
mented LTE: A privacy-preserving scheme for vehicular LTE commu-
nication,” in Proceeding of the Tenth ACM International Workshop
on Vehicular Inter-Networking, Systems, and Applications, ser. VA-
NET ’13, event-place: Taipei, Taiwan, New York, NY, USA: Associa-
tion for Computing Machinery, 2013, pp. 1–10, isbn: 978-1-4503-
//doi.org/10.1145/2482967.2482975.

[34] X. Liu and Y. Zhu, “Privacy and utility preserving trajectory data
publishing for intelligent transportation systems,” IEEE Access, vol. 8,
pp. 176 454–176 466, 2020, issn: 2169-3536. DOI: 10.1109/ACCESS.
2020.3027299.

privacy preservation for intelligent vehicle network in mobile IoT,”
IEEE Access, vol. 8, pp. 81 698–81 710, 2020. DOI: 10.1109/ACCESS.
2020.2990932.

[36] Y. Y. Liu, A. Cooke, and P. Thulasiraman, “Location privacy preser-
vation of vehicle data in internet of vehicles,” [Online]. Available:

[37] I. Gudymenko, “A privacy-preserving e-ticketing system for public
transportation supporting fine-granular billing and local validation,” in
Proceedings of the 7th International Conference on Security of

for a crowd-sourced semantic accessibility database for public trans-
port,” in HCI International 2020–Late Breaking Papers: Universal
Access and Inclusive Design: 22nd HCI International Conference,
HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings
springer.com/chapter/10.1007/978-3-030-60149-2_54.


Appendix

Included Publications


F. Callegati, A. Campi, M. Prandini, et al., “PRIVACY ISSUES IN a CLEARING SYSTEM FOR a REGIONAL-SCALE PUBLIC TRANSPORT NETWORK,” in *Proceedings of the 5th INTERNATIONAL CONFERENCE ON INTERNET TECHNOLOGIES & SOCIETY*, IADIS Press, 2014,


Y. Li, K. Ouyang, N. Li, et al., “A blockchain-assisted intelligent transportation system promoting data services with privacy protection,” *Sensors*, vol. 20, no. 9, p. 2483, 2020, Publisher: MDPI.


