KIT SCIENTIFIC REPORTS 7692

Proceedings of the First Karlsruhe Service Summit Research Workshop - Advances in Service Research

Karlsruhe, Germany, February 2015

Roland Görlitz, Valentin Bertsch, Simon Caton, Niels Feldmann, Patrick Jochem, Maria Maleshkova, Melanie Reuter-Oppermann (eds.)

KIT Scientific Publishing

Roland Görlitz, Valentin Bertsch, Simon Caton,
Niels Feldmann, Patrick Jochem, Maria Maleshkova,
Melanie Reuter-Oppermann (eds.)

Proceedings of the First Karlsruhe Service Summit Research Workshop - Advances in Service Research

Karlsruhe, Germany, February 2015

Karlsruhe Institute of Technology
KIT SCIENTIFIC REPORTS 7692

Advances in Service Research

Series Editors

Wolf Fichtner
Karlsruhe Institute of Technology (KIT)

Kai Furmans
Karlsruhe Institute of Technology (KIT)

Stefan Nickel
Karlsruhe Institute of Technology (KIT)

Ralf Reussner
Karlsruhe Institute of Technology (KIT)

Gerhard Satzger
Karlsruhe Institute of Technology (KIT) and IBM

Rudi Studer
Karlsruhe Institute of Technology (KIT)

Christof Weinhardt
Karlsruhe Institute of Technology (KIT)

Helmut Wlcek
Karlsruhe Institute of Technology (KIT) and Bosch

Proceedings of the First Karlsruhe Service Summit Research Workshop - Advances in Service Research

Karlsruhe, Germany, February 2015

by
Roland Görlitz
Valentin Bertsch
Simon Caton
Niels Feldmann
Patrick Jochem
Maria Maleshkova
Melanie Reuter-Oppermann (eds.)

KIT Scientific Publishing

Report-Nr. KIT-SR 7692

Volume Editors

Valentin Bertsch
Senior Researcher at Karlsruhe Institute of Technology
Valentin.Bertsch@kit.edu

Simon Caton
Lecturer at the National College of Ireland
Simon.Caton@ncirl.ie

Niels Feldmann
Senior Researcher at Karlsruhe Institute of Technology
Niels.Feldmann@kit.edu

Roland Görlitz
General Manager Karlsruhe Service Research Institute
Roland.Goerlitz@kit.edu

Patrick Jochem
Senior Researcher at Karlsruhe Institute of Technology
Patrick.Jochem@kit.edu

Maria Maleshkova
Senior Researcher at Karlsruhe Institute of Technology
Maria.Maleshkova@kit.edu

Melanie Reuter-Oppermann
Senior Researcher at Karlsruhe Institute of Technology
Melanie.Reuter@kit.edu

Impressum

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

KIT Scientific Publishing is a registered trademark of Karlsruhe Institute of Technology. Reprint using the book cover is not allowed.

www.ksp.kit.edu

This document – excluding the cover – is licensed under the Creative Commons Attribution-Share Alike 3.0 DE License (CC BY-SA 3.0 DE): http://creativecommons.org/licenses/by-sa/3.0/de/

The cover page is licensed under the Creative Commons Attribution-No Derivatives 3.0 DE License (CC BY-ND 3.0 DE): http://creativecommons.org/licenses/by-nd/3.0/de/

Print on Demand 2015

ISSN 1869-9669
ISBN 978-3-7315-0344-6
DOI: 10.5445/KSP/1000045634

Conference Organization

Track and Session Chairs

Energy and Mobility Services
Valentin Bertsch
Patrick Jochem
Christof Weinhardt

Healthcare Services, Logistics and Information Systems
Melanie Reuter-Oppermann
Stefan Nickel

Social Collaboration and Service Innovation
Simon Caton
Niels Feldman
Gerhard Satzger

Semantic Web Technologies and Web Services
Maria Maleshkova
Rudi Studer

Program Committee

Alexander Mädche	University of Mannheim, Germany
Anne Zander	Karlsruhe Institute of Technology, Germany
Barbara Carminati	University of Insubria, Italy
Bastian Chlond	Karlsruhe Institute of Technology, Germany
Christian Wernz	Virginia Tech, USA
Christof Weinhardt	Karlsruhe Institute of Technology, Germany
Eric Demeulemeester	KU Leuven, Belgium
Erwin Hans	University of Twente, Netherlands
Evrim Gunes	KOC University, Turkey
Felix Leif Keppmann	Karlsruhe Institute of Technology, Germany
Georg Groh	TU München, Germany
Gerhard Satzger	Karlsruhe Institute of Technology, Germany
Henner Gimpel	University of Augsburg, Germany
Heidi Heinrichs	Forschungszentrum Jülich, Germany
Herbert Kotzab	Uni Bremen, Germany
Honora Smith	University of Southampton, United Kingdom
Ines Arnolds	Karlsruhe Institute of Technology, Germany
Ingela Tietze	Hochschule Niederrhein, Germany
Jacek Kopecky	University of Portsmouth, United Kingdom
Jakob Wachsmuth	Smart Grids Plattform Baden-Württemberg e.V., Germany
Jan Treur	University of Amsterdam, Netherlands

Jan-Marco Leimeister	University of Kassel, Germany
Jens Hogreve	Catholic University of Eichstätt, Germany
Jens Schippl	Karlsruhe Institute of Technology, Germany
Kai Furmans	Karlsruhe Institute of Technology, Germany
Kai Hufendiek	University of Stuttgart, Germany
Kathrin M. Möslein	University of Erlangen, Germany
Kyle Chard	University of Chicago, USA
Lars-Peter Lauven	University of Göttingen, Germany
Margarete Hall	Karlsruhe Institute of Technology, Germany
Maria Esther Vidal	University of Bolivar, Venezuela
Maria Maleshkova	Karlsruhe Institute of Technology, Germany
Markus Frank	Energy Solution Center e.V., Germany
Markus Krause	University of Hannover, Germany
Martin Junghans	Karlsruhe Institute of Technology, Germany
Melanie Reuter-Oppermann	Karlsruhe Institute of Technology, Germany
Michael ten Hompel	IML Fraunhofer, Germany
Mike W. Carter	University of Toronto, Canada
Monica Oliveira	University of Lisboa, Portugal
Nancy Wünderlich	University of Paderborn, Germany
Niels Feldmann	Karlsruhe Institute of Technology, Germany
Patrick Jochem	Karlsruhe Institute of Technology, Germany
Patrick Plötz	ISI Fraunhofer, Germany
Paul Fremantle	University of Portsmouth, United Kingdom
Peter Hottum	Karlsruhe Institute of Technology, Germany
Roberta Cuel	University of Trento, Italy
Roberto Aringhieri	University of Torino, Italy
Roland Görlitz	Karlsruhe Institute of Technology, Germany
Rudi Studer	Karlsruhe Institute of Technology, Germany
Sally Brailsford	University of Southampton, United Kingdom
Simon Caton	National College of Ireland, Ireland
Stefan Nickel	Karlsruhe Institute of Technology, Germany
Steffen Stadtmüller	Karlsruhe Institute of Technology, Germany
Stephen Kwan	San Jose State University, USA
Tilo Böhmann	University of Hamburg, Germany
Valentin Bertsch	Karlsruhe Institute of Technology, Germany
Vince Knight	Cardiff University, United Kingdom
Xiaolan Xie	Ecole Nationale Superierue Saint-Etienne, France

Preface

Since April 2008 KSRI fosters interdisciplinary research in order to support and advance the progress in the service domain. As an industry-on-campus partnership between KIT, IBM and Bosch, KSRI brings together academia and industry while serving as a European research hub with respect to service science. In the past years we have established a vivid speaker series and an enriching exchange with guest professors from leading global universities and practitioners with responsibilities for the services business of their respective companies. Our regular Service Summits and Summer Schools became a well-known place for collaboration and source of inspiration for our partners from the service science and service business communities. This year we decided to co-locate our Fifth Service Summit with a service research workshop. For KSRI's First Karlsruhe Service Summit Research Workshop, we invited submissions of theoretical and empirical research dealing with the relevant topics in the context of services including energy, mobility, health care, social collaboration, and web technologies. With the help of the reviewers, we selected valuable submissions and grouped them in the following four tracks.

1) Energy and Mobility Services: The energy sector continues to undergo substantial structural changes. The expanding usage of renewable energy sources, the decentralization of energy supply and the market penetration of electric vehicles will have a significant impact on the future development of energy service requirements. In the transport sector, especially in road transport and intralogistics, the understanding of energy consumption and the efficient usage of energy in material handling technology is of increasing importance. In order to actively integrate consumers into the energy system of the future, appropriate incentives (e.g. electricity tariffs), market schemes, and service level concepts need to be developed and introduced. This requires, among others, new products in electricity retail markets, innovative marketing and comprehensive acceptance research and the investigation of future business models. Furthermore, efficient usage of energy, efficiency in costs and flexibility in mobility and reconfigurability need to be combined to create successful future mobility services, e.g. in logistics. This track enhances the understanding of the future role of services in energy economics, e-mobility and logistics.

2) Healthcare Services, Logistics and Information Systems: Demographic changes cause higher patient demands alongside severe cost pressure and increasing quality requirements. Therefore, more efficient healthcare services and logistics are desirable. Even though underlying planning problems in the area of Operations Research resemble the ones from other service or manufacturing industries (e.g., scheduling of different tasks, processes or appointments) healthcare services are especially challenging, because patients need different care than, for example, parts of cars. In addition, particularly interdisciplinary approaches are necessary for research on and improvement of healthcare services. Since Information Systems have high potentials for improving efficiency, they play an important role and are of high interest in this track.

3) Social Collaboration and Service Innovation: Collaborative approaches are playing an increasingly important role for individuals as well as corporations in tackling innovation endeavors. Today a variety of platforms that

facilitate the multi-faceted innovation management process exist. They employ approaches contributing to the generation, conceptualization, evaluation, funding, and implementation of ideas and knowledge artefacts. However, there is a lack of interdisciplinary and mixed method approaches for disentangling and understanding the social, cognitive and collaborative processes that underpin these platforms. This track aims to juxtapose the social collaboration and service innovation communities with the intent to shed light on the foundational crowd aspects of social collaboration in service innovation and vice versa.

4) Semantic Web Technologies and Web Services: Traditional Web services based on WSDL and SOAP have dominated the world of services on the Web for a long time. Inspired by the Semantic Web community and aiming to provide a higher level of task automation through semantics, researchers focused on bridging the gap between services and semantics. However, with the proliferation of Web APIs the solutions devised in the context of traditional Web services are unsuitable for addressing the challenges faced by resource-oriented APIs, commonly called REST or RESTful services. Still, up to date, a number of challenges related to using Web services, and especially the support for APIs as well as the facilitating of their integration with data on the Web, remain unaddressed. Contributions in this track deal with some of these unaddressed topics.

Table of Contents

Energy and Mobility Services

Healthcare Services and Logistics

Social Collaboration and Innovation

Semantic Web Technologies and Web Services

A concept for service level indicators in residential electricity tariffs with variable capacity prices

Marian Hayn, marian.hayn@kit.edu, Chair of Energy Economics, Institute for Industrial Production (IIP), Karlsruhe Institute of Technology (KIT)

Valentin Bertsch, valentin.bertsch@kit.edu, Chair of Energy Economics, IIP, KIT

Wolf Fichtner, wolf.fichtner@kit.edu, Chair of Energy Economics, IIP, KIT

The increasing share of renewable energy sources in European electricity markets raises the interest in leveraging demand flexibility also on a residential level. While in traditional residential electricity tariffs, the provided service consists mainly in the delivery of electricity in the right amount and at the right time, more recent tariff structures, especially innovative ones with variable capacity prices, are more sophisticated. In the context of these new tariffs, service level agreements offer a good opportunity to ensure that providers and consumers of that service have a common understanding of its quality. Therefore, a set of four suitable service level indicators for tariffs with variable capacity prices is developed in this paper. Furthermore, possible approaches to derive related service level objectives and possibilities for providing decision support to customers in choosing a tariff are introduced.

1 Introduction

The residential sector in Europe is responsible for about 29% of Europe's total electricity demand (Eurostat European Commission, 2013). In the context of an increasing share of electricity from renewable energy sources (RES) in Europe's electricity markets, leveraging demand flexibility gains importance in order to keep a certain level of supply security (cf. Hayn et al., 2014a). Therefore, new electricity tariffs for residential customers, e.g., tariffs with variable electricity or capacity prices[1], become more prominent in research and political discussions.

In traditional residential electricity tariffs the main service provided to the customer consists of the delivery of the requested amount of electricity at the right time at a pre-defined price. More recent electricity tariffs offer further differentiations, e.g., regarding the used generation mix or with varying electricity prices. According to Parasuraman et al. (1985) services are characterized through their intangibility, heterogeneity and inseparability.

[1] Tariffs with variable capacity prices are characterized through a threshold for the contracted guaranteed capacity available for use at all times (see section 3.3).

As electricity tariffs, especially the more recent ones, fulfill all three characteristics, they can be defined as a service:

- Intangibility: Even though it is possible to measure how much electricity is used, electricity is not tangible for most consumers. When using electricity, consumers do not intend to consume electricity itself but the related service offered through the electricity consumption, e.g., cooking dinner or watching TV (Wilson & Dowlatabadi, 2007).
- Heterogeneity: While electricity itself is, from a consumers' point of view, a homogeneous good[2], the related service of its delivery, is heterogeneous. Consumers can choose from a wide range of different electricity tariffs from different providers. The heterogeneity exists on various levels of the service relation. The contracted service may differ, for instance, with regard to the accessibility and quality of the customer service, the price structure, or the used generation mix for providing the delivered electricity, to name just a few. Although differences in today's contracts are rather marginal, it is likely that changes in the energy system will increase the heterogeneity in future (cf. section 4).
- Inseparability: Production and consumption of electricity are, too a very large extent, inseparable as electricity storage is nowadays still limited, both from a technical and an economic point of view. In current energy markets electricity production and consumption must always be balanced. Additionally, the contracted service of delivering electricity can only be fulfilled satisfactorily when both involved parties, the provider and the consumer, collaborate. The service provider needs to know the consumers' preferences in order to provide the right amount of electricity at the right time from the right energy sources; the consumer needs to demand electricity. While nowadays service providers have almost no knowledge about the consumer's individual preferences, this is likely to change in the future, e.g., through the utilization of smart meters.

As shown, residential electricity tariffs represent a service contract between a service provider, e.g., a utility company, and a customer, i.e., households, which specify the conditions of the delivery of electricity. While nowadays these conditions are mostly limited to the specification of different price components, e.g., fix and variable price components, new electricity tariffs, especially tariffs with variable capacity prices, require the definition of service level agreements (SLAs) between providers and customers. Based on the general concept of SLAs, this article develops a concept of suitable service level indicators (SLIs) for residential electricity tariffs with variable capacity prices. Therefore, we shortly present the relevant basics of SLAs, SLIs and related service level objectives (SLOs) in section 2, followed by a short overview on different existing and new residential electricity tariffs in section 3. Afterwards, in section 4, we discuss a new concept for SLIs and introduce approaches to derive adequate SLOs of tariffs with variable capacity prices. The paper ends with a conclusion and an outlook on future work in section 5.

[2] Once the electricity is available in the grid, consumers cannot differentiate where it comes from and it has a completely uniform quality when used.

2 Service level agreements

Several different definitions for the term "service level agreement" exist in literature, see Berger (2007) for an overview. We follow the definition of Berger (2007), who states that a SLA describes specific target values for commonly defined performance indicators related to the contracted service. In this paper we focus on these performance indicators while other characteristics of SLAs, also described in Berger (2007), are out of scope. Performance indicators and their target values are also called "service level indicators" (SLIs) and their respective "service level objectives" (SLOs) (cf. Kieninger et al., 2011). Below, we will shortly introduce these two constructs.

2.1 Service level indicators

SLIs are used in service contracts to establish well defined performance indicators which enable both the provider and the consumer of services to evaluate the quality of the specific service (cf. Kieninger et al., 2011). Parasuraman et al. (1985) identify the following ten determinants of perceived service quality: reliability, responsiveness, competence, access, courtesy, communication, credibility, security, understanding/ knowing the customer and tangibles. While these determinants are strongly dependent on the individual perception of the service by consumers and can be evaluated through surveys for instance (cf. Parasuraman et al., 1988), SLAs have the objective to define quantifiable service related criteria allowing both the provider and the consumer of a service to evaluate the quality of a service (Berger, 2007). Therefore, Berger (2007) highlights six requirements for a meaningful definition of SLIs used to evaluate services – SLIs must be:

- Fully defined
- Relevant with regard to the value of the service
- Proportionally related to the described circumstances
- Meaningful for the customer
- Fully controllable by the service provider
- Economic

For a more detailed explanation of these requirements, please refer to the stated reference (Berger, 2007). These requirements are used for the development of adequate SLIs for new residential electricity tariffs in section 4.

2.2 Service level objectives

When appropriate SLIs have been chosen for a service, their specific target value still needs to be defined. This target value is referred to as "service level objective" (Kieninger et al., 2011). Kieninger et al. (2011) propose an approach called "service level engineering" (SLE) to analytically determine the cost-efficient value for SLOs. They suggest deriving the value for a SLI which minimizes both variable service costs, from a provider's point

of view, and business opportunity costs, from a customer's point of view. This value represents the Pareto efficient value of the SLI and should be set as the SLO (cf. Kieninger et al., 2011).

3 Residential electricity tariffs

In Europe, a huge number of different electricity tariffs exists for the residential sector. A good overview on the majority of publicly discussed or available tariffs is given in Ecofys et al. (2009). In this paper, we focus on a brief introduction of selected tariffs. Figure 1 shows the principal price components of residential electricity tariffs and a more detailed breakdown based on German tariffs. While tariffs with variable capacity prices (section 3.3) would constitute a new element of the base price, tariffs with variable electricity prices (section 3.2) only alter the energy price in a tariff.

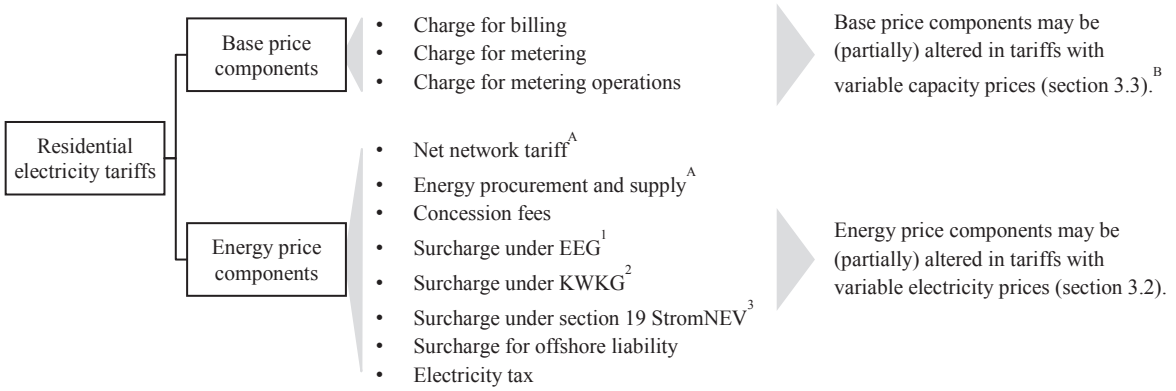


Figure 1: Price components of residential electricity tariffs in Germany (based on (Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen (BNetzA) & Bundeskartellamt, 2014))

3.1 Tariffs with flat electricity prices

The tariff with the largest penetration rate in European households is a tariff with flat electricity prices. This means, that each electricity unit (in kWh) can be consumed for the same price (in €/kWh) independent from any other restrictions. The maximum amount of consumable electricity in each time step (capacity in kW) is only limited by the technical specifications of the main fuse (over current protection). In Germany, for instance, single family houses must be equipped with main fuses with 63 A per phase allowing a maximum load of about 45 kW (DIN Deutsches Institut für Normung e.V., 2013). In other countries, e.g., France or Spain, the allowed capacity might be limited to a contracted threshold (cf. Électricité de France, 2014; Iberdrola, 2014).

3.2 Tariffs with variable electricity prices

Besides those rather simple tariffs described above, more complex tariffs exist with variable electricity prices. Here, time- and load-variable electricity tariffs can be distinguished (Ecofys et al., 2009). Due to the brevity of this paper, we only give a short overview of these tariffs. For more details, please refer to Ecofys et al. (2009).

The easiest form of time-variable tariffs is a so called time of use (TOU) tariff which has different price levels for the electricity unit for different hours of the day. These price levels are valid in pre-defined time ranges for the contract duration. More complex examples of time-variable tariffs, in terms of occurrence and number of different price levels, include real time pricing (RTP) and critical peak pricing (CPP). (cf. Ecofys et al., 2009)

In load-variable electricity tariffs, on the contrary, the price per electricity unit depends on the electricity consumption of households within a pre-defined time interval (cf. Ecofys et al., 2009). Examples for this tariff form include tariffs with different price levels depending on the total amount of electricity consumed, e.g., consumption in one month, or the price levels vary according to the average power used over a specific time, e.g., average power in 15 minute time slot. Load variable tariffs originate from energy systems with a high share of conventional power plants in which increasing demand is covered through peak load power plants with higher variable generation costs. In energy systems with an increasing share of RES with negligible variable generation costs, tariffs with variable capacity prices might more adequately reflect capacity dependent generation costs (cf. Hayn et al., 2014c). The concept of such tariffs will be explained in the following.

3.3 Tariffs with variable capacity prices

Another kind of tariffs is not described through different price levels for the consumed electricity units but for capacity limits (in kW). These tariffs can be referred to as curtailable load tariffs or, if the limit is set to zero, as interruptible load tariffs (cf. Ecofys et al., 2009). Within this tariff class, two more concepts can be distinguished: i) self-rationing, for instance described by Panzar & Sibley (1978), and ii) priority service, for instance described by Marchand (1974).

In the first self-rationing concepts, consumers decide on their contracted capacity limit by purchasing a fuse which limits their load at all times (cf. Panzar & Sibley, 1978). In these contracts, the capacity limit is an upper threshold representing the maximum usable capacity amount. Such tariffs on a residential level exist already in some European countries, e.g., in France or Spain (cf. Électricité de France, 2014; Iberdrola, 2014), or in contracts for industrial or large commercial consumers (cf. Oren & Smith, 1992). Based on the contracted capacity limit, customers pay a different capacity price (in €/kW). In tariffs with priority services, however, consumers need to specify an interruption order of their load shares. This is realized through different probabilities for the delivery of that load (cf. Marchand, 1974). Hence, in contrary to self-rationing, consumers purchase a minimum of guaranteed capacity which will be available with a certain probability. For a good overview on these concepts including their advantages and disadvantages, please refer to Woo (1990).

Based on the highlighted disadvantages, mainly the complexity of priority services and the untimely curtailment in the first self-rationing concepts[3], Woo (1990) proposes an enhanced self-rationing concept which allows the provider of electricity services to control (activate/ deactivate) a circuit breaker at the consumers' place in case of capacity shortages. A slightly more complex concept is introduced by Hamlen & Jen (1983); they propose the so called "limiter method" making use of a complex fuse whose output is, when activated, a fraction of the available input. Both approaches enhance the self-rationing concept by substituting the maximum capacity threshold through a minimum threshold making their operating mode more similar to the concept of priority services. While the technical implementation of such tariffs was more challenging in the past, the utilization of smart meters or other controlling soft- and hardware nowadays better supports the implementation. Although smart meters are a precondition for all types of variable electricity tariffs, their large-scale roll-out still lacks behind in Germany due to the related costs and missing standards (Hoffknecht et al., 2012). The commonality between all approaches is the consideration of consumers' willingness to pay for guaranteed capacity which motivates us to use the term "tariffs with variable capacity prices".

4 A concept for service level indicators

Nowadays, residential electricity tariffs with flat or variable electricity prices do not require dedicated SLAs. This is mainly due to the fact that providers do not offer different service levels to their customers in these kinds of contracts. Each customer has the same right to consume electricity and providers cannot exclude selected customers from consumption since security of supply, which is in various aspects influenced by the electricity demand, is treated as a public good. Some researchers already discuss whether supply security should remain a public good or whether it was better marketed as a private good in the context of the changing energy system as in traditional energy-only markets providing capacity is in most cases not refunded (cf. Hayn et al., 2014a; Jacobsen & Jensen, 2012). This thought forms the basis of tariffs with variable capacity prices as described above. The individual customer must decide which service level fits best to its individual willingness to pay and to its individual requirements. This cannot be generalized for all households as the individual behavior and preferences determine the requested service level. Consequently, appropriate SLIs need to be defined for these tariffs in order to allow providers and customers to decide on the contracted service level. In this section we will discuss some conceptual SLIs and respective SLOs for different households, taking the requirements stated in section 2 into account, in order to define the service level in tariffs with variable capacity prices.

Various SLIs can be envisaged to define the service level of tariffs with variable capacity prices. With regard to interruptible and curtailable electricity contracts Smith (1989) and Oren & Smith (1992) name three core elements of these contracts: The frequency of interruptions/ curtailments, the duration per interruption/ curtailment and the warning time a customer receives. Depending on the preferences of different customers, each of these elements can have a different value. For instance, some customers might be willing to accept daily curtailments

[3] Untimely curtailments occur when consumers are limited in their load even if no capacity shortage exists in the system. Panzar & Sibley (1978) give the example of the insomniac who blows his fuse at 4 am.

with a maximum duration of one hour while other customers would only accept weekly curtailments with a maximum of four hours. Consequently, following the terminology of SLAs, the abovementioned elements represent three SLIs to describe the contracted service. An additional SLI, not yet made explicit, is the minimum guaranteed capacity in case of curtailments. In interruptible load contracts this minimum is zero as the service can be interrupted completely. However, in curtailable load contracts it needs to be specified to which level each customer can be curtailed. As already mentioned by Woo (1990) a menu structure for electricity tariffs needs to be simple. Therefore, we suggest the following four SLIs adequately describing the service level of tariffs with variable capacity prices and fulfilling the requirements stated in section 2.1:

- Guaranteed capacity limit: Defines the minimum guaranteed capacity of the consumer which is available at all times of the contract duration. The SLI is given in power units, e.g., watt.
- Duration of curtailment: Defines the maximum length of a single curtailment. The SLI is given in time units, e.g., hours.
- Frequency of curtailment: Defines the maximum number of curtailments during a specified time period, e.g., during one year. The SLI is given in number of curtailments per time period.
- Advance warning time: Defines the minimum warning time for customers to be informed about upcoming curtailments. The SLI is given in time units, e.g., hours.

Theoretically, it is possible to define an unlimited number of different tariffs with varying SLOs for these SLIs in order to satisfy the specific requirements of different customers. In practice, we suggest creating a simple menu of different tariffs to satisfy the majority of residential customers specifying dedicated SLOs, i.e., the respective target values of these SLIs. Therefore, it is necessary to acquire a good understanding of different consumer groups. According to Hayn et al. (2014a) the three most important socio-demographic factors influencing households' electricity demand are the number of occupants living in a household (household size), the net income and the employment status as these factors not only influence the equipment of a household with electric appliances but also the utilization of these appliances. Hence, it can be assumed that these factors also play a vital role in the decision of households on different service levels in electricity tariffs with variable capacity prices. Consequently, they might be an appropriate means to define a tariff menu for different customer segments described through these factors.

A possible approach to derive adequate SLOs for the guaranteed capacity limit, the duration of curtailment and the frequency of curtailment can be the use of models for analyzing the load profiles of households with certain socio-demographic characteristics. Nowadays, German utility companies use standard load profiles to estimate the electricity demand of residential customers, e.g., the H0 profile of the German Association of Energy and Water Industries (BDEW formerly known as VDEW) but it is questionable whether these load profiles can deliver satisfying results with regard to tariffs with capacity limits. An alternative method could be to base the analysis on simulated load profiles of different households from a bottom-up load model as individual differences between households are better distinguishable. Figure 2 a) shows the simulated load duration curves of one four-person household and averaged of 1.000 four-person households; Figure 2 b) shows the related box plot.

Both charts use simulated load profiles with a 15-minute resolution from a residential bottom-up load model developed by Hayn et al. (2014b). It becomes obvious that the average amongst a large group of customers, which is similar to a standard load profile, underestimates the peak demand of households. The peaks occurring from the simultaneous utilization of large electric appliances, e.g., dishwasher and stove, in one household diminish when observing a larger group of households. Analyzing individual load profiles of different households can hold valuable insights regarding appropriate SLOs for the first three SLIs. For the last SLI, the advance warning time, the simulated load profiles cannot be used as this SLI mainly depends on individual preferences and requirements of different households. In this context a household survey can reveal additional insights as individual preferences can be investigated and translated into SLOs. Furthermore, the insights from the analysis and possible surveys can be used to define decision support tools for customers choosing such a tariff. For instance, the service provider could offer making a proposal of appropriate SLOs for a specific customer based on some selected socio-demographic characteristics of that customer, e.g., the household size.

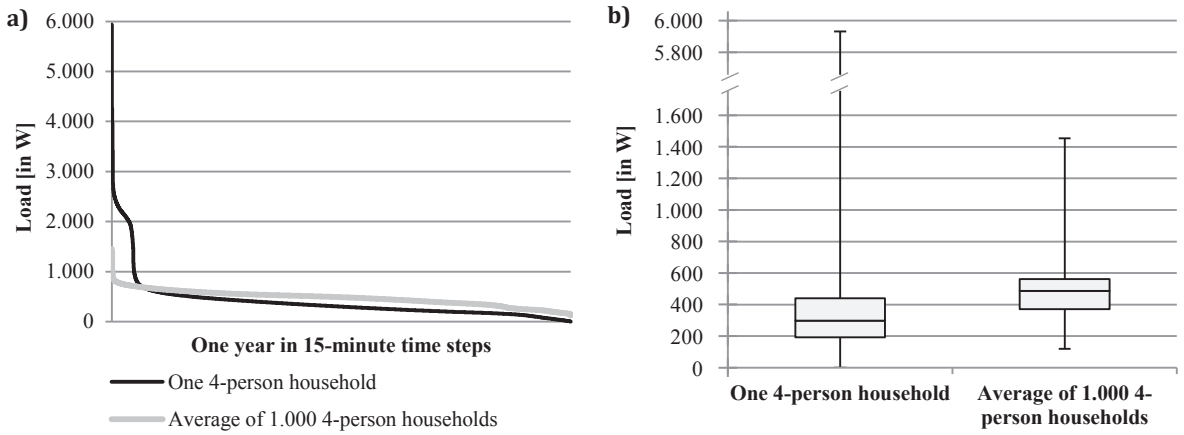


Figure 2 a): Simulated load duration curves; **b):** Box plot of simulated residential load (both based on 15-minute time step bottom-up load model)

5 Conclusion and outlook

This paper proposes a concept of SLIs for residential electricity tariffs. Therefore, it is argued that electricity tariffs are a service and that tariffs with variable capacity prices require the definition of SLIs with related SLOs in order to ensure that commonly agreed service levels between provider and consumer can be met. As a first step in this context four possible appropriate SLIs for tariffs with variable capacity prices are defined: Guaranteed capacity limit, duration of curtailment, frequency of curtailment and advance warning time. With regard to the definition of related SLOs it is proposed to create a simple tariff menu structure based on the requirements of households with different socio-demographic characteristics, e.g., household size. Further, it is suggested to use simulated load profiles to derive adequate SLOs for the first three SLIs and a household survey for the last.

Obviously, this concept is also bounded by some limitations. As currently security of supply is still treated as a public good, the interest in using tariffs with variable capacity prices is, especially from a customer's point of view, very limited. Also from a technical perspective some hurdles still exist before rolling out this kind of electricity tariffs, e.g., the development of appropriate metering hardware. However, the ongoing changes in the energy system require developing new concepts to cope with the upcoming challenges. Additional research with regard to the long-term economic impact and the social acceptance of such tariffs can further clarify the relevance of this concept in future energy systems.

Going forward, further research should follow three directions. First, load profiles of households with different socio-demographic characteristics should be analyzed in order to derive possible SLOs for the suggested SLIs. Second, the impact of the equipment with electric appliances and technological trends should be incorporated in the analysis of appropriate SLOs. Especially when new technologies for residential electricity and heat generation or electric vehicles become more common in households, their demand in the defined SLIs can alter significantly. And third, the possibility of residential load management should be considered when defining SLOs. Recent findings from researchers show that residential demand is, at least to some extent, flexible and therefore might lead to a different consumption behavior with regard to the required guaranteed minimum capacity of households.

References

Berger, T. G. (2007). *Service-Level-Agreements: Konzeption und Management von Service-Level-Agreements für IT-Dienstleistungen*. (1. Aufl). Saarbrücken: VDM Verlag Dr. Müller.

Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen (BNetzA), & Bundeskartellamt (2014). *Monitoringreport 2013: in accordance with § 63 Abs. 3 i. V. m. § 35 EnWG and § 48 Abs. 3 i. V. m. § 53 Abs. 3 GWB*. As of January 2014. Bonn.

DIN Deutsches Institut für Normung e.V. (2013). *DIN 18015-1:2013-09, Electrical installations in residential buildings - Part 1: Planning principles*. Berlin: Beuth Verlag GmbH.

Électricité de France (2014). Les offres d'électricité. https://particuliers.edf.com/offres-d-energie/electricite-47378.html. Accessed 01.07.2014.

Eurostat European Commission (2013). *Energy balance sheets 2010-2011*. (2013 edition). Luxembourg: Publications Office of the European Union.

Hamlen, W. A., & Jen, F. (1983). An Alternative Model of Interruptible Service Pricing and Rationing. *Southern Economic Journal*, *49*, 1108–1121.

Hayn, M., Bertsch, V., & Fichtner, W. (2014a). Electricity load profiles in Europe: The importance of household segmentation. *Energy Research & Social Science*, *3*, 30–45.

Hayn, M., Bertsch, V., & Fichtner, W. (2014b). Residential bottom-up load modeling with price elasticity. In *Sustainable Energy Policy and Strategies for Europe: Proceedings of the 14th IAEE European Energy Conference*. Rom: IAEE.

Hayn, M., Bertsch, V., & Fichtner, W. (2014c). Stromtarife und Technologien im Endkundenmarkt und deren Einfluss auf den Leistungsbedarf von Haushalten aus dem Netz. *uwf UmweltWirtschaftsForum*, *22*, 249–255.

Hoffknecht, A., Wengeler, F., & Wunderer, A. (2012). Herausforderungen und Chancen für einen regionalen Versorger. In Servatius, H.-G., Schneidewind, U., & Rohlfing, D. (Eds.), *Smart Energy*, 113–129. Berlin, Heidelberg: Springer.

Iberdrola (2014). Electricity offers and rates for home. https://www.iberdrola.es/customers/home/electricity. Accessed 01.07.2014.

Jacobsen, H. K., & Jensen, S. G. (2012). Security of supply in electricity markets: Improving cost efficiency of supplying security and possible welfare gains. *International Journal of Electrical Power & Energy Systems*, *43*, 680–687.

Kieninger, A., Westernhagen, J., & Satzger, G. (2011). The Economics of Service Level Engineering. In Sprague, R. H. (Ed.), *2011 44th Hawaii International Conference on System Sciences: (HICSS); 4 - 7 Jan. 2011, Koloa, Kauai, Hawaii*. Piscataway, NJ: IEEE.

Marchand, M. G. (1974). Pricing power supplied on an interruptible basis. *European Economic Review*, *5*, 263–274.

Ecofys, ENCT & BBH (2009). *Einführung von lastvariablen und zeitvariablen Tarifen*. Im Auftrag der Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen.

Oren, S. S., & Smith, S. A. (1992). Design and Management of Curtailable Electricity Service to Reduce Annual Peaks. *Operations Research*, *40*, 213–228.

Panzar, J. C., & Sibley, D. S. (1978). Public Utility Pricing under Risk: The Case of Self-Rationing. *The American Economic Review*, *68*, 888–895.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, *49*, 41–50.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing*, *64*, 12–40.

Smith, S. A. (1989). Efficient menu structures for pricing interruptible electric power service. *Journal of Regulatory Economics*, *1*, 203–223.

Wilson, C., & Dowlatabadi, H. (2007). Models of Decision Making and Residential Energy Use. *Annual Review of Environment and Resources*, *32*, 169–203.

Woo, C.-K. (1990). Efficient electricity pricing with self-rationing. *Journal of Regulatory Economics*, *2*, 69–81.

Quality of Service Product Differentiation in Smart Grids

Alexander Schuller, schuller@fzi.de, FZI Research Center for Information Technology
Florian Salah, florian.salah@kit.edu, Karlsruhe Institute of Technology
Christian Will, christian.will@student.kit.edu, Karlsruhe Institute of Technology
Christoph M. Flath, christoph.flath@uni-wuerzburg.de, University of Würzburg

The Smart Grid paradigm enables bidirectional communication and control between the demand and supply side in the power system. Metering, information and communication technology (ICT) and control infrastructure is increasingly rolled out, but the economic implications of this roll-out have not been considered sufficiently yet. Demand side management offers a framework for changing the role of the consumer in the power system. In particular demand side flexibility which can be achieved by shifting, curtailing, or in some cases the increase in power draw needs to be harvested in an efficient manner. In this paper we propose a basic classification framework for quality differentiated products which enable consumers to self-select with respect to their individual valuation for the particular end energy usage.

1 Introduction

The Smart Grid enables bidirectional communication between distributed actors and resources in the power system. Through the increase of variable renewable energy sources on the supply side, the share of potentially uncontrollable generators requires a more flexible demand side. Currently, the system balance is predominantly maintained by the supply side, in particular the system balancing risk (e.g. the deviation between forecast generation and demand) is only addressed through controllable generators and system reserves. In the near future the generation risk (i.e. the output uncertainty of variable renewable energy sources) needs to be borne to a higher extent to the currently rather passive demand side. In order to activate demand side flexibility, economic incentives need to be designed. These incentives need to address the individual valuation and application scenario of each customer group. This requires the development of new products and considerations about the appropriate market environment.

The goal of this paper is to present a structured morphological approach (following (Zwicky, 1967)) that systematically captures the design options for power products based on quality of service differentiation. Quality of service in our context considers different risk components, i.e. the service quality is affected by these components. The morphologic approach captures the possible and viable combinations for such products. Each of the given combinations can then be assessed from the perspective of the involved stakeholders (e.g. system operators, customers, energy service companies), in our case under consideration of the German regulatory framework and (retail) electricity market conditions.

2 Related Work

Quality of service (QoS) has been analyzed form different perspectives in service research (Cronin Jr & Taylor, 1992), indicators and several models that define and evaluate service quality have been proposed (Ghobadian, Speller, & Jones, 1994; Seth, Deshmukh, & Vrat, 2005) and the role of the customer has been evaluated in given service related environments (Barrutia & Gilsanz, 2013). For the energy domain, quality of service is yet still mainly perceived to be a strictly technical property (e.g. voltage limits). This conception has not been altered so far, even though the extension of the QoS term has promising possibilities.

Product differentiation in the power sector so far mainly focuses on transmitting some sort of scarcity signal for the availability of electricity in a given time span. Depending on the customer size and type, the variable generation cost of the system marginal power plant can be communicated via a variable pricing scheme to the customers (C. Woo et al., 2014; Albadi & El-Saadany, 2008). Variable pricing schemes such as real time pricing (RTP) are well known instances of this approach. Other, simpler forms of tariffs set higher prices if high overall load needs to be covered.

In addition, research and practice focus on direct load control or price-based coordination schemes for managing flexible loads (Albadi & El-Saadany, 2008). However, these approaches induce uncertainty for individual customers energy prices may fluctuate or the energy quantity delivered is curtailed. These uncertainties expose retail customers to concrete risk situations. For the standard case of risk-averse customers this may result in a potential hindrance for the establishment and acceptance of innovative electricity rates (C.-K. Woo, Kollman, Orans, Price, & Horii, 2008). Automation technology can address these obstacles, but customers still appear to be inclined towards less complex electricity products (Duetschke & Paetz, 2013).

Further work concentrates on the consumer acceptance of different conventional attributes of electricity (Kaenzig, Heinzle, & Wuestenhagen, 2013), or assesses the willingness to pay for these attributes. In particular the assessment of environmentally friendly generation is at the center of attention (Ozaki, 2011). Several publications suggest that customers (not only in Germany) have a higher willingness to pay for electricity supply from renewable sources (Borchers, Duke, & Parsons, 2007; Roe, Teisl, Levy, & Russell, 2001).

Supplementary to the uncertainty aspects and the different aspects of electricity delivery, additional components like the mentioned source of power supply and in particular usage restrictions need to be considered for a comprehensive analysis of the options of quality of service product differentiation in a Smart Grid environment, a gap this work addresses.

3 Quality Differentiated Products in the Smart Grid

To better account for customer heterogeneity with respect to risk preferences and flexibility endowments, greater differentiation of energy pricing is required. Using a structured morphological approach we want to systematically identify and subsequently discuss design options for differentiated transaction objects for future retail energy services. To this end, we isolate individual components of these retail energy services and compile possible design options in matrix representation. Then, each combination of the different design options yields a potential solution candidate which can be evaluated. In the following, we group design options in three categories and present exemplary morphological boxes for each of the categories (Figure 1).

3.1 Design Options for Service Risk

The design options for service risk encompass in particular the (limited) uncertainty about execution, volume, time of delivery, price and interruption. The uncertainty regarding the execution is a dimension that is not considered in the current supply paradigm, since every load has to be served, which could potentially lead to very high overall system

Characteristic		Design Option			
Uncertainty	Execution	No	Yes		
	Volume	No	Limited (minimum volume)	Unlimited	
	Time of delivery	No	Yes (deadline)		
	Price	No	Limited (limit price)	Unlimited	
	Interruption	No	Yes		
Pricing	Energy	Without	Linear (kWh)	kWh package	
	Power	Without	Linear (kW)	Maximum value	Median value
	Service	Without	km	Heating / cooling power	
	Temporal differentiation	Static	Time based	Event based	
	Locational differentiation	Uniform price	Zonal price	Nodal price	Roaming price
Additional Components	Generation mix	Without	Regional	Green	Non-nuclear
	Extended electrical quality (e.g. reactive power guarantee, lower voltage range)	No	Yes		
	Usage restriction	No	Mobility	Heating	Cooling
	Product bundle	Without	Parking	Comfort & Security (Smart home)	Several energy sources

Figure 1: Overview of the morphological design options.

costs. But the addition of this possibility can greatly increase demand side flexibility and overall supply security. Similar in its system effect to the potential non-execution is interruption of a started service following a predefined process. A volume uncertainty describes the possibility of load curtailment. The uncertainty about the time of delivery is a dimension which addresses the increasingly volatile character of the generation side. Finally, the price which is realized could also be uncertain, a characteristic that needs to be addressed in a satisfactory manner without high transaction costs for retail customers as already mentioned above.

3.2 Design Options for Service Pricing

The main characteristics for service pricing are the energy and power component, the service or end-usage component and temporal and locational differentiation of service delivery. Traditional electricity rates mainly encompass the energy component which is mostly constant and linear in its shape, accounting for every unit of energy used. Power or demand charges are added in order to set incentives to reduce the synchronization of applications with a high power rating. However, these incentives can be contrary to the system needs in situations in which load has to be increased in order to use available fluctuating renewable energy sources (Caramanis, Bohn, & Schweppe, 1982). A further differentiation can be performed by the end-usage type of energy. For electric mobility applications km instead of kWh can be priced, for heating or cooling services, pricing can be based on cooling and heating amounts. Finally, a differentiation by time of delivery, which captures the mentioned variable generation costs in the power system, and a differentiation by geographical location can be pursued. The geographical location can have specific demand and supply features which need to be accounted for in the price, which is currently implemented e.g. by zonal or nodal pricing schemes.

3.3 Design Options for Additional Components

Additional design components with potential for service differentiation are the supply source or generation mix, the technical power quality (e.g. voltage deviations and reactive power ratios), usage restrictions and contractual bundling possibilities with other products. The differentiation by generation source can enable regional power to be more attractive and support the integration of green generation sources, but can be difficult from a physical perspective since the power supply route can only be controlled in a very limited manner or in geographically constrained grid segments. Furthermore, the extended electrical quality, i.e. the (local) voltage deviations and the reactive power characteristics at a location can be used for service differentiation, in particular in regional settings. Power exclusively used to charge electric vehicles with a specific charging infrastructure or for heating or cooling applications, represents usage restriction based differentiation. Last but not least, bundling of the energy service with other products, like smart home security and comfort options, can be employed to complete the picture of energy service product differentiation in the Smart Grid.

4 Conclusion and Outlook

This paper makes a first step to set a frame for product characteristics and their design options for the description of energy services, which can be differentiated by quality of service attributes. The valid and innovative combination of the service design options can help to activate the flexibility potential of the demand side in a Smart Grid environment. Our work thus addresses the design of economically sound products, building on the infrastructure provided by the Smart Grid. Further work must address the design options in more detail, propose new products and consider current and future regulatory environments for a valid frame of action.

References

Albadi, M. H., & El-Saadany, E. (2008). A summary of demand response in electricity markets. *Electric Power Systems Research*, *78*(11), 1989–1996.

Barrutia, J. M., & Gilsanz, A. (2013). Electronic service quality and value: Do consumer knowledge-related resources matter? *Journal of Service Research*, *16*(2), 231-246. Retrieved from http://jsr.sagepub.com/content/16/2/231.abstract doi: 10.1177/1094670512468294

Borchers, A. M., Duke, J. M., & Parsons, G. R. (2007). Does willingness to pay for green energy differ by source? *Energy Policy*, *35*(6), 3327 - 3334. Retrieved from http://www.sciencedirect.com/science/article/pii/S0301421506005131 doi: http://dx.doi.org/10.1016/j.enpol.2006.12.009

Caramanis, M. C., Bohn, R. E., & Schweppe, F. C. (1982). Optimal spot pricing: practice and theory. *Power Apparatus and Systems, IEEE Transactions on*(9), 3234–3245.

Cronin Jr, J. J., & Taylor, S. A. (1992). Measuring service quality: a reexamination and extension. *The journal of marketing*, 55–68.

Duetschke, E., & Paetz, A.-G. (2013). Dynamic electricity pricing - which programs do consumers prefer? *Energy Policy*, *59*(0), 226 - 234. Retrieved from http://www.sciencedirect.com/science/article/pii/S0301421513001791 doi: http://dx.doi.org/10.1016/j.enpol.2013.03.025

Ghobadian, A., Speller, S., & Jones, M. (1994). Service quality: concepts and models. *International Journal of Quality & Reliability Management*, *11*(9), 43–66.

Kaenzig, J., Heinzle, S. L., & Wuestenhagen, R. (2013). Whatever the customer wants, the customer gets? exploring the gap between consumer preferences and default electricity products in germany. *Energy Policy*, *53*(0), 311 - 322. Retrieved from http://www.sciencedirect.com/science/article/pii/S0301421512009469 doi: http://dx.doi.org/10.1016/j.enpol.2012.10.061

Ozaki, R. (2011). Adopting sustainable innovation: what makes consumers sign up to green electricity? *Business Strategy and the Environment*, *20*(1), 1–17. Retrieved from http://dx.doi.org/10.1002/bse.650 doi: 10.1002/bse.650

Roe, B., Teisl, M. F., Levy, A., & Russell, M. (2001). {US} consumers willingness to pay for green electricity. *Energy Policy*, *29*(11), 917 - 925. Retrieved from http://www.sciencedirect.com/science/article/pii/S0301421501000064 doi: http://dx.doi.org/10.1016/S0301-4215(01)00006-4

Seth, N., Deshmukh, S., & Vrat, P. (2005). Service quality models: a review. *International Journal of Quality & Reliability Management*, *22*(9), 913–949.

Woo, C., Sreedharan, P., Hargreaves, J., Kahrl, F., Wang, J., & Horowitz, I. (2014). A review of electricity product differentiation. *Applied Energy*, *114*(0), 262 - 272. Retrieved from http://www.sciencedirect.com/science/article/pii/S0306261913008179 doi: http://dx.doi.org/10.1016/j.apenergy.2013.09.070

Woo, C.-K., Kollman, E., Orans, R., Price, S., & Horii, B. (2008). Now that california has ami, what can the state do with it? *Energy Policy*, *36*(4), 1366–1374.

Zwicky, F. (1967). The morphological approach to discovery, invention, research and construction. In F. Zwicky & A. Wilson (Eds.), *New methods of thought and procedure* (p. 273-297). Springer Berlin Heidelberg.

Holistically Defining E-Mobility: A Modern Approach to Systematic Literature Reviews

Jan Scheurenbrand, jan.scheurenbrand@student.kit.edu, Karlsruhe Institute of Technology
Christian Engel, christian.engel2@student.kit.edu, Karlsruhe Institute of Technology
Florin Peters, florin.peters@student.kit.edu, Karlsruhe Institute of Technology
Niklas Kühl, niklas.kuehl@kit.edu, Karlsruhe Institute of Technology

This work presents two main contributions to the fields of literature and e-mobility research. As there is no existing common and consistent definition of e-mobility, we aim at acquiring relevant literature in order to suggest a holistic definition of the term "e-mobility", incorporating multiple perspectives on the topic. This definition is intended to be able to serve as a common ground for further research in the field of e-mobility. To achieve this goal we introduce a software tool for conducting the search component of systematic literature reviews with a special focus on handling large amounts of search results efficiently. It addresses all researchers with an interest or need in literature reviews.

1 Introduction

E-mobility is a widely discussed topic in recent research, with almost every automotive company acting on this emerging market (Schlick, Hertel, Hagemann, Maiser, & Kramer, 2011). Even though e-mobility is very present in modern economies, there is no holistic definition existing which comprehensively covers the core aspects of e-mobility paired with co-opted businesses. The necessity of a consistent definition is underlined by the fact that polysemous expressions can cause different associations depending on the receiving individual (Tremmel, 2004). To approach a comprehensive solution, we use existing literature in a structured way to find a suitable definition by means of a systematic literature review (SLR). As the large amount of present literature on this topic has to be dealt with in a structured manner, we approach an efficient way to solve this problem. To face this challenge, our work presents a methodology of finding and analyzing literature by using a custom-designed software solution and a statistics-based filtering and synthesizing method which are explained in a detailed way. This multi-step methodology is then used to solve the initial problem of finding a holistic definition for e-mobility.

2 Methodology: Systematic Literature Review

In evidence-based research, systematic literature review (SLR) is an important element that is attempting to collate all evidence that fits pre-specified eligibility criteria in order to address a specific research question with the aim to minimize bias by using explicit, systematic methods (Higgins, Green, et al., 2008).
One key characteristic of systematic literature reviews is an explicit, reproducible methodology for systematic search that attempts to identify all studies that would meet the eligibility criteria (Higgins et al., 2008).
Finding relevant literature for a specific topic has always been hard and cumbersome work. Until ten to twenty years ago, it meant going to a library and searching relevant publications in printed media. Through the Internet, literature research changed: Since publishers and independent search engines offer online libraries with extensive search capabilities, it is now easier than ever to find a great amount of distinct sources.

One of the main contributions of this work is to introduce a tool for conducting the search component of systematic literature reviews with special focus on handling large amounts of online literature search results.
The data basis for the tool are online libraries and search engines such as Google Scholar. As those search engines are offered as web services, their feature-set and speed is traditionally inferior to native applications. They especially lack typical SLR functionalities like filtering, sorting and exporting capabilities. Google Scholar, for example, offers neither advanced result sorting nor export functionality. Additionally, the underlying ranking algorithms are often unknown and subject to change and therefore not suited for SLR. Google for instance only states, that they "rank documents the way researchers do, weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature" (Google, 2015). Previous research (Beel & Gipp, 2009) has shown that the citation count is the highest weighed factor in Google Scholar's ranking algorithm. As a consequence, Google Scholar seems to be more suitable for finding standard literature than gems or articles by authors advancing a new or different view from the mainstream (Beel & Gipp, 2009).
With these drawbacks in mind, the tool is designed with the following objectives

1. to provide a systematic approach to find as many relevant sources as possible,
2. to minimize bias,
3. to provide a thorough, objective and reproducible workflow,
4. to be conducted in several, clearly defined phases,
5. and to use various sources.

In order to achieve these objectives, we developed a modular tool offering a convenient workflow for finding, filtering, downloading and exporting relevant literature. To the best of our knowledge, there exist no other solutions that offer a similar functionality.

2.1 Data retrieval

The tool is designed to support multiple data providers like search engines and online libraries in order to perform a federated search for relevant publications. Search options include searching for multiple search terms, using logical operators like "and" and "or" in the search term and searching only in titles ("allintitle").
The objective is to fetch the complete result set $\mathbb{P}_i$ from the search engine provider i, in order to overcome the problem of missing results that arises by loading only the first j pages.
Retrieving the complete result set $\mathbb{P}_i$ leads to optimum precision and recall values of 1, excluding false negatives. Nevertheless, $\mathbb{P}_i$ may be fetched unsorted or improperly sorted and include false positives (reported results, which are not relevant) – problems, we address in further steps.
If available, public application programming interfaces (APIs) are used for the purpose of result acquisition. Otherwise raw data is acquired by means of web scraping techniques like HTTP programming and HTML parsing. For speeding up the process, results can be loaded in parallel from different providers, resulting in significantly reduced retrieval times.
In a first step, a data retrieval module for Google Scholar is implemented. Google Scholar is chosen because of its massive multidisciplinary index and broad popularity.

2.2 Database normalization and merging of the results

The acquisition of the data is followed by a database normalization process in order to be able to work with the results locally in a uniform manner. Therefore, a normalization function $normalize(p) := rawdata \rightarrow source$ is used to compute the normalized partial result set $\mathbb{P}'_i$ from provider i as follows: $\mathbb{P}'_i = \bigcup_{p_i^j \in \mathbb{P}_i} normalize(p_i^j)$. In practice, the

normalization process includes extracting information like the list of authors or URLs to direct PDF downloads in a sensible way. Results from different providers can then be merged to a complete and normalized result set $\mathbb{P}' = \bigcup_i \mathbb{P}'_i$ for further use.

2.3 Faceted search by filtering results

To narrow down the result set from possibly thousands of results, the first step is to filter the results locally by different properties. The filter process must correspond with the objective – e.g. considering only articles published in the last ten years or with at least 100 citations. Therefore a module is provided, in which m Boolean filters can be executed consecutively in order to obtain the filtered result set of articles $\mathbb{P}'' = \{p \in \mathbb{P}' \mid filter_i(p)\ \forall i \in m\}$ with a Boolean filter functions such as e.g. $filter_1(p) = citecount(p) > 100$.

2.4 Marking relevance

After narrowing down the list of possibly relevant articles by applying the filters above, the next iteration of reducing the result set is a manual filtering step by relevance for the topic. Irrelevant search results might e.g. include polysemous expressions. For this time-consuming task, the tool provides a user interface-assisted skim-and-mark process, which enables going through metadata such as title and abstract in an efficient way. This process is designed to traverse the result set as fast as possible with a maximum of accuracy. Relevance marking leads to the filtered and marked result set $\mathbb{P}''' = \{p \in \mathbb{P}'' \mid selected(p)\}$.

2.5 Exporting results for further use

An important step in the SLR process is the documentation of the search. Therefore the review has to be instantly documented in sufficient detail and the unfiltered search results should be saved and retained for possible reanalysis (Kitchenham, 2004). Considering that, our tool allows to export the retrieved search results and additional meta data like the search term(s) and search timestamp in the Microsoft Excel format for further data processing and statistical analysis. Additionally, the tool allows for an easy download of single and multiple articles if the particular PDF URLs were acquired in the first step. Export functionality is generally available for both $\mathbb{P}''$ and $\mathbb{P}'''$.

2.6 Workflow-driven user interface

In order to conduct the search component of systematic literature reviews in an efficient and simple-to-use way, we have to significantly improve its usability for the end-user. Therefore a graphical user interface guiding the user through the three-step process of searching, filtering and exporting relevant literature is implemented.
Especially the skim-and-mark process for marking relevance needs a radical performance improvement. This gives the user a fast and reliable method to quickly grasp the essence of an article and either mark it as relevant or skip it. Therefore, results are displayed in a table with an additional single detail view displaying the currently selected entry. With this layout, the user gets an overview of the results and a fast way to read through the abstract in order to identify relevance. With keyboard shortcuts for going up and down in the result set and toggling the marking, one can quickly traverse the result set.

2.7 Technical details

Our tool is implemented in Java, leveraging Java 8's new features. For the user interface JavaFX 2 was used. The HTTP communication is implemented on top of Apache HTTP client. HTML content is parsed with JSoup. Because of Java's

platform independence, our tool is usable on every major operating system and has only minimal system requirements. In future work, we will incorporate additional features in the tool, including the retrieval of data from more sources, adding more filtering techniques and integration with literature management software like Papers.app and Citavi.
The presented tool is available on `https://paperfinderapp.com`. The source of the tool is licensed under the terms of the GPL v2.0 and available on GitHub: `https://github.com/scheja/paperfinder`.

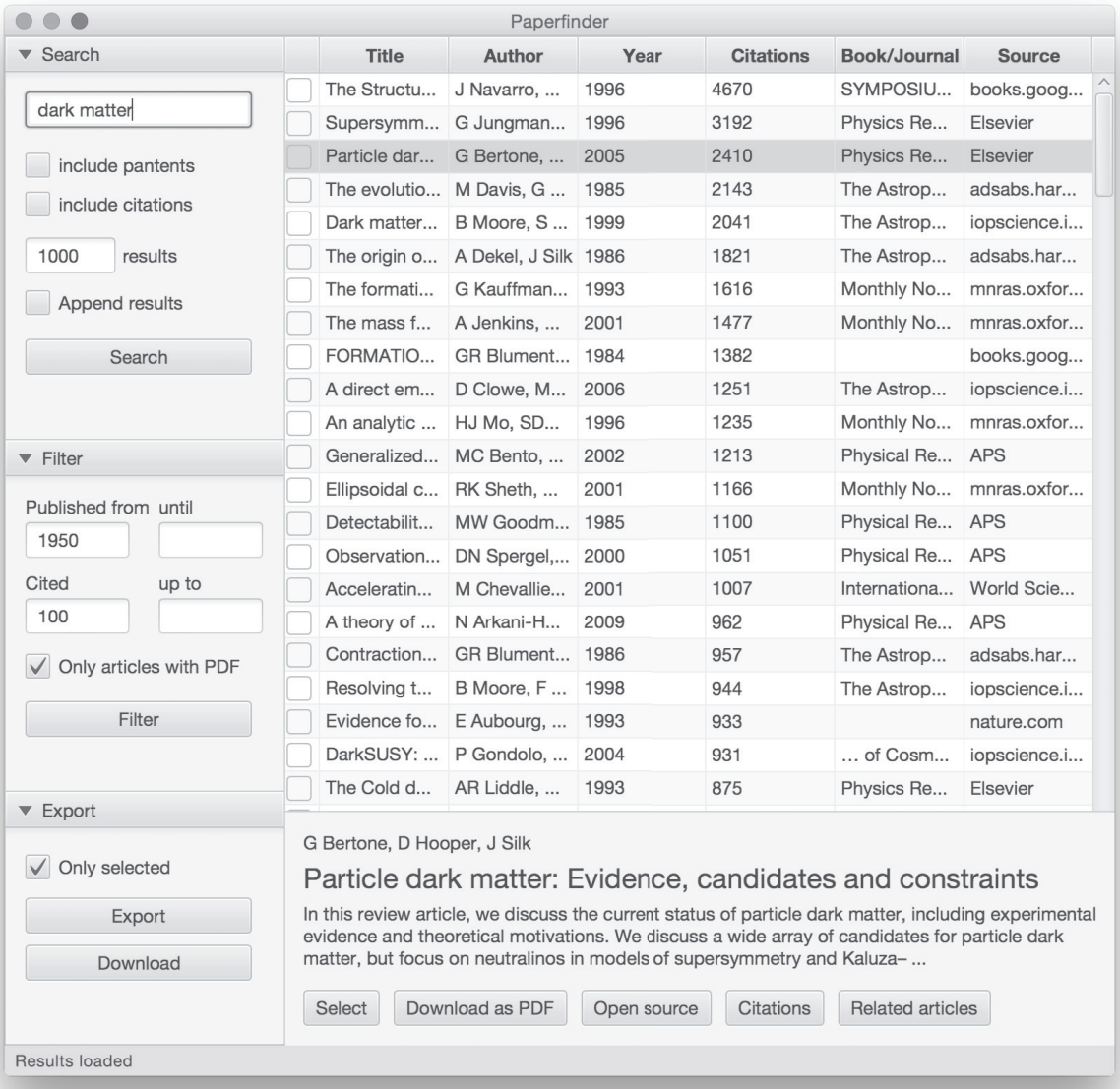


Figure 1: Screenshot of the user interface

3 Analysis of e-mobility findings

To apply the SLR-based technique on the large amount of available data in the wide field of e-mobility (see Table 1), the need of core keywords emerges in order to efficiently find relevant literature. Therefore, we select three keywords which

are directly linked to e-mobility. Due to language barriers we restrict our selection to German and English. The direct linkage is fulfilled by the word "e-mobility" itself and additionally "electric mobility" and "Elektromobilität" (which is German for electric mobility). To strengthen the relevance of our research we focus on papers where our particular keywords are completely mentioned in the title. Furthermore, the "allintitle"-search is conducted due to the assumption that if the whole keyword occurs in the title of a found paper it is likely to discover defining criteria in it. To structurally access the literature, we use the IT-based tool described above in the Google Scholar Database. The data was retrieved on December 7th, 2014. Even though our research terms are restricted by language and the keywords themselves, the raw data exported from Google Scholar is still not manageable, as the following overview of the searching hits shows:

"electric mobility":	216 [number of papers]
"e-mobility":	149 [number of papers]
"Elektromobilität":	270 [number of papers]

Table 1: Results of the keyword search

In order to optimize the research effort and to analyze only literature which is exclusively relevant for the goal of finding a holistic definition of e-mobility, a filtering system is established.

3.1 Four-level-filtering

We extend the filtering techniques described above: $\mathbb{P}'' = \{p \in \mathbb{P}' \mid filter_i(p)\ \forall i \in m\}$. We employ $m = 4$ Boolean filter functions $filter_i(p)$:

1. On a first glance at the exports, the problem of irrelevant domains like biology, chemistry or physics being included in the searching results make a relevance check as a first step necessary. Consequently all papers with irrelevant domains are deleted from the list of results.

$$filter_1(p) = domain(p) \notin \{biology, chemistry, physics\}$$

2. Secondly, literature which is still written in other languages than German or English (less than five sources) is eliminated.

$$filter_2(p) = language(p) \in \{english, german\}$$

3. Analyzing the quantity of publications over a time span from the early 1950s until 2014 a large growth since 2007 is discovered which shows that this is the relevant period to focus on (Figure 2).

$$filter_3(p) = yearOfPublication(p) > 2007$$

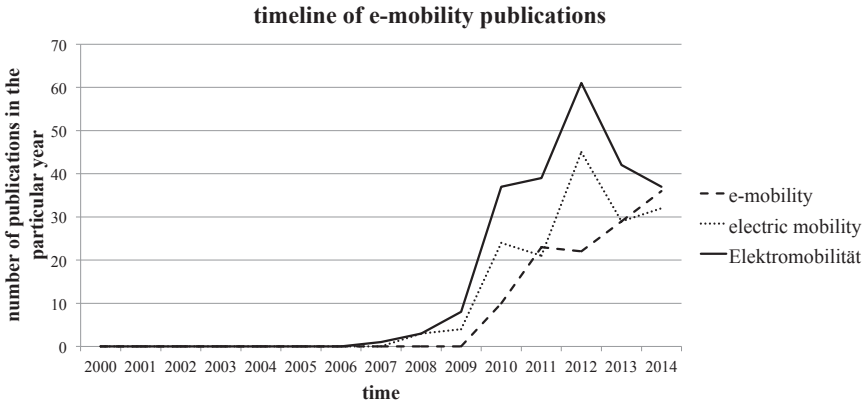


Figure 2: e-mobility publications plotted over time

4. As a last step, we filter the sources by citation count. The sum of citations of the literature in the set P is defined as follows:

$$citesum(P) := \sum_{p \in P} citecount(p)$$

The following fraction describes the proportion of citations, literature with a higher citation count than source p accumulate in relation to the total citation count of all sources in $\mathbb{P}'$:

$$citefrac(p) := \frac{citesum(q \in \mathbb{P}' \mid citecount(q) \geq citecount(p))}{citesum(\mathbb{P}')}$$

The sources we like to focus on are those in the lower quantile of the $citefrac$-measure. For illustrative reasons we show the literature-citations graph for every keyword (Figure 3) and additionally present the cumulative citations graph (Figure 4) with a line which visualizes the 50 percent border of the overall citations. This results in the intersection of the two graphs which then highlights the lower area quantile. Due to the assumption that this lower area quantile of the overall citations represents the view on e-mobility of the scholary majority, the resulting number of relevant papers is equivalent to the x-coordinates of the intersection mentioned above.

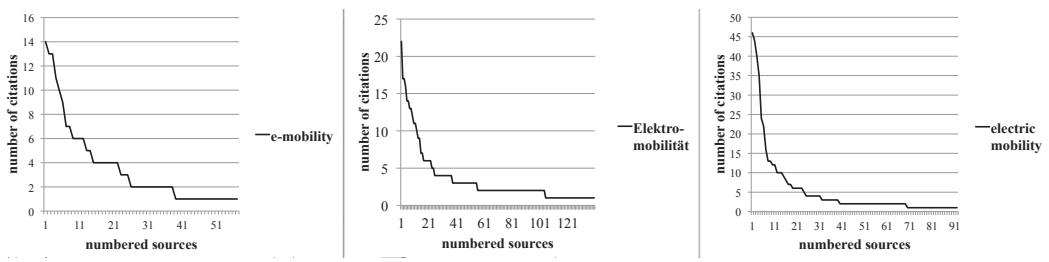


Figure 3: citation curves

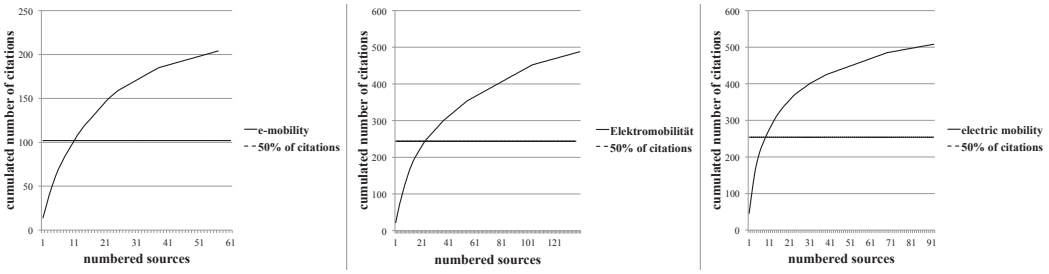


Figure 4: cumulative citation curves

The number of papers which are finally analyzed in order to create a definition is shown below:

"electric mobility":	10 [number of papers]
"e-mobility":	11 [number of papers]
"Elektromobilität":	24 [number of papers]

3.2 Synthesizing approach of the filtered result set

Studying the selected literature aims at resulting in a dual layer system which consists of a top level represented by a basic clustering and particularly lower levels of more detailed components. The lower level components are sequentially generated while analyzing the sources and subsequently structured in a concept matrix in order to discover commonalities for the top level clustering (Webster & Watson, 2002). Every mind-set or concrete definition of e-mobility that can be found during the research process serves as a new lower level component for the clustering or is matched to an existing one if the particular mind-set on e-mobility is the same.This then results in a holistic definition of the broad e-mobility sector.

To achieve a reasonable precision in mutual matching and checking of the results, we recommend a multiple number of researchers to independently create individual research tables and merge them in the end. In our case we realised this approach with two researchers. Generally speaking, the larger the number of independently working researcher the more objective are the results.

4 Perspectives of the definition

In this chapter firstly the distinct components of the definition are presented clustered in their top level categories, as three perspectives are discovered in literature: The technological, market-oriented and social perspective. Afterwards the complete comprehensive definition of e-mobility is orchestrated from its components. Generically we focus on basic definition techniques as we firstly analyze and then synthesize the findings (Borkowski, 1956). The results below reflect the widely spread understanding that researchers have on e-mobility.

4.1 Technological perspective

A basic technological characteristic of e-mobility is that the particular vehicles use an electric drive as a substitution of the classic combustion engine (Zanker, Lay, & Stahlecker, 2011; Maia, Silva, Araújo, & Nunes, 2011). In addition the main energy source (e.g. a battery) is portable (Martiny & Schwab, 2011; Rua et al., 2010) and can differ on available range and charging speed which leads to various usage types (Leitinger & Litzlbauer, 2011; Brauner, 2008). Another common understanding of e-mobility is the continuum view on the degree of electrification (Yay, 2010; Link, 2012). In

practice, researchers distinguish between electric vehicles which vary in the share of distinct energy sources. Speaking on a macroscopic top-level, vehicles differ mainly in four types of energy supply: The internal combustion engine vehicle (ICEV), the hybrid electric vehicle (HEV), the plug-in hybrid electric vehicle (PHEV) and the battery electric vehicle (BEV). The degree of electrification continuously increases with this mentioned order as figure 5 shows. But even though the degrees of electrification differ, the range of e-mobility comprises HEV, PHEV and BEV as electromotive entities (Bioly, Kuchshaus, & Klumpp, 2012; Franke & Krems, 2013; Schwedes, Kettner, & Tiedtke, 2013).

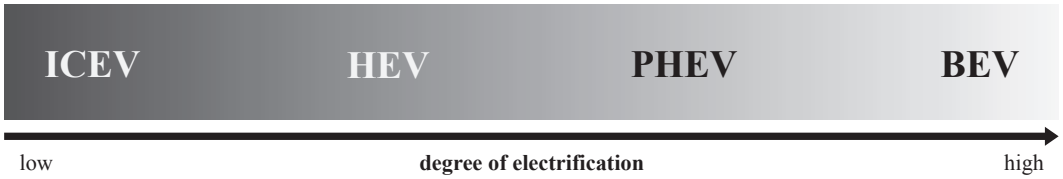


Figure 5: continuum view on electrification

4.2 Market-oriented perspective

From a market-based view, the high interdisciplinary cross-linkage between different industries and institutions such as automotive industry, mobility services, information technology, energy suppliers and the government, is characteristic of e-mobility (Hanselka & Jöckel, 2010; Galus et al., 2012; Leurent & Windisch, 2011). Researchers agree on e-mobility being a highly connective element for still existing market incumbents. Consequently new markets and business models are created in the field of e-mobility (Figure 6).

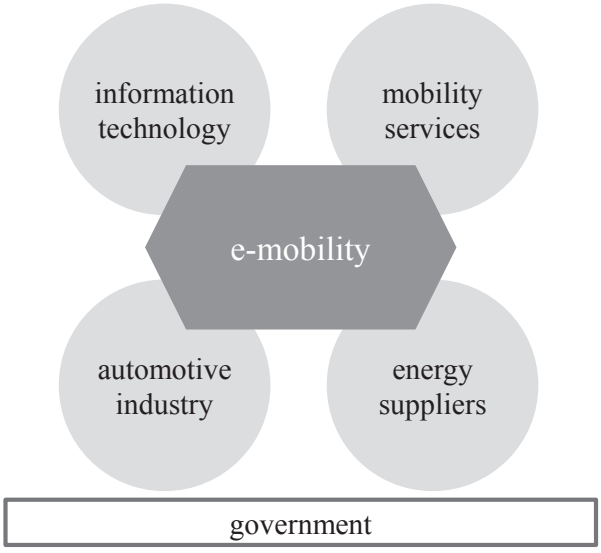


Figure 6: e-mobility as an interface

Another widely mentioned aspect are the current intensive innovation dynamics in the domain of e-mobility (Schlick et al., 2011; Schiavo, Delfanti, Fumagalli, & Olivieri, 2013). Although this topic is very present, we still exclude it from our definition because we do not want to predict any future expectations and to keep our results atemporal, as far as possible.

4.3 Social perspective

Regarding e-mobility from a social perspective, mobility in general and e-mobility in our case serve people's mobility needs by transporting them from one place to another (Bioly et al., 2012). The need for mobility has been deeply rooted in societies all over the world for eras. People do not need mobility per se but the utility that derives from it, e.g. in order to pursue a business or to conduct social activities (Kranton, 1991). E-mobility addresses a possibility to fulfill these needs in an energetically sustainable way if the demand for sustainability is met by mainly using renewable energy as a power source which integrates green energy into the minds of the people (Pehnt, Höpfner, & Merten, 2007; Schlick et al., 2011; Faria, Moura, Delgado, & de Almeida, 2012; Stamp, Lang, & Wäger, 2012; Franke, Cocron, Bühler, Neumann, & Krems, 2012; Bures et al., 2013).

5 Definition

In respect of all components mentioned above, we suggest the following holistically merged definition of e-mobility:

> ***"E-mobility (electric mobility) is a highly connective industry which focuses on serving mobility needs under the aspect of sustainability with a vehicle using a portable energy source and an electric drive that can vary in the degree of electrification."***

6 Conclusion and prospect

In order to come up with a holistic definition of e-mobility by combining existing mind-sets and definitions, we employed a systematic literature review approach. To achieve this goal, we introduced a software tool with special focus on efficiency. This tool offers a convenient workflow for finding, filtering, downloading and exporting relevant literature. The tool is both able and designed to be applied to other fields of science to handle large amounts of sources. The support of the tool makes the systematic discovering of relevant papers possible and gives a sufficient overview about the existing literature. The limiting factor hereby lies in the index of the leveraged search engine providers. Looking at future projects, this problem can be overcome by increasing the number and variety of search engine providers. In addition to this, another factor, the choice of new emerging keywords of "e-mobility", could have an impact on the resulting definition.

Based on the results acquired with the tool, we employed a methodology with a focus on several selection criteria to manage and structure the literature in an efficient and transparent way. This procedure enabled us to discover diverse definition components, which were clustered to top-level categories and merged to one overall definition of e-mobility. In doing so, its complexity and multi-dimensionality were considered and handled resulting in the definition described above. The holistic character of the e-mobility definition was determined by the macro-perspective on the broad range of e-mobility. Additionally, the distinct perspectives could serve as entry points to multiple sub-definitions to enable every researcher to pick the aspect he would like to focus on. This concept then could provide more detail on a micro perspective. In this context, the applicability of the above developed definition can be examined by e-mobility experts to check whether the intended definition is able to serve as a common and solid ground for further research in the field of e-mobility.

References

Beel, J., & Gipp, B. (2009). Google scholar's ranking algorithm: The impact of articles' age (an empirical study). *2014 11th International Conference on Information Technology: New Generations*, *0*, 160-164. doi: http://doi.ieeecomputersociety.org/10.1109/ITNG.2009.317

Bioly, S., Kuchshaus, V., & Klumpp, M. (2012). *Elektromobilität und ladesäulenstandortbestimmung: Eine exemplarische analyse mit dem beispiel der stadt duisburg* (Tech. Rep.). ild Schriftenreihe Logistikforschung.

Borkowski, L. (1956). Über analytische und synthetische definitionen. *Studia Logica*, *4*(1), 7–61.

Brauner, G. (2008). Infrastrukturen der elektromobilität. *e & i Elektrotechnik und Informationstechnik*, *125*(11), 382–386.

Bures, T., Nicola, R. D., Gerostathopoulos, I., Hoch, N., Kit, M., Koch, N., ... others (2013). A life cycle for the development of autonomic systems: The e-mobility showcase. In *Self-adaptation and self-organizing systems workshops (sasow), 2013 ieee 7th international conference on* (pp. 71–76).

Faria, R., Moura, P., Delgado, J., & de Almeida, A. T. (2012). A sustainability assessment of electric vehicles as a personal mobility system. *Energy Conversion and Management*, *61*, 19–30.

Franke, T., Cocron, P., Bühler, F., Neumann, I., & Krems, J. (2012). Adapting to the range of an electric vehicle–the relation of experience to subjectively available mobility resources. In *Proceedings of the european conference on human centred design for intelligent transport systems, valencia, spain* (pp. 95–103).

Franke, T., & Krems, J. F. (2013). Interacting with limited mobility resources: Psychological range levels in electric vehicle use. *Transportation Research Part A: Policy and Practice*, *48*, 109–122.

Galus, M. D., Waraich, R. A., Noembrini, F., Steurs, K., Georges, G., Boulouchos, K., ... Andersson, G. (2012). Integrating power systems, transport systems and vehicle technology for electric mobility impact assessment and efficient control. *Smart Grid, IEEE Transactions on*, *3*(2), 934–949.

Google. (2015). *About google scholar.* Retrieved from `http://scholar.google.com/intl/en/scholar/about.html`

Hanselka, H., & Jöckel, M. (2010). Elektromobilität—elemente, herausforderungen, potenziale. *Elektromobilität*, 21–38.

Higgins, J. P., Green, S., et al. (2008). *Cochrane handbook for systematic reviews of interventions* (Vol. 5). Wiley Online Library.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*, 2004.

Kranton, R. E. (1991). Transport and the mobility needs of the urban poor. *Infrastructure and Urban Report No. INU*, *86*.

Leitinger, C., & Litzlbauer, M. (2011). Netzintegration und ladestrategien der elektromobilität. *e & i Elektrotechnik und Informationstechnik*, *128*(1-2), 10–15.

Leurent, F., & Windisch, E. (2011). Triggering the development of electric mobility: a review of public policies. *European Transport Research Review*, *3*(4), 221–235.

Link, J. (2012). Elektromobilität und erneuerbare energien: Lokal optimierter einsatz von netzgekoppelten fahrzeugen.

Maia, R., Silva, M., Araújo, R., & Nunes, U. (2011). Electric vehicle simulator for energy consumption studies in electric mobility systems. In *Integrated and sustainable transportation system (fists), 2011 ieee forum on* (pp. 227–232).

Martiny, N., & Schwab, A. (2011). E-mobility for tropical megacities–the tum create centre for electro mobility. *Editorial Team*, 16.

Pehnt, M., Höpfner, U., & Merten, F. (2007). *Elektromobilität und erneuerbare energien*. Wuppertal Inst. für Klima, Umwelt, Energie GmbH.

Rua, D., Issicaba, D., Soares, F. J., Almeida, P. M. R., Rei, R. J., & Peas Lopes, J. (2010). Advanced metering infrastructure functionalities for electric mobility. In *Innovative smart grid technologies conference europe (isgt europe), 2010 ieee pes* (pp. 1–7).

Schiavo, L. L., Delfanti, M., Fumagalli, E., & Olivieri, V. (2013). Changing the regulation for regulating the change: Innovation-driven regulatory developments for smart grids, smart metering and e-mobility in italy. *Energy policy*, *57*, 506–517.

Schlick, T., Hertel, G., Hagemann, B., Maiser, E., & Kramer, M. (2011). Zukunftsfeld elektromobilität. *Chancen und Herausforderungen für den deutschen Maschinenund Anlagenbau. Roland Berger Strategy Consultants, Düsseldorf, Hamburg, Frankfurt*.

Schwedes, O., Kettner, S., & Tiedtke, B. (2013). E-mobility in germany: White hope for a sustainable development or fig leaf for particular interests? *Environmental Science & Policy*, *30*, 72–80.

Stamp, A., Lang, D. J., & Wäger, P. A. (2012). Environmental impacts of a transition toward e-mobility: the present and future role of lithium carbonate production. *Journal of Cleaner Production*, *23*(1), 104–112.

Tremmel, J. (2004). " nachhaltigkeit"–definiert nach einem kriteriengebundenen verfahren. *GAIA-Ecological Perspectives for Science and Society*, *13*(1), 26–34.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *Management Information Systems Quarterly*, *26*(2), 3.

Yay, M. (2010). *Elektromobilität: theoretische grundlagen, herausforderungen sowie chancen und risiken der elektromobilität, diskutiert an den umsetzungsmöglichkeiten in die praxis*. Peter Lang.

Zanker, C., Lay, G., & Stahlecker, T. (2011). *Automobilzulieferer in baden-württemberg unter strom? perspektiven der automobilzulieferindustrie für den übergang zur elektromobilität* (Tech. Rep.). Mitteilungen aus der ISI-Erhebung zur Modernisierung der Produktion.

Discrete time analysis of automated storage and retrieval systems

Dipl.-Ing. Martin Epp, martin.epp@kit.edu, Karlsruhe Institute of Technology - Institute for Material Handling and Logistics

Dr.-Ing. Eda Özden, eda.oezden@kit.edu, Karlsruhe Institute of Technology - Institute for Material Handling and Logistics

M.Sc. Benedikt Fuß, benedikt.fuss@kit.edu, Karlsruhe Institute of Technology - Institute for Material Handling and Logistics

M.Sc. Jiaqi Chen, chen.jiaqi@outlook.com, Karlsruhe Institute of Technology - Institute for Material Handling and Logistics

Prof. Dr.-Ing. Kai Furmans, kai.furmans@kit.edu, Karlsruhe Institute of Technology - Institute for Material Handling and Logistics

In this paper, we present a method for the performance evaluation of automated storage and retrieval systems. For this purpose, a discrete time queueing approach is applied. This approach allows the computation of the complete probability distributions of key performance measures such as the transaction cycle time, thus helping practitioners to determine efficient system configurations and control policies during the design phase of these systems.

1 Introduction and problem description

Automated storage and retrieval systems (AS/RSs) are central elements of many logistical systems. They usually consist of storage and retrieval machines (SRMs) automatically serving the racks by running through aisles between the racks (see figure 1). During the design phase of a system containing an AS/RS it is important to assure that the system configuration is capable of fulfilling the operational requirements in an efficient way. These requirements are mostly storage capacity and throughput. Another important requirement is the service level which is usually defined as the possibility to fulfill a storage or a retrieval transaction within a given time span. The method presented in this paper can be used to determine the service level of a given AS/RS configuration. The design aspects of AS/RSs such as velocity, acceleration/deceleration rates and vertical/horizontal dimensions also affect the energy efficiency of these systems. Being used in combination with an energy model, the method presented in this paper constitutes assistance for system designers.

Research on automated storage and retrieval systems using SRMs has been done since the 1960s. Zschau (1964) and Schaab (1968) were the first to calculate mean service times for systems with random storage policies. Afterwards, Gudehus (1972) developed the first approach for the exact computation of cycle times of AS/RSs with single and dual command cycles taking into account the shape factor of the rack. Graves et al. (1977) and Hausman et al. (1976) published first models that focused on alternative storage assignment policies in the 1970s. Bozer and White (1984) developed models for single and dual command cycles regarding the rack as a continuous surface with normed and scaled coordinates. Based on these models numerous models have been developed since then. They extend the basic models by studying different control policies, configurations of AS/RSs and/or operational characteristics. For further information about these models the reader is referred to the survey of Roodbergen and Vis (2009).

As shown in the literature overview there is no paper dealing with the calculation of the complete probability distributions of AS/RS performance measures. Existing models are either dealing with the computation of mean cycle times or the calculation of characteristic performance measures only on the basis of means and variances. The calculation of the quantiles of performance measures such as the probability to perform 99% of the retrieval transactions within a given time span is not possible. Thus, we propose a discrete time queueing system approach to model automated storage and retrieval systems. In contrast to continuous time calculation methods (e.g. classical queuing models) all input and output variables are described with discrete probability distributions. Events occur only at discrete moments which are multiples of a constant time increment t_{inc}. Thereby the discretization of the time is not a limitation. On the contrary, operational times in material handling systems very often exist in form of discrete values, such as in the case of a storage and retrieval machine. As the level of detail is increased significantly, the discretization is an advantage for the determination of key figures (Schleyer 2007). Due to the advantages offered by a discrete time approach, numerous models for the basic elements branching, merging (Furmans 2004), single server station (Grassmann and Jain 1989), and parallel processing stations (Matzka 2011) were developed. In addition, collecting processes and handling of batches were extensively analyzed (Schleyer 2007) (Özden 2011). Using the existing methods we are able to model an automated storage and retrieval system. Thus, we obtain the complete probability distributions of the performance measures.

The paper is organized as follows. In section 2, a system description and the modeling approach as well as the calculation of the performance measures are presented. Afterwards, a numerical study compares the results of the analytical model with a discrete event simulation in continuous time (see section 3). Finally, section 4 summarizes the paper and provides directions for future research.

2 System description and modeling approach

The investigated AS/RS consists of a single aisle served by one SRM. The number of racks in horizontal and vertical direction is variable. Furthermore, the buffer in front of the I/O point for incoming transactions is as-

sumed to be infinite. The SRM has one load handling device that can hold one unit load. The racks are equally sized, single deep and can store one unit load.

Figure 1 depicts the AS/RS. The SRM travels with a maximum constant velocity v_x and acceleration/deceleration rate a_x/b_x in horizontal direction as well as v_z and a_z/b_z in vertical direction. The racks are indexed (i, j) with i = 0, ..., n_x-1 and j = 0, ..., n_z-1 where n_x is the number of racks in horizontal direction and n_z the number of racks in vertical direction. The width and height of a rack are given by w and h, respectively.

Regarding the control policies random, class-based[1] and turnover-based[2] storage assignment rules as well as a FIFO transaction serving rule can be applied. The system can be operated by single command or dual command cycles. Therefore, the percentage of dual command cycles has to be specified. It is assumed that the number of storage transactions equals the number of retrieval transactions. This is also the prerequisite for the steady state, thus the assumption makes sense for real systems. Moreover, the model assumes a return to I/O dwell point policy.

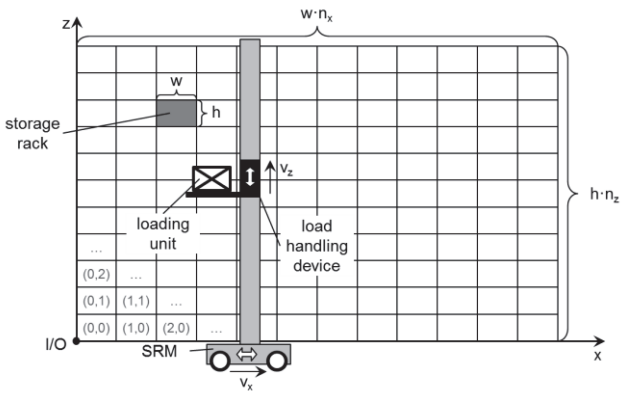


Figure 1: aisle of an AS/RS

The SRM is modeled as a discrete time G|G|1 queue. The generally distributed arrival stream consists of both single command and dual command cycles. They are assumed to be customers of a single class, thus they are waiting in the same queue before they are served by the SRM. The service time distribution is determined by the system configuration (number of racks, velocity of SRM etc.) and the control policies. Therefore, the modeling approach is based on existing methods to calculate performance measures of a G|G|1 service station (e.g. the distributions of the number of customers at the arrival instant, the waiting time and the interdeparture time (Furmans and Zillus 1996) (Grassmann and Jain 1989) (Jain and Grassmann 1988)). In the following, the steps to determine the performance measures are explained in detail.

- Step 1: calculation of the service time distribution of single and dual command cycles

[1] Class-based storage assignment: each product is assigned to a specific area of the rack system. Within the area a random storage assignment rule is applied.
[2] Turnover-based storage assignment: the storage locations are determined based on their demand frequency. Frequently requested products are stored at the fastest accessible locations.

The cycle time of a single command consists of the dead time t_0, the travelling times from the I/O point to the rack and back as well as the operating times of the load handling device (LHD) at the I/O point and at the rack. We obtain the cycle time t_{ij} for each rack (i, j) by building the sum of these times.

$$t_{ij} = t_0 + 2 \cdot t_{travel,ij} + 2 \cdot t_{LHD}$$

The probability p_{ij} of travelling to rack (i, j) is based on the storage assignment rule. For example, using a random storage assignment rule results in a probability of $p_{ij} = \frac{1}{n_x \cdot n_z}$ for each rack.

Finally, the distribution of the service time $\vec{S}_{SC}$ is calculated as follows, where $S_{SC,k}$ denotes the probability of a service time of k time increments.

$$S_{SC,k} = \sum_{i=0}^{n_x-1} \sum_{j=0}^{n_z-1} p_{ij} \quad \forall \; t_{ij} = k$$

For example, if we assume that there are 3 out of 525 racks in the aisle ($n_x = 35$, $n_z = 15$) which we can serve in $10 \cdot t_{inc}$ under a random storage assignment rule, the probability for a service time of of 10 time increments will be calculated as follows.

$$S_{SC,10} = 3 \cdot \frac{1}{35 \cdot 15} = 0.00571$$

The service time distribution of dual command cycles $\vec{S}_{DC}$ is calculated analogously.

- Step 2: calculation of the overall service time distribution

Since the AS/RS can be operated by single command and dual command cycles, the service time distributions of both types of cycle have to be weighted. Given the percentage of dual command cycles p_{DC} and single command cycles $p_{SC} = 1 - p_{DC}$, the overall service time distribution can be computed.

$$\vec{S} = p_{SC} \cdot \vec{S}_{SC} + p_{DC} \cdot \vec{S}_{DC}$$

- Step 3: calculation of the waiting and transaction time distributions

The calculated service time distribution is used as input for the G|G|1 service station. The waiting time distribution $\overrightarrow{W}$ (which is equal for all transaction types) is computed using the algorithm from Grassman and Jain (1989). The transaction time distributions $\vec{T}$ can be obtained by the convolution of the waiting time distribution and the service time distribution.

$$\vec{T} = \vec{S} \otimes \overrightarrow{W}$$

- Step 4: calculation of the distribution of storage transactions waiting at the arrival instant

In practice, the buffer area in front of an AS/RS is usually dimensioned based on a chosen quantile (e.g. 95%) of the number of waiting storage transactions. To compute this measure, we first use a G|G|1 model to calculate the distribution of the number of customers in the system at the arrival instant $\vec{N}$ (Furmans and Zillus 1996). The number of waiting customers $\vec{Q}$ is derived from this distribution, where Q_k denotes the probability of k waiting customers.

$$Q_k = \begin{cases} N_0 + N_1 & if\ k = 0 \\ N_{k+1} & else \end{cases}$$

Subsequently, we use the binomial distribution and the law of total probability to calculate the distribution of the number of waiting storage transactions at the arrival instant $\vec{Q}_s$. For example, if there is just one customer waiting, the probability for a waiting storage transaction is equal to the sum of the probability for a dual command cycle and the probability for a single command storage transaction.

$$Q_{S,k} = \sum_{i=k}^{Q_{max}} Q_i \cdot \binom{i}{k} \cdot (p_{DC} + 0.5 \cdot p_{SC})^k \cdot (0.5 \cdot p_{SC})^{i-k}$$

- Step 5: calculation of the retrieval interdeparture time distribution

To compute the retrieval interdeparture time distribution, the model of (Furmans 2004) is used. This model distributes a stream of arriving customers stochastically to several directions (a probability has to be specified for each direction). We use the interdeparture time distribution of the G|G|1 service station as model input. Storages and retrievals are modeled as customers. However, just the retrievals leave the system physically and correspond to the physical departure process. That's why we assume a split into two directions. The probability of the first direction equals the percentage of retrieval transactions $p_R = p_{DC} + 0.5 \cdot p_{SC}$, the probability of the second direction equals $1 - p_R$. The retrieval interdeparture time distribution of the AS/RS then is given by the interdeparture time distribution of the first direction.

3 Numerical study

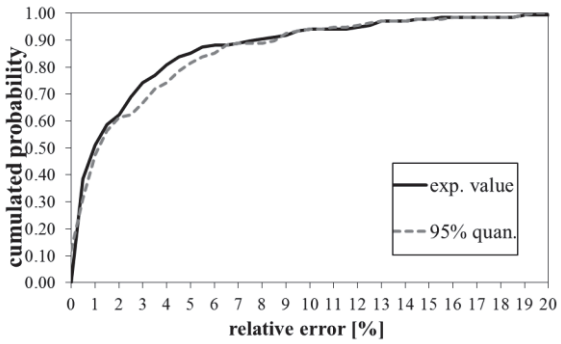
In this section, a numerical study demonstrates the accuracy of the model by comparison with a discrete event simulation in continuous time. The approximation quality is measured by the deviation of the analytical model to the simulation for the mean and 95% quantile of the following performance measures:

- Q_S: number of waiting storage transactions at the arrival instant
- T_R: retrieval transaction time

The following table depicts the tested parameter configurations chosen to reflect a broad variety of different settings.

Parameter	Values
Number of racks in horizontal direction	$n_x = 35$
Number of racks in vertical direction	$n_z = 15$
Horizontal distance between two racks	w = 0.6 m
Vertical distance between two racks	h = 0.8 m
Operating time load handling device	$t_{LHD} = 4$ s
Dead time	$t_0 = 2.3$ s
Maximum horizontal velocity	$v_x = 4.0$ m/s
Maximum vertical velocity	$v_z = 2.0$ m/s
Maximum horizontal acceleration/deceleration	$a_x = b_x = 2.0$ m/s^2
Maximum vertical acceleration/deceleration	$a_z = b_z = 1.5$ m/s^2
Storage assignment rules	Random, Class-based, Turnover-based
Arrival stream distribution types	Uniform, Discrete random values, Exponential
Utilization rates	$\rho \in \{0.75, 0.85, 0.90, 0.95, 0.97\}$
Probability single command cycles	$p_{SC} \in \{0.10, 0.50, 0.90\}$
Time increment	$t_{inc} = 0.2$s

In general, the discrete time approximation delivers a high approximation quality. Figures 2 and 3 show the plots for the cumulated distributions of the relative and absolute errors of the discrete time approximation for T_R and Q_S, respectively. The average relative errors for the expected value and the 95% quantile of T_R are 2.55% and 2.84%, respectively. The average absulute errors for the expected value and the 95% quantile of Q_S are 0.19 and 0.53 storage units, respectively. An additional analysis reveals that the highest deviations occur in the test cases with high utilization rates ($\rho \geq 0.95$).



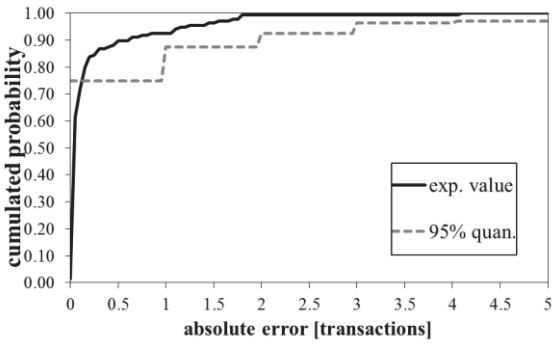


Figure 2 (left side): cumulated distribution of the relative errors for T_R

Figure 3 (right side): cumulated distribution of the absolute errors for Q_S

4 Conclusion

In this paper we presented a method for the calculation of AS/RS performance measures. By modeling the system as a discrete time queueing system it is possible to determine the complete probability distributions of the retrieval transaction time, the storage transactions waiting at the arrival instant and the retrieval interdeparture time. A numerical study demonstrated the accuracy of the model in comparison to a discrete event simulation. The study revealed that high utilization rates, which should be avoided when designing an AS/RS, lead to the largest deviations. The deviations for utilization rates that are more relevant in practice were rather small. In general, the discrete time approximation deliveres a high approximation quality. Therefore, the method can be used to help practitioners determining efficient system configurations and control policies during the design phase of intralogistics systems. Further research is needed regarding alternative system configurations and control policies such as varying I/O locations and double deep racks. Additionally, an energy model could be added to help designers planning energy efficient AS/RSs that fulfill the operational requirements.

5 Acknowledgment

This research is supported by the research project “Analytical computation of sojourn time distributions in large-scale conveyor systems”, funded by the Deutsche Forschungsgemeinschaft (DFG) (reference number FU 273/12-1).

References

Bozer, Y. A. and J. A. White. Travel-time models for automated storage/retrieval systems. IIE Transactions 16 (4), p. 329–338, 1984.

Furmans, K. A framework of stochastic finite elements for models of material handling systems. In: 8th International Material Handling Research Colloquium, Graz 2004.

Furmans, K. and A. Zillus. Modeling independent production buffers in discrete time queueing networks. In: Proceedings of CIMAT ’96, Grenoble, p. 275–280, 1996.

Graves, S. C., W. H. Hausman, and L. B. Schwarz. Storage retrieval interleaving in automatic warehousing systems. Management Science 23 (9), p. 935–945, 1977.

Grassmann, W. K. and J. L. Jain. Numerical solutions of the waiting time distribution and idle time distribution of the arithmetic GI/G/1 queue. Operations Research 37 (1), p. 141–150, 1989.

Gudehus, T. Grundlagen der Spielzeitberechnung für automatisierte Hochregallager. Deutsche Hebe- und Fördertechnik 18 (64), p. 63–68, 1972.

Hausman, W. H., L. B. Schwarz, and S. C. Graves. Optimal storage assignment in automatic warehousing systems. Management Science 22 (6), p. 629–638, 1976.

Jain, J. L. and W. K. Grassmann. Numerical solution for the departure process from the GI/G/1 queue. Computers & OR 15 (3), p. 293–296, 1988.

Matzka, J. Discrete Time Analysis of Multi-Server Queueing Systems in Material Handling and Service. Dissertation, Karlsruhe Institute of Technology, 2011.

Özden, E. Discrete time Analysis of Consolidated Transport Processes. Dissertation, Karlsruhe Institute of Technology, 2011.

Roodbergen, K. J. and I. F. Vis. A survey of literature on automated storage and retrieval systems. European Journal of Operational Research 194, p. 343–362, 2009.

Schaab, W. Technisch-wirtschaftliche Studie über die optimalen Abmessungen automatischer Hochregallager unter besonderer Berücksichtigung der Regalförderzeuge. Dissertation, Technische Universität Berlin, 1968.

Schleyer, M. Discrete time analysis of batch processes in material flow systems. Dissertation, Universität Karlsruhe, 2007.

Zschau, U. Technisch-wirtschaftliche Studie über die Anwendbarkeit von Stapelkranen im Lagerbetrieb. Dissertation, Technische Universität Berlin, 1964.

A Note on Consumer Flexibility in Car-sharing

Philipp Ströhle, philipp.stroehle@kit.edu, KIT
Johannes Gärttner, gaerttner@fzi.de, FZI

Growing importance of *usage instead of ownership* enables servicification of larger parts of the economy and gives rise to economic coordination challenges. The mobility sector is expected to be one of these affected parts and thus will traverse through significant changes in the near future. Due to the reduced cost of multimodal mobility, it may gain notable shares of so-far purely individual mobility. Hence, car-sharing may play the central role of a system enabler in multimodal mobility systems. In order to improve its economics, flexibility provision by consumers will become increasingly important. Therefore, we propose the study of consumer flexibility in car-sharing. So far, research on flexibility has mostly been confined to industrial, production settings. However, the integration of consumer (or demand-side) flexibility may foster more efficient capacity utilization, improved service provision, and thus facilitate the overall proliferation of the concept. Irrespective of the technological shortcomings of electric vehicles, electrification has received significant attention in recent years. Based on appropriate incentives with the goal of revealing consumers' flexibility potentials, the latter may be leveraged to match uncertain demand more precisely with multi-technology fleets in car-sharing systems, as well as assist in overcoming the current technological limitations of electric vehicles. This paper outlines some approaches to incentive design with the goal of harnessing consumer flexibility in car-sharing.

1 Introduction

Future mobility systems will make use of reduced transaction costs and allow consumers to select appropriate means of transportation on a regular, possibly per-trip, basis. With information about travel modes becoming increasingly accessible, multi-modal travel will gain importance (Kuhnimhof et al., 2012). In this vein, electrification promises sustainable individual mobility, especially if the electrical energy required is generated from renewable energy sources (wind, solar, etc.). However, electric vehicles generally are limited in driving range and require time-consuming recharging. A large share of car-sharing reservations features only short distances and thus may not be adversely affected by limited range. Accordingly, car-sharing poses one of the economically most interesting use-cases for electric mobility. High utilization ratios of shared vehicle fleets partially offset higher initial capital expenditures for electric vehicles.

In multi-modal mobility systems, efficient capacity utilization requires coordination of fluctuating demand with fixed supply. Through approaches incorporating both algorithmics and economics, demand heterogeneity can explicitely be taken into account. Then, mobility demand can appropriately be assigned to conventional and electric parts of the fleet. As a result, the overwhelming part of aggregate mobility demand may be served, while both operating expenses and emissions are reduced.

However, these desired results hinge critically on successfully harnessing consumer flexibility in station-based car-sharing. Up to this point, the dimensions comprising consumer flexibility and each user's flexibility endowments are not well understood (Kuhnimhof et al., 2006). To further the understanding of consumer flexibility in car-sharing, we propose mining empirical data in order to identify and quantify consumer flexibility. A detailed understanding of flexibility is necessary as input to sophisticated (static or dynamic) assignment schemes. Goals may comprise achieving "good" trade-offs between high fleet utilization and high service satisfaction levels among consumers.

2 Related Work

Under individual ownership of (electric) vehicles, each vehicle must satisfy highly heterogeneous mobility demands, comprised of both short and long distance trips. Due to their limited range, electric vehicles are effectively sidelined as the process of choosing a vehicle technology is dominated by few long distance trips. Moreover, EV adoption suffers from (anticipated) *range anxiety* (Eberle & von Helmolt, 2010), the fear of running out of energy to complete a trip. However, EV charging can be controlled to fulfill a plethora of goals that may compensate for the drawbacks of electric mobility. The literature provides rich examples, such as cost minimization under variable electricity prices through appropriate timing of the charging decisions, maximizing sustainability (Schuller et al., 2014), or maximizing the consumption of self-generated electricity.

The question of optimally assigning reservations to vehicles yields the well-known bin-packing problem (Nemhauser & Wolsey, 1988; Dyckhoff, 1990). Via the use of consumer flexibility, more efficient assignment decisions become possible, as all items are assigned using a reduced number of bins. This is a popular problem, found in numerous other domains, e.g., assignment of virtual machines to physical hosts. For the offline bin-packing problem, a simple *next fit* algorithm yields a competitive ratio of two. In practical applications, however, this lower bound can often be beat by the use of a model of future arrivals. Setzer & Bichler (2013), for example, propose the use of dimensionality reduction techniques for an online bin-packing where the workloads exhibit seasonalities. Through succinct representation of the corresponding optimization problem, larger instances may be solved, while the simpler problem representation barely affects solution quality.

In the literature on Revenue Management, Choudhary et al. (2005) report ambiguous effects of personalized pricing on profits in a two-firm setting. If the cost of providing higher quality is sufficiently convex, both firms may be better off by not introducing personalized pricing, and avoiding the cost of higher quality provision. In car-sharing, higher quality provision may exhibit highly convex costs structures, effectively prohibiting higher quality for consumers. However, through the use of consumer flexibility, provision of high-quality service may be possible at lower cost, effectively reducing the convexity of the cost function. As a consequence, firms may find themselves in a prisoner's dilemma in competitive settings. For the monopolist's case, and refraining from a long-term perspective, we believe that the use of consumer flexibility is a dominant strategy.

In the marketing literature, the seminal work of Cronin & Taylor (1992) highlights the difference between service quality and consumer satisfaction, with the latter having a larger influence on purchase intentions. Following this result, demand differentiation and segmentation in car-sharing may enhance customer satisfaction and thus provide positive feedback on further uptake of the concept.

3 Quality-Tiering Access to Car-sharing systems

Capacity utilization and consumer service in car-sharing systems may be improved by appropriately substituting consumer flexibility for operator capacity. Following Jordan & Graves (1995), adding even small amounts of flexibility can assist in coping with demand uncertainty, potentially leading to significant improvements regarding economic outcomes. Adding system flexibility on the operator's part, however, is not straight-forward, as it may involve vehicles' re-positioning in order to respond to shifting demand patterns. Consumers, on the other side, may exhibit a certain degree of flexibility which becomes apparent in their usage choices regarding time, location, and type of vehicle. This kind of flexibility may be employed to improve operational objectives, e.g., fleet utilization. Alternatively, optimization goals may be formulated following consumers' perspective, e.g., better fleet availability or other forms of higher service quality.

Overall, consumer flexibility may provide an important lever for the electrification of car-sharing fleets. Ceteris paribus, the denser the station network becomes, the smaller the amount of spatial flexibility required to achieve efficient outcomes.

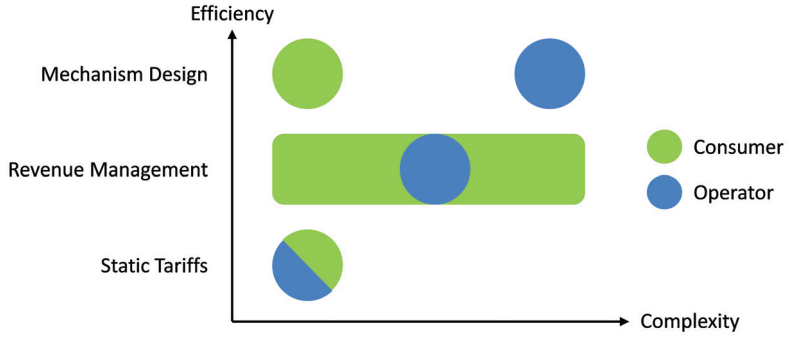


Figure 1: Efficiency-complexity trade-off for the proposed coordination approaches

By assigning "fitting" reservations to electric vehicles, the latter are utilized more intensively, faster recouping higher initial investments. Clearly, dispersed flexibility potentials are initially unknown, requiring the design of appropriate incentives, for example in the form of variable tariffs, to reveal consumer flexibility. Under economic mechanisms that make revelation of private information – regarding spatial and/or temporal flexibility as well as expected driving range – a good strategy for the user, car-sharing may grow beyond a niche application and render individual mobility more sustainable.

Through appropriate statistical methods, already present flexibility in space, time, and vehicle type may be identified from empirical consumer decision data. In more detail, at each user interaction a specific choice set is available, which may be obtained from user interactions with the reservation system. The state of the system (availability and stock-outs) can then be identified from a stream of reservation data given the fleet.

In current systems, the user may leverage different kinds of his flexibility to meet his respective mobility demands, contingent on system state. To identify the corresponding type and amount of flexibility, empirical reservation streams, of which we have access to, provide a rich source of censored data that allows detailed exploration and modeling of consumer choice. In future reservation systems, the task of leveraging different kinds of users' flexibility, contingent on the state of the system, may be addressed more efficiently by the system operator.

The state of the shared fleet can, to some extent, be re-constructed from reservation data, available for billing purposes, if the time of reservation, as well as beginning and end date of each reservation (and possibly the distance) are available. Note that non-observations, i.e., intended, but unrealized reservations are not included in the data (Talluri & Van Ryzin, 2005, cf. p.474). However, such information could potentially be derived from logging user interactions with the reservations system, but is out of the scope of this work.[1] In more detail, during times of scarcity, i.e., heavy fleet utilization, consumers deviate to some extent from their initially desired vehicle, either in time, space, or type. Given empirical customer choice, consumer flexibility clusters may be identified and appropriate tariffs provided. Fig. 1 illustrates the efficiency-complexity trade-off involved when selecting a tariff type.

Static Tariff Under a static tariff, the operator may have the right (option) to postpone or geographically re-assign the corresponding reservation to another vehicle. The goal of introducing such tariff structures lies in trading-off fleet utilization and consumer inconvenience by explicitly employing available flexibility. The advantage of a static flexibility tariff lies in its low complexity (see Fig. 1), requiring consumers to choose a certain tariff bracket only once (or infrequently). This low complexity, however, induces the lack of adaptability to a user's specific situation. For example, circumstances may reduce a user's flexibility. Demanding excessive flexibility in such situations may significantly affect user acceptance

[1] The lack of such observations, so called censored data, is a common problem in retailing and service systems.

of static pricing schemes. Nevertheless, as Cohen et al. (2014) indicate, static pricing decisions may already yield highly competitive outcomes compared to more dynamic mechanisms that introduce additional complexity.

Revenue Management Dynamic tariffs feature varying prices, based on two distinct components. First, prices may vary over time of day, day of week, or even season of the year, similar to electricity tariffs. Second, prices may reflect scarcity it space. Given those varying prices, consumers select their utility maximizing from a menu of options. Clearly, dynamic tariffs offer more specific choices and allow consumers to incorporate their corresponding, situative flexibility level into the decision making process. However, these are more complex to implement compared to static tariffs.
Besides price-based control, the operator's capacity could (partially) be reserved for high-valued consumers that reveal their demand only shortly before beginning a trip, following the literature on revenue management. Key to its successful application through appropriate model parametrization is detailed knowledge on future demand arrival and in particular its valuation. By means of such detailed knowledge, the risk of poor decision making on behalf of the operator can presumably be reduced massively. Modeling car-sharing operations via a newsvendor model, or, more realistically, via multiple, correlated newsvendor models, may give rise to detailed insights into car-sharing operations and – in space, time, and vehicle type – heterogeneous reservation classes.
Furthermore, overbooking, either dynamically or statically (Talluri & Van Ryzin, 2005), may provide an additional lever to improve the system's efficiency. In contrast to airline operations, the cost of denied service in car-sharing may be relatively low, rendering overbooking a realistic addition to the car-sharing revenue management toolbox.[2]
Besides heterogeneous flexibility endowments, demand may be heterogeneous with respect to the notification period prior to the reservation's begin. Earlier notification (under systems relying on consumers' flexibility revelation) reduces the amount of flexibility available to the operator to adapt its schedule, but clearly is more favorable for consumers. Accordingly, lead-time differentiation may provide yet another lever to both assign reservations more efficiently and skim consumers' willingness to pay.

Mechanism Design Alternatively, following a mechanism design approach, consumers may reveal their flexibility endowment with respect to space, time and vehicle type at each interaction instead of choosing a specific alternative directly. The mechanism then computes allocations and corresponding payments, taking competing demand into account. As it can build on a greater information set, better trade-offs involving consumer inconvenience in the form of spatial or temporal deferral and operator objectives can be found. Flexibility revelation may comprise the following examples:

- Commit to a particular vehicle at a specific time and place.
- Allow postponement on the same vehicle class.
- Allow the use of the same vehicle type at a different station.

The main drawback of such an approach relates to more exhaustive information revelation by consumers. Therefore, it may be met with caution by those preferring data-sparse mechanisms.

4 Conclusion and Outlook

Car-sharing is expected to contribute to the servicification of the mobility sector, providing a richer choice menu and enabling comprehensive, sustainable, multi-modal transportation offers. Accordingly, the economic and social importance

[2] While the next flight may only be available hours or days from now, the next vehicle, given a sufficiently dense network of vehicles (stations), may be available with little offset/inconvenience to the consumer.

of car-sharing may be further increasing, calling for both methodologically sound and interdisciplinary research at the intersection of economics, computer science, and information systems.

We envision a number of interesting research avenues regarding the operations of car-sharing, as it becomes part of mainstream consumption. For one, a better understanding of customer choice is necessary. To this end, statistical modeling poses an important and challenging milestone. Spatio-temporal models may be the appropriate means to model the demand process and associated uncertainty in time and space. However, prior to building these models, a principled method to cope with erroneous and censored data must be applied. To this end, methods from statistics and machine learning will be valuable. Eventually, building on a solid to-be-developed understanding of consumer choice in car-sharing, appropriate incentive schemes may be designed. Here, we expect the application and extension of methods from both, revenue management (Talluri & Van Ryzin, 2005) and online mechanism design (Porter, 2004; Gerding et al., 2011) to the domain of car-sharing in order to tap into available flexibility endowments.

References

Choudhary, V., Ghose, A., Mukhopadhyay, T., & Rajan, U. (2005). Personalized pricing and quality differentiation. *Management Science*, *51*(7), 1120–1130.

Cohen, I., Eden, A., Fiat, A., & Jez, L. (2014). Pricing online decisions: Beyond auctions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on discrete algorithms* (pp. 73–91). Society for Industrial and Applied Mathematics.

Cronin, J. J., Jr., & Taylor, S. A. (1992). Measuring service quality: A reexamination and extension. *Journal of Marketing*, *56*(3), 55–68.

Dyckhoff, H. (1990). A typology of cutting and packing problems. *European Journal of Operational Research*, *44*(2), 145–159.

Eberle, U., & von Helmolt, R. (2010). Sustainable transportation based on electric vehicle concepts: a brief overview. *Energy & Environmental Science*, *3*(6), 689–699.

Gerding, E., Robu, V., Stein, S., Parkes, D., Rogers, A., & Jennings, N. (2011). Online mechanism design for electric vehicle charging. In *The tenth international joint conference on autonomous agents and multi-agent systems (AAMAS 2011)* (pp. 811–818).

Jordan, W. C., & Graves, S. C. (1995). Principles on the benefits of manufacturing process flexibility. *Management Science*, *41*(4), 577–594.

Kuhnimhof, T., Buehler, R., Wirtz, M., & Kalinowska, D. (2012). Travel trends among young adults in germany: increasing multimodality and declining car use for men. *Journal of Transport Geography*, *24*, 443–450.

Kuhnimhof, T., Chlond, B., & von der Ruhren, S. (2006). Users of transport modes and multimodal travel behavior steps toward understanding travelers' options and choices. *Transportation Research Record: Journal of the Transportation Research Board*, *1985*(1), 40–48.

Nemhauser, G. L., & Wolsey, L. A. (1988). *Integer and combinatorial optimization* (Vol. 18). Wiley New York.

Porter, R. (2004). Mechanism design for online real-time scheduling. In *Proceedings of the 5th ACM conference on electronic commerce* (pp. 61–70). New York, NY, USA: ACM.

Schuller, A., Dietz, B., Flath, C., & Weinhardt, C. (2014). Charging strategies for battery electric vehicles: Economic benchmark and V2G potential. *IEEE Transactions on Power Systems*, *29*(5), 2014–2022.

Setzer, T., & Bichler, M. (2013). Using matrix approximation for high-dimensional discrete optimization problems: Server consolidation based on cyclic time-series data. *European Journal of Operational Research*, *227*(1), 62–75.

Talluri, K. T., & Van Ryzin, G. J. (2005). *The theory and practice of revenue management* (Vol. 68). Springer.

Evaluating services in mobility markets: A business model approach

Christopher Lisson, christopher.lisson@kit.edu, Karlsruhe Service Research Institute

Wibke Michalk, wibke.michalk@bmw.de, BMW AG

Roland Görlitz, roland.goerlitz@kit.edu, Karlsruhe Service Research Institute

Developments in ICT lead to a wide range of new services in the dynamic mobility market. Especially emerging web-based mobility services (WBMS) like Uber or Moovel have a disruptive potential for satisfying customer's mobility needs. The absence of an established classification scheme is an obstacle in comparing these new services and thus in investigating their critical success factors. This paper applies a business model approach to generate such a classification framework for WBMS and thus closes the research gap. The concept exhibits promising features in structuring a market of WBMS and facilitates first comparative insights. It can be used for mapping current market structure in the area of mobility services and investigating critical success factors in each WBMS class.

1 Changing business model in mobility markets

Value generation by means of fast Internet connections and coordination of information is disrupting business models in many industries by restructuring their critical processes – e.g. online banking in banking and e-commerce in retail (LaValle et al. (2013) , Kagermann (2014)). This holds for passenger transportation and related mobility services, too. Advances in information and communication technologies (ICT), e.g. continuous web access via smartphones, facilitate the advent of new services like car- and ride-sharing. Additionally, new incumbents on the mobility market like Google[1] force established players to adjust (Stricker et al. 2011).

Success of ride coordination services like Uber[2] and Lyft[3] – in terms of a rapidly growing user base – are perfect examples of this novel development. Both are gaining ground and start to transform taxi markets and public transportation around the globe (Anderson 2014). They enable intelligent resource allocation for ride-sharing in real-time and facilitate booking as well as payment processes in one application. Information Systems for intermodal passenger

[1] www.google.com
[2] www.uber.com
[3] www.lyft.com

transportation like Moovel[4] or Qixxit[5] are currently gaining momentum in the German market. They focus on solving the multidimensional mobility problem to overcome a distance by coordinating different modes of transportation like trains, bicycles and cars while considering user preferences.

As stated above, the variety of the emerging services in the dynamic mobility market is high and comprises a wide range of aspects of individual mobility. To the knowledge of the authors, a taxonomy or established classification scheme for mobility services has not been devised until today. Without such a classification, it is it difficult to entirely compare these services, understand their user acceptance and thus gain insights about their critical success factors. This paper aims at establishing a framework, which enables a classification of WBMS in mobility markets. Furthermore, first insights into the current state of the German market for WBMS are given.

2 A comparison framework for WBMS

2.1 Web-based mobility services

Individual mobility can be seen as the spatial and timely motion of an individual in order to overcome distances – ways with an explicit destination and intended purpose consisting of a sequence of stages (Ammoser & Hoppe, 2006). The complex variety of reasons why individuals need to overcome distances can be summarized under the concept of mobility needs. The interaction of resources and instruments with which the mobility needs can be satisfied are defined as mobility services[6] (Ammoser & Hoppe, 2006). Mobility services in gerneral have already been widely investigated in literature. This paper focusses on the subset of web-based mobility services (WBMS) for passenger transportation, which have been enabled by recent developments in ICT and are defined follows:

Web-based mobility services support the customer to inform about, coordinate and/or realize a spacial movement – including the physical transition – in order to satisfy their individual mobility needs by using web-based technologies.

Based on this definition the corresponding services in the German mobility market are identified by Internet research and expert intetviews. Followig WBMS are considerd in this study: *Moovel, Qixxit, DB Navigator, Allryder, My Mobility Map, Google Maps, FromAtoB, GreenMobility, Waymate, GoEuro, Stuttgart Services, Mein Fernbus, Busliniensuche, Flixbus, Verkehrs- und Tarifverbund Stuttgart (VVS), Kölner Verkehrs-Betriebe (KVB), Hamburger Verkehrsverbund (HVV), Münchner Verkehrsverbund (MVV), Swoodoo, myTaxi, Uber,*

[4] www.moovel.com
[5] www.qixxit.de
[6] These services posses a potential-, a process-, and an outcome-dimension and can be material as well as immaterial. Material instruments are the means of transportation, e.g. cars, ships, airplanes, trucks as well as the infrastructural components, while immaterial instruments can be seen in information and coordination mechanisms, e.g. timetables or routing services.

DriveNow, Multicity, Car2go, Flinkster, Citeecar, Autonetzer, Nachbarschaftsauto, Cambio, Tamyca, DB Call a Bike, Nextbike, Rent-a-bike, Flinc, BlaBlaCar, Mitfahrgelegenheit, Matchrider.

2.2 The business model approach

Absence of an established classification scheme is characteristic in a phase of innovation within a young and dynamic market like the one for ICT-driven WBMS. According to Stähler (2002) such digital services cannot be compared by classic approaches, since they are more of a reconstructing and disruptive nature and posses a higher degree of inherent volatility. However, it is possible to compare these new services based on their business models – the business model approach. Business models describe the rationale of how an organization creates, delivers and captures value. Thereby they provide insights about the critical success factors. In literature, there exist different concepts about how a business model should be described (Scheer et al. 2003). Based on established concepts (Osterwalder & Pigneur (2011), Stähler (2002)) categories for the classification framework have been elaborated, which are capable to structure existing WBMS in the German market: customer segments, customer relationship, channels, cost structure, revenue streams, value proposition, key activities, key resources and key partners.

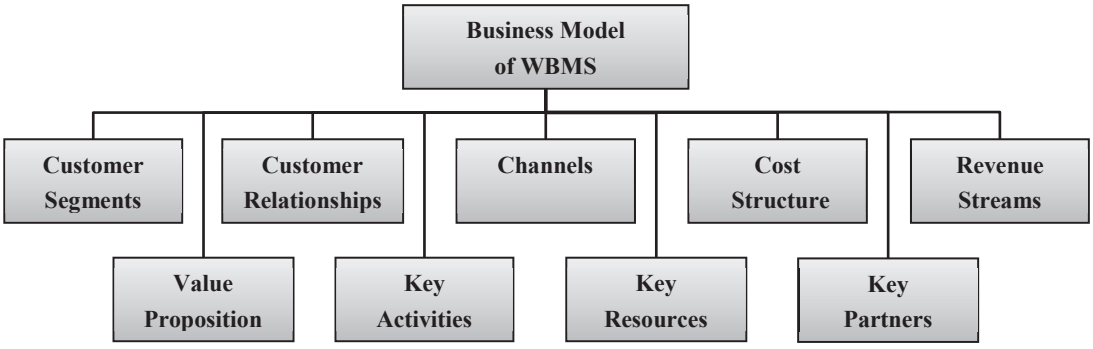


Figure 1: Categories for the comparison framework

While each category covers certain business aspects and consists of a number of attributes, their combination represents the business model of a distinct WBMS, see Figure 1. In order to compare the WBMS their categories' attributes need to be adjusted to meet the requirements and specifications of the application context.

2.3 Context specific categories

In accordance to the classifiaction framework the context specific adjustments of each category's attributes have been elaborated and are illustrated in Table 1. Customer segments help to define different customer groups and gain insights about their distinctive needs. They can distinguish customers by spacial, socio-economic or usage characteristics. The customer relationship aims at establishing an adequate type of relationship between provider and customer, which generates a fit between both sides' interests. The channels determine how the WBMS are offered and marketed, how the customer can access them and which kind of payment options are accepted. The cost structure describes the

WBMS's expenditures and can be taken as an indicator for its financial manoeuvrability, critical activities and strategic options. The revenue streams document the cash flows, which are related to service generation and provision.

Table 1: Categories of the comparison framework and ther context specific attributes.

Category	Attributes & Explanation
Customer segments	− **local:** WBMS is available only in certain regions. − **regional:** WBMS is available only in certain regions. − **global:** WBMS is available throughout the country. − **Sigma milieu:** WBMS is especially used by a certain socio-economic group.
Customer relationship	− **Personal assistant**: Each customer is connected to the service via a personal assistant, which takes care for the customer, allows personalized treatment, and thus the highest information gain in terms of customer insights, e.g. Visa travel assistant. − **Account owned by provider**: The customer is connected to the service via an account owned by the provider and a high level of customer insights is gained – e.g. moovel. − **Account owned by intermediary**: The customer is connected to the service via an account owned by a third party and a low level of customer insights is gained – e.g. swoodoo. − **Anonymous**: The customer is not connected, as far as possible anonymous and can use the service without any preconditions – e.g. Deutsche Bahn navigator. Thereby no individual customer insights can be gained.
Channels	− **Service access**: Online via Webpage, Online via App. − **Payment by**: cash, credit card, debit card, bank transfer, PayPal. − **Payment on**: service provider's page, sub provider's page, stage-dependent.
Cost structure	− **Information:** Costs related to data generation and processing (e.g. network coordination or programming). − **Transportation:** Costs related to transportation activities (e.g. infrastructure generation and maintenance of means of transportation). − **Supplementary:** Costs related to supplementary expenses (e.g. licences, personal, rents, patents).
Revenue Streams	− **(Revenue Model) Free:** All service's functions can be used for free. − **(Revenue Model) Freemium:** A part of the service's function can be used for free, while additional functions are liable to pay costs. − **(Revenue Model) Premium:** All service's functions are liable to pay costs. − **B2C-related:** Leasing, subscription fee, registration fee, and payment for distinct transportation service. − **B2B-related:** Brokerage fee, advertisement, and subsidies.
Value proposition	− **Information related service attributes**: Routing with and without modal changes, evaluation of alternative routes, showing route's characteristics (e.g. length, changes, costs, probability to be on time), showing position, price information and mapping the routes, provision of traffic information (e.g. warning of time delays due to traffic jams) and an delay alarm, provision of in-house navigation, provision of weather information, a customizable user interface (e.g. adjusting functionalities and points of interests). − **Transportation related service attribute**s: Provision of all kinds of mode of transportation, e.g. Bus, Tram, Underground, Train, Long Distance Bus, Plane, Ship, Walk, Bike, Bike sharing, privately owned Car, Car sharing, Ride sharing, Rental Car and Taxi. − **Supplementary related service attributes**: Stage related booking, overall ticketing, learning preference (e.g. routes, modes of transportation, point of interest), education related services, e.g. reading newspapers, insurance related services, e.g. insurance against delays, Social Media related services, e.g. possibility to meet friends on the same way or community blogs, food related services, e.g. possibility to have food on the way, travel related services, e.g. integration of hotel finding and booking, personalization related services, e.g. enabling to choose which services should be offered, entertainment related services, e.g. lottery and events advertisements, customer loyalty programs and feedback options.

Key Activities	− Network integration and coordination. − Data generation and processing. − Operational excellence in information related activities: e.g. User interface design, programming of algorithms, reducing complexity, and enabling service access. − Operational excellence in transportation related activities: e.g. Fleet managements in terms of ensuring punctuality, offering low cost and high quality services, offering many options for means of transportation and routes, and maintenance of infrastructure and means of transportation. − Operational excellence in supplementary related activities: e.g. Personalization of services, adding additional services to satisfy non-primary needs, and payment processes.
Key Resources	− Data generation, processing and provision, e.g. right to use and distribution of customer data. − Possession of the means of transportation, e.g. owning a fleet. − Influence and coordination power in the network, e.g. ability to convince cities to participate in a car-sharing project through the provision of parking slots. − Strong brand and excellent customer standing, e.g. long history of success in other branches that generates mutually trust. − Technical excellence in specific processes, e.g. design of interfaces or reliability of processes. − Human and intellectual property, e.g. patents and developing capabilities. − Financial resources, e.g. large enterprises as strategic investors.
Key Partners	− **Information related** Partners are partners, who help to realize the required information provision, e.g. supplier's traffic-, geo- and weather-data like Google or programmers for the design of routing-algorisms. − **Transportation related** Partners are partners, who help to realize the transportation or the required infrastructure, e.g. taxi companies, Deutsche Bahn or public transportation. − **Supplementary related** Partners are partners, which help to realize the supplementary services, e.g. providing information about points of interest, social services or payment providers.

The value proposition describes all actions and instruments delivered by a provider that enable the customers to satisfy their needs and to solve their problems. In the context of WBMS the offered services are divided into the categories of information, transportation and supplementary related services. While the key activities list all activities that are required to achieve the promised value proposition, the key resources document the compulsory inputs. The key partners are indispensable for the service provision and further help to optimize the process as well as they reduce risk and uncertainty. Based on those categories and attributes existing WBMS in the German market have been investigated and rated. For example, if a service offers car and ride sharing both attributes are marked as active in the classification scheme. By evaluating each attribute's characteristics – given in Table 1 – of a service a picture evoles in the framework, which's insights are discussed in the next chapter.

3 First insights of the German WBMS market

Conducting the business model approach to compare WBMS in the German mobility market revealed that the additional category modality – the portfolio of modes of transportation, which is offered by a WBMS – is a criterion with significant discriminatory power. The modality describes a service's capability to connect different modes of transportation and can be divided into mono-, multi- and intermodal. A WBMS is monomodal when it offers only a single mode of transportation for a trip, while a multimodal WBMS offers a selection of modes of transportation for a trip. An intermodal WBMS further offers different – at least two – modes of transportation for different stages of a trip

(Zumkeller et al. 2005). Up to now, more monomodal than multi- and intermodal mobility services are available in Germany.

Looking at the key resources it is characteristic that monomodal service providers generally own the means of transportation and are responsible for the actual transportation. This determines their major cost structure, i.e. operating and maintaining a fleet, and their revenue model, which is focussed on transportation fees. They do not depend on many partners to provide their service. Their corresponding key activities and resources aim at developing a strong operational excellence. In contrast the multi- and intermodal services focus on coordination aspects and function as an integrator, while combining monomodal services to one solution for the customers' mobility needs. The information gathering, structuring and provisioning processes determine their major cost. The induced revenue model is based on free information for the customers and brokerage fees or advertisement for business partners. They depend on a large network of partners to fulfil their service promise and are therefore focussed on developing a large and stable network. Accordingly, their key competences are organizing the transportation chain and connecting with the customer.

Looking at the key partners it can be observed that due to their network position and monopoly aspects local public transportation services and Deutsche Bahn are key players with a high degree of negotiation power when it comes to intermodal services. Investigating the value proposition the majority of services focuses on the areas of information and transportation, while supplementary related services are only occasionally used. Therefore the full potential regarding services satisfying the non-primary needs is not exploited by now.

4 Discussion and future work

Applying the business model approach to structure existing WBMS in the German market reveals that the discriminatory power of categories as well as attributes needs to be increased. A clearer and more detailed definition of both reduces the probability of misinterpretation during the evaluation process of a WBMS and increases results' quality. Therefore, future research should focus on improving or rather narrowing down the categories to differentiate WBMS. The highly dynamic market makes a regular up-date of the categories' attributes indispensable. This holds especially for the functionalities as a part of the WBMS's value proposition, which are a critical factor for customers' usage intention.

Albeit large potentials for improvement, the proposed framework and the carried out market analysis revealed first insights regarding the German mobility market and existing WBMS. Thus, trying to classify the different services using a business-model-based framework seems promising. It provides the foundation to compare these services regarding nine business relevant categories and thus enables to facilitate insights on a wide range of aspects.

References

Ammoser, H., & Hoppe, M. (2006). *Glossar Verkehrswesen und Verkehrswissenschaften*. Dresden: Institut für Wirtschaft und Verkehr, Technische Universität Dresden.

Anderson, D. N. (2014). "Not just a taxi"? For-profit ridesharing, driver strategies, and VMT. *Transportation*, *41*(5), 1099-1117.

Kagermann, H. et al. (2014). Smart Service Welt. Umsetzungsempfehlungen für das Zukunftsprojekt Internetbasierte Dienste für die Wirtschaft. Berlin: acatech – Deutsche Akademie der Technikwissenschaften.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2013). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, *21*.

Osterwalder, A., & Pigneur, Y. (2011). *Business Model Generation: Ein Handbuch für Visionäre, Spielveränderer und Herausforderer*. Frankfurt: Campus Verlag.

Scheer, C., Deelmann, T., & Loos, P. (2003). *Geschäftsmodelle und internetbasierte Geschäftsmodelle – Begriffsbestimmung und Teilnehmermodell*. Mainz: ISYM – Information Systems & Management, Universität Mainz.

Stähler, P. (2002). *Geschäftsmodelle in der digitalen Ökonomie: Merkmale, Strategien und Auswirkungen* (Vol. 7). Köln: Josef Eul Verlag.

Stricker, V. K., Matthies, G., & Tsang, R. (2011). *Vom Automobilbauer zum Mobilitätsdienstleister*. München: Bain & Company Germany.

Zumkeller, D., Manz, W., Last, J., & Chlond, B. (2005). *Die intermodale Vernetzung von Personenverkehrsmitteln unter Berücksichtigung der Nutzerbedürfnisse (INVERMO)*. Karlsruhe: Institut für Verkehrswesen, Universität Karlsruhe (TH).

Histopathology laboratory operations analysis and improvement

A.G. Leeftink MSc[1], prof. dr. R.J. Boucherie[1], prof. dr. Ir. E.W. Hans[1], M.A.M. Verdaasdonk[2], dr. Ir. I.M.H. Vliegen[1], prof. dr. P.J. Van Diest[2]

[1] Centre for Healthcare Operations Improvement & Research (CHOIR), University of Twente, Enschede, The Netherlands, [2] Department of Pathology, University Medical Center Utrecht, Utrecht

Abstract

Histopathology laboratories aim to deliver high quality diagnoses based on patient tissue samples. Indicators for quality are the accuracy of the diagnoses and the diagnostic turnaround times. However, challenges exist regarding employee workload and turnaround times in the histopathology laboratory. This paper proposes a decomposed planning and scheduling method for the histopathology laboratory using (mixed) integer linear programming ((M)ILP) to improve the spread of workload and reduce the diagnostic turnaround times. First, the batching problem is considered, in which batch completion times are equally divided over the day to spread the workload. This reduces the peaks of physical work available in the laboratory. Thereafter, the remaining processes are scheduled to minimize the tardiness of orders. Preliminary results show that using this decomposition method, the peaks in histopathology workload in UMC Utrecht, a large university medical center in the Netherlands, are potentially reduced with up to 50% by better spreading the workload over the day. Furthermore, turnaround times are potentially reduced with up to 20% compared to current practices.

1 Introduction

The histopathology and anatomic pathology laboratories consist of a sequence of labor intensive processes. Therefore, resources and personnel in the laboratories should be used effectively (Buesa, 2009). However, challenges exist regarding turnaround times and employee workload (Muirhead et al., 2010). In this study we aim to reduce the peaks in workload for histopathology technicians, while ensuring turnaround times within the required norms (see Stotler et al., 2012; Buesa, 2004), by analyzing planning and scheduling solutions for histopathology resources. This is particularly relevant for patients awaiting a cancer diagnosis, since a long lead time of pathology processes may lead to emotional and physical distress (Paul et al., 2012).

Histopathology processes are complex processes (Brown, 2004). The process can be divided into five main steps: grossing, tissue processing, embedding, sectioning and staining, and examination. This system of processes can be defined as a multi-stage, multiproduct flow shop, in which all specimens go through a predefined order of stages in which only their parameter values vary, as known from the process industry (Harjunkoski et al. 2014; Méndez et al., 2006; Gupta and Karimi, 2003). All stages consist of several single-unit parallel processors, except for the tissue processing stage. Here, batch processors are to be scheduled with large processing times compared to the other stages.

The multi-stage, multiproduct flow shop planning and scheduling is a difficult problem to solve, due to the large amount of solutions (Prasad and Maravelias, 2008). Frequently used exact approaches to solve these problems are Mixed-Integer Linear Programming (MILP) and Mixed-Integer Non-Linear Programming (MINLP). Many approaches consider batch size and batch scheduling decisions separately, for complexity reasons. A few approaches exist that combine batch size, batch assignment, and batch sequencing decisions (Prasad and Maravelias, 2008). However, these approaches only allow for very small instances, with limited number of resources and orders (Harjunkoski and Grossmann, 2002). For real life settings, with more orders to be scheduled, heuristics are used.

The lead-time optimization of histopathology laboratory processes requires a system-wide approach. Existing approaches consist of lean or rapid improvement events focusing on operational bottlenecks, and trial-and-error experimentation with interventions on the operational level of control (i.e., Brown, 2004). Other work focusses on optimizing tissue processing machines (i.e., Vernon, 2005). In this research we aim to integrally optimize histopathology processes by considering all resources involved, and addressing the tactical level of control in addition to the operational (Hans et al., 2012). More specifically, at a tactical level we optimize the batch completion times in order to spread the workload, and at an operational level we reschedule the orders in the histopathology laboratory such that the tardiness of orders is minimized. For both problems we use an (M)ILP approach.

The remainder of this paper is organized as follows. Section 2 gives a description of the histopathology laboratory. In Section 3, we define the problem, and give the mathematical formulation of the problem. Section 4 presents preliminary results of the application of our method in an academic histopathology laboratory. Section 5 ends with conclusions, discussion, and opportunities for further research.

2 Histopathology processes

The histopathology process can be divided into five main steps: grossing, tissue processing, embedding, sectioning and staining, and examination, as shown in Figure 1. Depending on the size and the moment of arrival of a tissue sample, tissue becomes available for the grossing stage immediately, or the next day. Information on tissue arrival is unknown.

In the grossing stage, tissues are trimmed in representative parts by a technician, and put into cassettes. In the automated tissue processing stage, the tissue in these cassettes is fixated and dehydrated using various chemicals. This process takes up to 12 hours depending on the tissue size. After tissue processing, the tissues are embedded in paraffin wax by a technician, to be sectioned in very thin sections (+/- 4μ) by another technician. When these sections are put on slides, the slides receive a staining using an automated stainer, which is required for the residents and pathologists to subsequently examine the slides under the microscope or using digital examination.

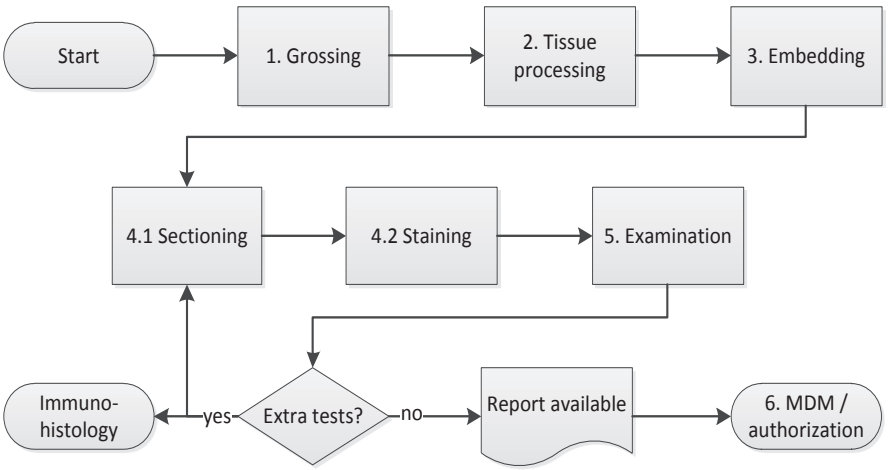


Figure 1: Histopathology processes

In many academic hospitals in the Netherlands, the tissue processing is regularly done in batches during the night, due to the large processing time of the conventional tissue processors. By overnight tissue processing, turnaround times are unnecessarily increased with one night. Currently, schedules of pathologists and technicians accommodate this delay for diagnosis, by facilitating the overnight tissue processing (Vernon, 2005). This results in batch processing throughout all stages of the histopathology laboratory. The implications of overnight tissue processing for the diagnostic workload in the histopathology laboratory are a buzzy environment in the morning, with lower work pressure in the afternoon (Buesa, 2009). However, when introducing tissue processing during the day, specific activities, such as sectioning, will shift from the early morning towards the afternoon (Vernon, 2005), which has consequences for the spread of workload over the day.

As a case study we consider the histopathology laboratory of the department of Pathology of University Medical Center Utrecht (UMCU). UMCU is a 1042 bed academic hospital which is committed to patient care, research, and education. In UMCU's department of Pathology there are several laboratories, such as the histopathology laboratory, the immunochemistry laboratory, the DNA-laboratory, and cytology. The histopathology laboratory evaluates tissue of close to 30.000 patients each year, resulting in the examination of some 140,000 slides each year.

3 Problem description

This study considers the scheduling of histopathology processes, using a decomposed, two-phase approach, since exact approaches to solve the batching and the scheduling problem simultaneously, only allow for very small instances, with limited number of resources, batches, and orders (Harjunkoski and Grossmann, 2002). First, batching moments are determined to minimize the workload. This is called the *batching problem* (Section 3.1). Second, orders

are scheduled for all resources to minimize the tardiness, given the start times of batches from the first phase. This is called the *scheduling problem* (Section 3.2). To solve the batching and scheduling problem we propose two (M)ILP models. Furthermore, we use an approximation method for solving larger instances of the scheduling problem.

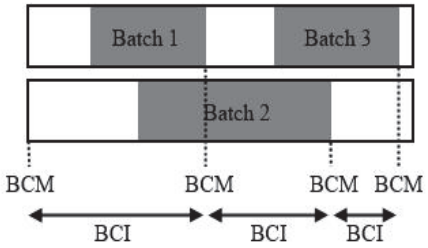


Figure 2: BCMs and BCIs for a 2 machine problem

3.1 Batching problem

The batching problem focuses on scheduling tissue processing batches on multiple machines (tissue processors) aiming to minimize the workload for employees. This problem is considered separately, since the tissue processors experience very high processing times compared to the remaining processes, and since they are the only batch processors in the system. The expected duration of the batches might differ, but is known. All batches can be processed on all machines, and preemption is not allowed. The moment that a batch is finished is referred to as batch completion moment (BCM). The interval between two subsequent BCMs is defined as the batch completion interval (BCI): see Figure 2. The length of the BCIs depends on the assignment, sequence, and timing of the batches.

In this research, we aim to spread the BCMs over the day, such that peaks in workload in the subsequent stages are minimized. Consider a set of B batches (b=1,…,B). We then maximize the minimum batch completion interval: $\max \min_{b \in B} BCI_b$.

Under our objective, the workload is most effectively divided over the day when all batches contain the same number of slides, i.e., lead to the same workload in subsequent stages. In practice, if two batches of the same batch type are scheduled within a small time frame, only a few new arrivals have occurred, and thus the workload resulting from the second batch will be small compared to the workload resulting from the first batch. Therefore, the time between the completion of subsequent batches of the same type should be maximized. Consider a set of T batch types (t=1,…,T), with for each batch type t a corresponding set of batches B_t ($B_t \subseteq B$). This gives a second objective: $\max \sum_{t \in T} \min_{b \in B_t} \{BCI_{b,t}\}$, where $BCI_{b,t}$ equals the interval between two subsequent BCMs of the same batch type, which is minimized for all batch types.

To determine the maximum minimum BCI, we formulated an ILP that not only decides upon the batch sequencing on each machine and the batch timing (e.g. the completion time of all batches), as proposed in Van Essen et al. (2012), but also considers the batch-machine assignment. This way, we can determine the BCIs, using the sequence in which all batches are finished by taking the interval in between subsequent batches. We consider the following as given:

- A set of B batches ($b \in B$);
- A set of T batch types ($t \in T$), and each batch type t has its own set of batches B_t ($B_t \subseteq B$);
- A set of M machines ($m \in M$), with known start time *s* and end time *e*.

Considering the machine assignment, we introduce a binary variable $X_{b,m}$:

$$X_{b,m} = \begin{cases} 1 \; if \; batch \; b \; is \; scheduled \; on \; machine \; m \\ 0 \qquad\qquad otherwise \end{cases}$$

Each batch $b \in B$ should be assigned to exactly one machine $m \in M$. This gives:

$$\sum_m X_{b,m} = 1 \quad \forall b \in B \; (1)$$

Considering the batch sequencing on each machine, we define a position variable P_b indicating the overall completion position of a batch bϵB, and we introduce a binary variable $Y_{b,b'}$:

$$Y_{b,b'} = \begin{cases} 1 \; if \; batch \; b \; is \; scheduled \; somewhere \; before \; batch \; b' \\ 0 \qquad\qquad otherwise \end{cases}$$

The position of a batch $b \in B$ equals one plus the number of batches scheduled before this batch. Furthermore, a batch $b \in B$is either scheduled before batch $b' \in B$, or after batch $b' \in B$. This gives:

$$P_b = \sum_{b'} Y_{b',b} + 1 \quad \forall b \in B \; (2)$$

$$Y_{b,b'} + Y_{b',b} = 1 \quad \forall b, b' \in B, b < b' \; (3)$$

Since cycles in the positioning are not allowed, and no batch can be on the same position as one of its successors, we introduce the following big-M constraint:

$$P_b \leq P_{b'} - 1 + BigM * Y_{b',b} \quad \forall b, b' \in B \; (4)$$

Now the batch assignment and sequencing are guaranteed, we consider the batch timing. The completion time C_b and starting time S_b of a batch $b \in B$ depend on the processing time p_b. This gives:

$$C_b = S_b + p_b \quad \forall b \in B \; (5)$$

A batch $b \in B$ can only start processing after the machines' starting time *s* and should be finished before the end time *e*. We consider the same start and end time for all machines, which gives:

$$S_b \geq s \quad \forall b \in B \; (6)$$

$$C_b \leq e \quad \forall b \in B \; (7)$$

The completion time and starting time of two successive batches scheduled on the same machine mϵM, cannot overlap. This gives:

$$C_b - BigM * (1 - Y_{b,b'}) \leq S_{b'} + BigM * (2 - X_{b,m} - X_{b',m}) \quad \forall b, b' \in B, m \in M \; (8)$$

Now the assignment, sequencing, and timing is assured, we can determine the batch completion intervals. Let the first objective, $\min_{b \in B} BCI_b$, be represented by OBJ1, and the second objective, $\min_{b \in B_t}\{BCI_{b,t}\}$ be represented by $OBJ2_t$.

$$OBJ1 \leq C_{b'} - C_b + BigM * \left(1 - Y_{b,b'}\right) \quad \forall b, b' \in B \text{ (9)}$$

$$OBJ2_t \leq C_{b'} - C_b + BigM * \left(1 - Y_{b,b'}\right) \quad \forall t \in T, b, b' \in B_t \text{ (10)}$$

When necessary, one can include the start time of the interval as batch completion moment, which gives two additional constraints:

$$OBJ1 \leq C_b - s \quad \forall b \in B \text{ (11)}$$

$$OBJ2_t \leq C_b - s \quad \forall t \in T, b \in B_t \text{ (12)}$$

Furthermore, we can set a lower bound to the objective, since it cannot become negative:

$$OBJ1 \geq 0 \text{ (13)}$$

The objective of the ILP is a weighted sum of the two objectives mentioned, i.e. maximize the minimum batch completion interval and maximize the minimum interval between the completions of two batches of the same type. This gives:

$$\max(\alpha * BCI + \beta \sum_{t \in T} BCI_t) \text{ (14)}$$

3.2 Scheduling problem

The scheduling problem encompasses three decisions to minimize the tardiness of orders: The sequencing of orders, the timing of all processes, and the order assignment to resources.

We consider multiple stages and multiple resources per stage, as shown in Figure 1. Orders arrive to the system, with known target due dates. Furthermore, it is known which resources are allowed to be used to process which orders. Preemption of orders is not allowed, since it can cause contamination of specimens, which causes diagnostic errors.

To solve the scheduling problem to optimality we propose an extended MILP formulation of the problem of Gupta and Karimi (2003) that decides upon the order assignment to resources in each stage, order sequencing on each resource, and the order timing. We consider the following as given:

- A set of G stages ($g \in G$).
- A set of J resources ($j \in J$), and each stage s has its own set of resources J_s.
- A set of B batches ($b \in B$), with known resource J_b, and start times S_b.

- A set of I different orders (corresponding to the incoming specimens) ($i \in I$), with known target due dates d_i, and known sets of resources J_i and batches B_i, which are allowed to process this order.
- A set of T different order types ($t \in T$). Each order type $t \in T$ has its own set of orders I_t, consisting of all orders of that type, and its own set of batches B_t ($B_t \subseteq B$).

The scheduling problem can be written as a MILP. The sequencing of orders in the non-batching stages can be modeled using adaptions to the constraints presented by Gupta and Karimi (2003). Furthermore, we need to decide upon the assignment of orders to batches and resources, and the timing of orders.

We define three binary variables Z_{ij}, ZF_{ij}, and $A_{ii'g}$ as follows:

$$Z_{i,j} = \begin{cases} 1 & if\ order\ i\ is\ processed\ by\ unit\ j \\ 0 & otherwise \end{cases}$$

$$ZF_{i,j} = \begin{cases} 1 & if\ order\ i\ is\ processed\ first\ by\ unit\ j \\ 0 & otherwise \end{cases}$$

$$A_{i,i',g} = \begin{cases} 1 & if\ order\ i\ is\ processed\ directly\ before\ order\ i'\ in\ stage\ g \\ 0 & otherwise \end{cases}$$

First of all, each order needs to be assigned to exactly one resource in each stage, since an order has to be processed in each stage exactly once (15). From all orders assigned to an operating resource j, one order has to be processed first (16). Since not all resources have to be operating, the left hand side of constraint (16) can also be zero.

$$\sum_{j \in J_{ig}} Z_{ij} = 1 \quad \forall i, g \quad (15)$$

$$\sum_{i \in I_j} ZF_{ij} \leq 1 \quad \forall j \quad (16)$$

Order $i \in I_j$ can only be processed first on resource $j \in J$ if it is assigned to that resource (17).

$$Z_{ij} \geq ZF_{ij} \qquad \forall i \in I_j \quad (17)$$

An order cannot have more than one feasible predecessor and one feasible successor in each stage. Each order can be processed first on a specific resource, or it succeeds another order (18). Furthermore, orders cannot have more than one direct successor (19).

$$\sum_{i' \notin NC_{ig}} A_{i',i,g} + \sum_{j \in J_{ig}} ZF_{ij} = 1 \quad \forall i, g \quad (18)$$

$$\sum_{i' \notin NC_{ig}} A_{i,i',g} \leq 1 \quad \forall i, g \quad (19)$$

To assign resources to a specific resource $j \in J$, it should hold that successive orders $i \in I$ and $i' \in I$ cannot be processed by resources that cannot process them both, but should be processed by a single resource $j \in J_{ig} \cap J_{i'g}$

(20) (21). The combination of constraints (20) and (21) performed best in the review of Gupta and Karimi (2003), and were therefore included in our model.

$$A_{i,i',g} + A_{i',i,g} + \sum_{j \notin J_{ig} \cap J_{i'g}} Z_{i,j} \leq 1 \quad \forall g, i, i' > i, (i,i') \in I_g, i' \notin NC_{ig} \text{ (20)}$$

$$Z_{i'j} \leq Z_{ij} + 1 - A_{i,i'g} - A_{i',i,g} \quad \forall g, i, i' > i, (i,i') \in I_g, i' \notin NC_{ig}, j \in J_{ig} \cap J_{i'g} \text{ (21)}$$

Now the order assignment and sequencing is accounted for, the start times of the orders should be set in each stage, as follows from the continuous time representation. Therefore, we define a decision variable S_{ig} as follows:

$$S_{i,g} = start\ time\ at\ which\ order\ i\ starts\ processing\ in\ stage\ g$$

To assign an order to a batch in the batching stage, we need an indicator for an order to be assigned to a specific time slot. Therefore, we define variable $Q_{i,j,b}$ as follows:

$$Q_{i,j,b} = \begin{cases} 1 & if\ order\ i\ is\ processed\ in\ batch\ b\ on\ unit\ j \\ 0 & otherwise \end{cases}$$

An order $i \in I$ can only start processing in the next stage, after order $i \in I$ has finished processing in the previous stage, and is transported to the next stage. Therefore, stage sequencing constraints are introduced.

When a batch $b \in B$ is selected in a batching stage, this batch should start processing after order $i \in I$ has finished processing in the previous stage, and is transported to the batching stage (22).

$$\sum_j \sum_b Q_{i,j,b} * bs_{j,b} \geq S_{i,g-1} + \sum_{j \in J_{i,g-1}} \left(Z_{ij} * \left(f_{ij} * t_{ij} + tt_{ij} \right) \right) \quad \forall i, g \in G^{batch} \text{ (22)}$$

To start processing in a post-batch stage, all orders of the batch containing order i should be fully processed in the batching stage, and transported towards the post-batch stage (23), with ns_{ig} defined as the next processing stage of order $i \in I$, currently being processed in stage $g \in G$.

$$S_{ig'} \geq S_{ig} + \sum_{j \in J_{ig}} (Z_{ij} * (tb_j + tt_{ij})) \quad \forall i, g \in G^{batch}, g' \in ns_{ig} \backslash G^{batch} \text{ (23)}$$

In the stage sequencing relation between two non-batching stages, order $i \in I$ has to finish processing in stage $g \in G$ and be transported to the next stage before starting in next stage (24). The stage dependent timing constraints are adapted from the timing constraint of Gupta and Karimi (2003) to take the increasing order size into account, and to correct for batching influences.

$$S_{ig'} \geq S_{ig} + \sum_{j \in J_{ig}} (Z_{ij} * (f_{ij} * t_{ij} + tt_{ij})) \quad \forall i, g \notin G^{batch}, g' \in ns_{ig} \backslash G^{batch} \text{ (24)}$$

Not only relations between stages influence the timing of orders on processing resources, also the relation between orders should be taken into account.

In all non-batching stages, order $i' \in I$ can start processing on $j \in J$ after its predecessor order $i \in I$ is finished (25). This constraint is adapted from the constraint of Gupta and Karimi (2003) to take the increasing order size into account.

$$BigM * \left(1 - A_{i,i',g}\right) + S_{i',g} \geq S_{ig} + \sum_{j \in J_{ig}} (Z_{ij} * f_{ij} * t_{ij}) \quad \forall g \notin G^{batch}, i, i' \notin NC_{ig} \text{ (25)}$$

The timing of orders on resources is subject to some constraints. The first order $i \in I$ on resource $j \in J$ can only start processing after the release time of the resource (26). Furthermore, each order can only start processing after its release time (27). Setup times are not taken into account.

$$S_{ig} \geq \sum_{j \in J_{ig}} (ZF_{ij} * URT_j) \quad \forall g, i \text{ (26)}$$

$$S_{i,g} \geq ORT_i \quad \forall i, g = 1 \text{ (27)}$$

The assignment of orders to a specific batch on a specific resource, is subject to two constraints. First, all orders can only be assigned to one batch, which follows from constraint (28). Second, the corresponding batch starting time equals the order timing of order i in stage g (29).

$$\sum_j \sum_b Q_{ijb} = 1 \quad \forall i \text{ (28)}$$

$$S_{ig} = \sum_j \sum_b Q_{i,j,b} * bs_{j,b} \quad \forall i, g \in G^{batch} \text{ (29)}$$

Orders can only start processing on resources when the resources are available. Since resources are unavailable during night-hours, we consider D nights during the planning horizon. To indicate if order $i \in I$ is planned before or after a certain night $d \in D$, let W_{dig} be an auxiliary binary variable defined as follows:

$$W_{d,i,j} = \begin{cases} 1 & if\ order\ i\ is\ processed\ after\ night\ d\ on\ resource\ j \\ 0 & otherwise \end{cases}$$

Processing of any order in any stage cannot start at moments it cannot be finished before the closing hours of the resource. Therefore, processing of an order $i \in I$ in stage $g \in G$ should start before or after the non-working moments (30) (31), which does not involve the transfer time. These constraints only holds for non-batching stages, since the batch processors in the histopathology laboratory model are able to work during night hours, when the process is started before the start of the night.

$$S_{ig} \leq \sum_{j \in J_{ig}} (Z_{ij} * NW1_{dj} - f_{ij} * t_{ij}) + BigM * W_{dij} \quad \forall d, g \notin G^{batch}, i \in I_g \text{ (30)}$$

$$S_{ig} \geq (NW2_d + URT_j) * W_{dij} \quad \forall d, g \notin G^{batch}, j \in J_g, i \text{ (31)}$$

The objective is to minimize the weighted tardiness of all orders. Let us define Sd_i as follows:

$$Sd_i = delay\ of\ order\ i$$

The tardiness of order $i \in I$ equals the sum of the start time in last stage ($\bar{g} \in G$), the transfer time of order $i \in I$ in this stage, and the order factor times the processing time of order $i \in I$ in this stage, which together equals the completion time of order $i \in I$, minus the due date of this order (dd_i) (32). This constraint is adapted from Gupta and Karimi (2003).

$$Sd_i \geq [S_{i,g} + \sum_{j \in J_{i,g}} (Z_{ij} * f_{ij} * t_{ij} + tt_{ij})] - dd_i \quad \forall i, g \in \bar{g} \text{ (32)}$$

Specific specimen types are more important to finish on time than others. Therefore, the orders are prioritized, by priority factor δ_i. This makes the objective to minimize the sum of the weighted tardiness (33).

$$minimize \ \sum_{i \in I} (\delta_i * Sd_i) \text{ (33)}$$

Some additional constraints are proposed to increase the efficiency of the MILP. An upper bound on the order timing T_{is} can be given by the end time of the planning horizon H. A better upper bound is derived when subtracting the processing time of order i in the final stage. Since the processing times are equal in all resources, the last resource is chosen for the upper bound determination. This results in constraint (34).

$$S_{ig} \leq H - (f_{ij} * t_{ij} + tt_{ij}) \quad \forall g, i, j = 13 \text{ (34)}$$

When an order cannot be processed by resource $j \in J$, since it is not allowed to be processed by that resource (i.e. $i \notin I_j$), the order cannot be assigned to that resource (35).

$$Z_{ij} = 0 \quad \forall j, i \notin I_j \text{ (35)}$$

As mentioned, only small instances can be solved using the MILP, due to the large problem size of real life instances and the long computation time (Harjunkoski and Grossmann, 2002). Therefore, we propose a constructive heuristic based on several dispatching rules to find a feasible solution within reasonable time for real life instances (including up to 130 orders per time interval, 4 stages, and 13 resources). These dispatching rules include Earliest Due Date (EDD) and First In First Out (FIFO), since these are easy to implement in the histopathology practices and have shown to result in near optimal solutions (Haupt, 1988). In the remainder of this research, we will use EDD

4 Results

The histopathology laboratory of UMCU has provided real life data to evaluate the applicability and performance of the solution method. We consider 10 different problem instances based on historical data of 22,379 patients derived from January to December 2013. The instances differ in terms of number and type of orders. Each instance includes four order types, corresponding with large specimens (type 1), small specimens (including biopsies) (type 2), priority specimens (type 3), and external specimens (type 4).

The priority of the order type is reflected in their due date, as shown in Table 1. The turnaround time (TAT) targets per order type, and therefore the corresponding due dates, are set by hospital management, the Dutch government, and external parties, to ensure a timely diagnosis for all patients (Pathologie, 2013).

We consider two scenarios. First we consider the current situation, for which only the scheduling problem is solved. The batching problem is not solved since the batching moments are already known in the current situation. Second we consider the situation with the batching policy as derived from the batching model. In both scenarios we fix one batch of type 3, to 11:15 AM each day, due to hospital regulations.

All experiments are solved on a HP laptop personal computer with 2GB RAM, using CPLEX 12.6 in AIMMS 4.0.

Table 1: TAT targets per order type

Order type	TAT target
Order type 1	90% diagnosed within 7 days
Order type 2	90% diagnosed within 5 days
Order type 3	80% diagnosed within 24 hours
Order type 4	90% diagnosed within 3 days

4.1 Current situation

In the current situation, all orders are processed in batches during the night, except for type 3 orders, which are processed on fixed moments during the morning, but only consist of a very small amount of orders (1-3 slides per batch). This results in a high workload during the morning, as shown in Figure 3 for one representative instance.

The overall TAT results are shown in Table 2. One can see that only a small percentage of type 2 and type 4 orders are ready before their due date. This is a direct result of tissue processing during the night, which leads to a one-day delay for all orders.

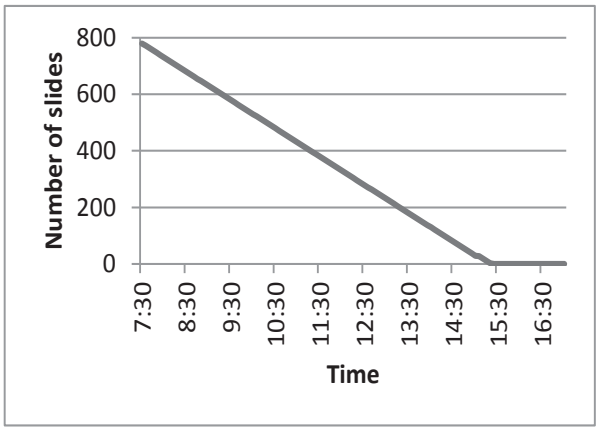


Figure 4: Workload performance current **situation**

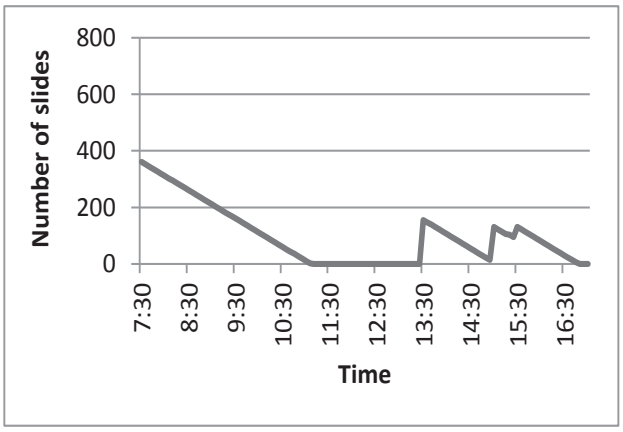


Figure 3: Workload performance batching policy

4.2 Batching policy

In the batching policy, we consider four interventions, based on the number of batches per order type per day allowed: (2-3-1), (1-3-1), (2-2-1), (1-2-1)[1]. Order type 1 batches are omitted, since type 1 orders are technically restricted to be processed during the night.

Figure 4 shows the spread in workload for one representative instance, including 6 batches. The TAT results are shown in Table 2. All interventions showed improved results regarding their norms, but specific patient types experience reduced performance compared to the current situation, such as type 1 patients. However, the results show that the performance of an intervention depends on the timing of the batches, especially in relation to the underlying arrival patterns of orders.

Table 2: TAT performance per scenario

Intervention	1: Current situation	2:(3-1-1)	3:(2-1-1)	4:(2-2-1)	5:(3-2-1)
Type 1 patients on time	99,4%	93,9%	93,6%	93,4%	93,4%
Type 2 patients on time	54,6%	88,4%	91,7%	94,7%	97,0%
Type 3 patients on time	98,0%	97,4%	97,4%	98,0%	97,4%
Type 4 patients on time	84,5%	92,5%	90,9%	90,4%	88,7%
TAT (in hours)	25,77	21,00	20,96	21,29	20,12

5 Conclusion / Discussion

We have introduced a decomposed solution method to optimize and prospectively assess the planning and scheduling of batches and orders in the histopathology laboratory. The results show that the turnaround time, which is the main performance indicator, can be reduced by 20% through eliminating unnecessary waiting during the night hours. Furthermore, peaks in workload can be reduced by more than 50% by shifting a part of the pile of work from the morning towards the afternoon.

The batches under the solution approach are not always equally filled, which in specific cases may result in larger or smaller peaks in workload depending on the patient arrival pattern, especially since different arrival patterns are encountered over the day. Therefore, future work will be dedicated to analyze the effect of weighing the BCIs according to the arrival distribution of orders per order type during the corresponding BCI.

By fixing the batch starting times of specific batches, the corresponding orders in that batch are prioritized, since they have a higher chance of being processed at a favorable time. The analysis showed evidence that prioritizing specific order types increases the TAT performance of those orders. However, this occurs at expense of others. Further research will be executed to analyze this relation.

[1] (# order type 2 batches, # order type 3 batches, # order type 4 batches)

Based on this work, UMC Utrecht is currently implementing planning and control approaches in the histopathology laboratory regarding the planning and scheduling of tissue processing batches and stage one resources.

References

Brown, L. (2004). Improving histopathology turnaround time: a process management approach. In: Current Diagnostic Pathology. 10: 444-452.

Buesa, R.J. (2009). Adapting lean to histology laboratories [Technical Note]. In: Annals of Diagnostic Pathology. 13: 322-333.

Essen, J.T. van, Hans, E.W., Hurink, J.L., and Oversberg, A. (2012). Minimizing the waiting time for emergency surgery. In: Operations Research for Health Care. 1: 34-44.

Gupta, S., and Karimi, I.A. (2003). An improved MILP formulation for scheduling multiproduct, multistage batch plants. In: Industrial & engineering chemistry research. 42(11): 2365-2380.

Hans, E. W., Van Houdenhoven, M., and Hulshof, P. J. (2012). A framework for healthcare planning and control. In Handbook of healthcare system scheduling (pp. 303-320). Springer US.

Harjunkoski, I., and Grossmann, I.E. (2002). Decomposition techniques for multistage scheduling problems using mixed-integer and constraint programming methods. In: Computers and Chemical Engineering. 26: 1533-1552.

Harjunkoski, I., Maravelias, C.T., Bongers, P., Castro, P.M., Engell, S., Grossmann, I.E., Hooker, J., Méndez, C., Sand, G., Wassick, J. (2014). Scope for industrial applications of production scheduling models and solution methods. In: Computers and Chemical Engineering. 62: 161-193.

Haupt, R. (1988). A survey of priority rule-based scheduling. In: OR Spectrum. 11: 3-16.

Méndez, C.A., Cerdá, J., Grossmann, I.E., Harjunkoski, I., and Fahl, M. (2006). State-of-the-art review of optimization methods for short-term scheduling of batch processes. In: Computers & Chemical Eningeering. 30(6): 913-946.

Muirhead, D., Aoun, P., Powell, M., Juncker, F., and Mollerup, J. (2010). Pathology Economic Model Tool. A novel approach to workflow and budget cost analysis in an anatomic pathology laboratory. In: Arch. Pathol. Lab. Med. 134: 1164-1169.

Pathologie. (2013). Kwaliteitshandboek pathologie. Retrieved from: http://magnus/kwaliteitsdocumenten/00025313.html

Paul, C., Carey, M., Anderson, A., Mackenzie, L., Sanson-Fisher, R., Courtney, R., and Clinton-McHarg, T. (2012). Cancer patients' concerns regarding access to cancer care: perceived impact of waiting times along the diagnosis and treatment journey. In: European Journal of Cancer Care. 21(3): 321-329.

Prasad, P., and Maravelias, C.T. (2008). Batch selection, assignment and sequencing in multi-stage multi-product processes. In: Computers and Chemical Engineering. 32: 1106-1119.

Stotler, B.A., Kratz, A. (2012). Determination of turnaround time in the clinical laboratory. In: American Journal of Clinical Pathology. 138: 724-729.

Vernon, S.E. (2005). Continuous throughput rapid tissue processing revolutionizes histopathology workflow. In: Labmedicine. 36(5): 300-302.

Predicting length of stay and assignment of diagnosis codes during hospital inpatient episodes

José Carlos Ferrão, jose.ferrao@tecnico.ulisboa.pt, Siemens Healthcare

Mónica Duarte Oliveira, monica.oliveira@tecnico.ulisboa.pt, CEG-IST

Filipe Janela, filipe.janela@siemens.com, Siemens Healthcare

Henrique Martins, hmartins@fcsaude.ubi.pt, CI2 - HFF

Electronic health record (EHR) data is becoming ubiquitous in the healthcare domain, with potential to provide valuable insights to clinicians and managers. Data mining methodologies have been largely unexplored to analyze EHR data retrospectively and to inform expected patterns of disease and utilization during the course of patient stay. In this work, we propose a data mining methodology based on feature selection and logistic regression models to predict if an episode's length-of-stay will be outside the expected interval, as well as the set of diagnosis codes (from the International Classification of Diseases – ICD) assigned to each episode along the course of patient stay. The experiments were carried out using EHR data from the records of 5089 episodes of inpatients admitted in a large (772-bed) hospital in Portugal. The predictive performance of models in terms of precision, recall and F1-score values showed the potential value of using decision support tools during patient stay, since in several experiments the developed models exhibited performance values with sufficient accuracy to provide support in clinical settings.

1 Introduction

In recent years, information systems have undergone extensive development in the healthcare domain (Ford, Menachemi, and Phillips 2006), where EHR systems have played a central role as main platform for recording clinical data. The large volumes of EHR data have induced interest in reusing these data for research and decision support (Hersh 2007) and in increasing the proportion of data captured in structured formats (Fernando et al. 2012) so as to surpass the challenges of using narrative data (Jaspers et al. 2011).

Data mining in medicine has been a hot topic in recent years (Bellazzi and Zupan 2008; Patel et al. 2009), making use of techniques to extract knowledge from data and support clinicians and managers on their decisions (Iavindrasana et al. 2009). The applications include length-of-stay (LOS) (Rowan et al. 2007) and clinical coding (Stanfill et al. 2010) prediction, amongst others, and typically encompass elements such as retrieval and preparation of historical data, defining a set of variables (features), applying feature selection techniques (when suitable)

and building models to extract patterns from data. However, most research adopts a retrospective batch approach whereby models are developed using all data available in the dataset, without evaluating the potential of developing decision support tools during the course of patient care, as more information becomes available. In the case of inpatient episodes, decision support on clinical and managerial elements such as diagnoses and LOS, respectively, during the course of the episode would be valuable to inform timely decisions.

In this scope, we sought to analyze the extent to which decision support tools may be developed to inform expected LOS (which can be regarded as a proxy of resource utilization) and episode coding (which may inform on disease patterns) during the course of the episode. For this purpose, we propose a data mining methodology to build models for predicting LOS and episode coding in different instants of patient stay. In effect, previous LOS prediction studies have analyzed LOS in different moments of patient stay, yet, up to our knowledge, no study has used a systematic approach of defining features and building models consistently across different moments (and comparing performance amongst them). On the other hand, as far as we could appraise, episode coding has only been addressed within a retrospective frame, without addressing coding support during the course of patient stay. The rest of this article is structured as follows: section 2 describes the proposed methodology, section 3 presents key results to illustrate an application with a real world dataset for LOS and episode coding prediction, and lastly section 4 presents our main concluding remarks.

2 Methods

2.1 EHR data properties and preparation

In this work, the dataset from which we set off consists of EHR database entries produced during clinical practice using the EHR system Soarian® (Haux et al. 2003). These entries contain patient information in structured formats, regarding demographic data, diagnoses, personal history, allergies, prescriptions and medication, as well as structured forms (assessments) composed of labeled fields through which health professionals record information using controlled formats (such as buttons, pick lists and dropdowns). The use of such system eliminates the need to extract computer-readable information from narratives (in which information is "locked" (Hripcsak et al. 1995)) and, thereby, the challenges associated with it.

The development of prediction models based upon structured data requires data to be represented in a data matrix format, i.e., with dataset instances represented in terms of values of the feature space (i.e., a matrix in which lines represent instances and columns represent features) (Bishop 2006). Since EHR data is natively represented as relational database entries (and not in a data matrix format), it was necessary to define features from data. In practice, defining features from data consists in defining the variables based on which clinical data can be represented and whose values (for each instance) are used to build prediction models. This feature definition process requires identifying the clinical concepts contained in the dataset, defining a feature for each concept and assessing its value for each instance (episode) in the dataset, thereby building and populating the data

matrix representation. However, for large datasets (with high number of features and instances), the process of building and populating a data matrix becomes laborious. In order to mitigate this issue, we implemented a framework with a set of routines that build and populate a data matrix automatically from a dataset (further details available in (Ferrão et al. 2013)). Therefore, after using this framework to automatically digest the dataset and build a data matrix from the original EHR database entries, it was then possible to work on the development of prediction models using the clinical dataset represented in a data matrix format.

2.2 Feature selection and model development

In this study, the decision support tools to predict LOS and episode coding were based on prediction models developed from historical data (i.e., clinical records from past episodes) aiming to predict outcomes for new episodes. To this end, we adopted a supervised learning approach (Bishop 2006), which consists in using data from past instances (episodes) and the corresponding known outcomes (in this case, LOS and assigned clinical codes) to fit models that are subsequently used to predict those outcomes for new instances. There are two main types of models within supervised learning models, depending on the type of outcome to be predicted: if the outcome is discrete or nominal, classification models are employed, as opposed to regression models used for continuous outcomes. In our approach, we modeled LOS and episode coding prediction as binary outcomes. Firstly, we modeled LOS as a binary variable according to whether or not the duration of patient stay is within the boundaries of the corresponding diagnosis-related group (DRG) (and thus within the fixed-rate payment scheme). Secondly, for episode coding, we defined a binary variable for each ICD code in the dataset: since multiple codes can be assigned to each episode, we built a model predicting the assignment of each ICD code to a given episode. We hereby describe the processes of data preparation, feature selection and model development.

The data matrix resulting from structured EHR data tends to contain a large number of features, which hinders model performance and, as such, feature selection methods come to play (Guyon and Elisseeff 2003). We chose to implement a filter method for its scalability to large datasets and independence of prediction models (Saeys, Inza, and Larrañaga 2007). We performed several tests with different filter methods by evaluating the predictive power of multiple filter methods, namely fast correlation-based filter (Yu and Liu 2004), information gain and chi-square (Yiming Yang 2014), Relief (Kira and Rendell 1992), symmetrical uncertainty (Press et al. 1992), correlation-based feature selection (Hall 1999) and minimal-redundancy maximal-relevance (mRMR) (Peng, Long, and Ding 2005). In these preliminary tests, mRMR revealed higher performance, and as such, we present results obtained with this method in section 3 (also due to space constraints). In practical terms, we used mRMR feature selection to select a subset of features of the original feature set according to the mRMR criterion, which defines a score based on mutual information and aims to minimize redundancy while maximizing relevance of the feature subset. The mRMR method outputs features sorted by decreasing order of mRMR score. From this ordered set, we selected a subset of the top 50 features. This 50-feature subset was then used to develop logistic regression models.

The prediction models used in this study – logistic regression models – have simple formulations, reasonable scalability and interpretability of results, and are tailored for binary outputs (Hosmer and Lemeshow 2000), being used in this work to predict LOS and ICD code assignment. Logistic regression models were developed in a forward selection, stepwise approach, adding one feature at a time by decreasing order of mRMR score. For each logistic regression model developed, we used classification thresholds ranging from 0 to 1 in steps of 0.005 (in order to compensate for class imbalances). Model performance was tested with 5-fold cross-validation, whereby the dataset is randomly split into 5 subsets, using 4 of these subsets as training set and the remaining one as test set. As evaluation metrics, we analyzed precision and recall, which are based on the proportion of false positives and false negatives, respectively, as well as the F1-score, which is the harmonic mean between the precision and recall.

Our experiment approach consisted in (1) extracting structured EHR data, (2) creating 8 separate datasets for each of the moments after patient admission (using the time stamps associated with each database entry), (3) building a data matrix for each separate dataset, (4) performing feature selection with the mRMR method and (5) developing prediction models and evaluating predictive power for each of the 8 datasets. Predictive power was evaluated at 8 different moments of patient stay: 1, 4, 8, 12, 18, 24, 36 and 48 hours after patient admission. For this purpose, we firstly determined the date/time boundaries and used the timestamps in EHR database records to filter data, thereby creating a dataset for each instant.

3 Results

Our real-world dataset contained 5089 inpatient episodes from medical wards of a large hospital in Portugal. EHR data from these episodes was extracted and prepared, yielding 4820 features with non-missing values. The distribution of positive and negative examples was quite imbalanced: 15.72% of positive examples in the LOS problem, one ICD code with 40% positive examples, 5 other codes with more than 15% of positive instances, while the remaining had less than 10% positive examples.

3.1 LOS prediction

In the first experiment, we developed models to predict whether an episode is within the boundaries of the DRG class it is assigned, in different instants of each episode. The results are depicted in Fig. 3.1. As expected, one may observe that model predictive power tends to increase along the course of the episode. This tendency is evident for precision, i.e., the number of false positives decreases as more information becomes available in the EHR. Recall seems to decrease in initial stages of each episode, recovering along the episode. It is also interesting to note that precision was tendentially higher than recall, which in practice yields that these models are more suitable to correctly spot LOS-outlier episodes. The overall performance (in terms of F1-score) has a steeper increasing trend in the first instants, reaching acceptable values (higher than 50%) in early stages of patient stay.

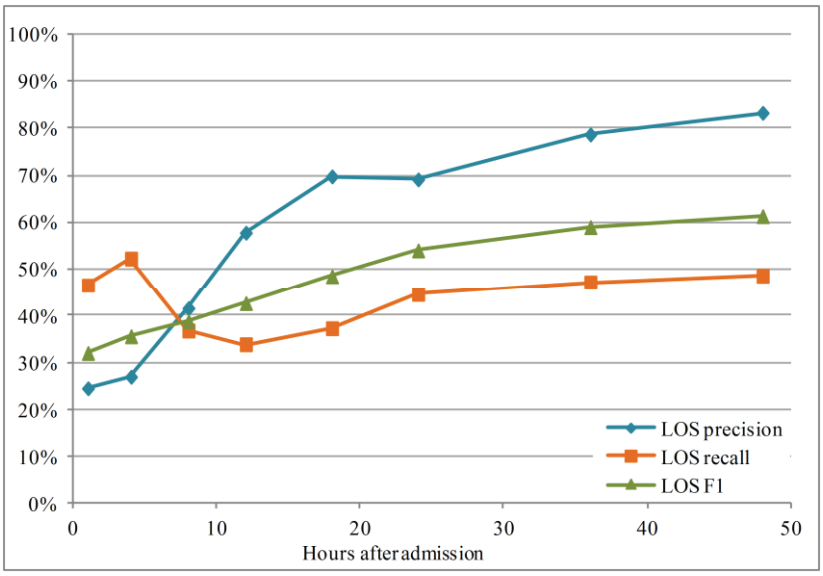
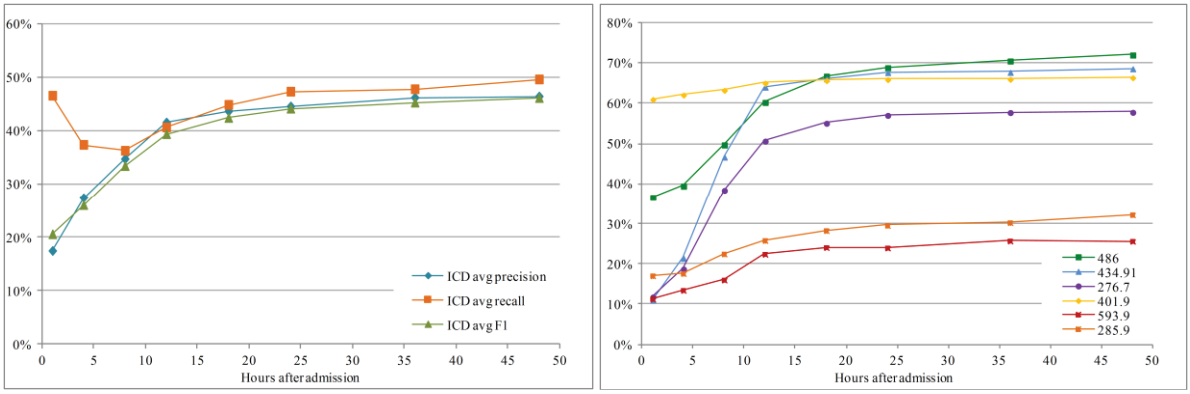


Figure 3.1:

Precision, recall and F1-score values (average of 5-fold cross validation) obtained with logistic regression models to predict if LOS falls within the boundaries of the assigned DRG class.

3.2 ICD codes prediction

We also developed logistic regression models to predict the assignment of ICD-9-CM diagnosis codes. In order to keep the analysis manageable, we focused on the 50 more frequent codes. We analyzed performance averaged across all 50 codes and also analyzed performance for selected codes that showed different tendencies (Fig. 3.2).



(a) (b)

Figure 3.2:

(a) Average performance of logistic regression models for the 50 most frequent ICD-9-CM codes in terms of precision, recall and F1-score; (b) F1-score values for selected ICD-9-CM codes (486 – pneumonia unspecified; 434.91 – cerebral artery occlusion, unspecified, with infarction; 276.7 – hyperpotassemia; 401.9 – essential hypertension, unspecified; 593.9 – renal and urethral disorders, unspecified; 285.9 – anemia, unspecified).

The trends shown in Fig. 3.2(a) exhibit similarities with the ones obtained with models for LOS prediction, namely in what concerns the increase in average predictive power in initial hours of patient stay. Comparing precision and recall, these measures now exhibit different behaviors, with recall being tendentially higher than precision, rendering models more prone to suggesting incorrect codes while decreasing the rate of overlooked codes. It is also interesting to observe the non-monotonic behavior of recall in initial stages of patient stay, suggesting that the patterns of EHR data produced in such periods may induce biases in models.

In order to further investigate the pool of analyzed ICD codes, we selected codes with different behaviors, which are depicted in Fig. 3.2(b). In this chart, it is possible to identify two different patterns, corresponding to codes exhibiting, or not, a steep increase in performance during the course of the episode. ICD codes referring to pneumonia, cerebral artery occlusion and anemia had a marked tendency to increase predictive power on early stages, while the other selected codes approximately maintained model results. In light of such results, we hypothesize that the potential to develop decision support tools and provide valuable insights during patient stay is highly dependent on the type of outcome being predicted, and especially on the characteristics of EHR data (e.g. data quality, feature subsets) underlying model development.

4 Conclusions and future work

In this study, we proposed a methodology to analyze the extent to which valuable decision support may be provided during the course of the episode, using a real-world dataset of inpatient episodes. Specifically, after data preparation and feature selection, we have built logistic regression models and analyzed model performance in predicting LOS-outliers and ICD code assignment in selected instants of patient stay. Our main conclusions refer to the fact that, as expectable, model performance tends to increase during the course of the episode, especially in the first hours after patient admission. Secondly, this behavior was observed in different magnitudes in predicting LOS and code assignment, which denotes the influence of modeling context and of dataset characteristics in model behavior. It was also possible to observe non-monotonic behaviors (Fig. 3.1 and 3.2(a)), which point to the possibility that statistical artifacts and data quality issues may be exerting influence on model results. Lastly, we could observe that model performance does not always exhibit a steadily increasing trend, either by starting off at higher values (e.g. code 401.9) or maintaining lower values in spite of the increasing availability of EHR data.

From these preliminary results, we wrap up by stating that the research path of developing decision support tools to provide on-the-fly insights in healthcare settings may be promising, as demonstrated by the existence of high-performing models on very early stages of patient stay, with strong correlation with the context of application and the availability of data.

In terms of future work, it should be relevant to firstly carry out a thorough analysis of data quality and data patterns produced in different stages of patient stay, in order to explore the value of information at different

stages of each episode. It may also be relevant to analyze feature subsets in further detail using domain knowledge from clinical experts as a means to improve the set of features that are used to build prediction models. Furthermore, it should also be relevant to test the behavior of different prediction models during the course of the episode. In effect, we have tested other models in the scope of coding support in our previous works (Ferrão et al. 2012; Ferrão et al. 2013), but only testing model behavior using all available EHR data. Lastly, additional methods to tackle issues of inter-label relationships and class imbalance may also be valuable in this context.

References

Bellazzi, Riccardo, and Blaz Zupan. 2008. "Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines." *International Journal of Medical Informatics* 77 (2) (February): 81–97.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Edited by Michael Jordan, Jon Kleinberg, and Bernhard Scholkopf. Singapore: Springer.

Fernando, Bernard, Dipak Kalra, Zoe Morrison, Emma Byrne, and Aziz Sheikh. 2012. "Benefits and Risks of Structuring And/or Coding the Presenting Patient History in the Electronic Health Record: Systematic Review." *BMJ Quality & Safety* 21 (4) (April): 337–46.

Ferrão, José Carlos, Monica Duarte Oliveira, Filipe Janela, and Henrique Manuel Gil Martins. 2012. "Clinical Coding Support Based on Structured Data Stored in Electronic Health Records." *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops* (October): 790–797.

Ferrão, José Carlos, Monica Duarte Oliveira, Filipe Janela, and Henrique Manuel Gil Martins. 2013. "Using Structured EHR Data and SVM to Support ICD-9-CM Coding." *2013 IEEE International Conference on Healthcare Informatics* (September): 511–516.

Ford, Eric W, Nir Menachemi, and Thad Phillips. 2006. "Predicting the Adoption of Electronic Health Records by Physicians: When Will Health Care Be Paperless ?" (13): 106–113.

Guyon, Isabelle, and André Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3: 1157–1182.

Hall, Mark A. 1999. "Correlation-Based Feature Selection for Machine Learning." The University of Waikato.

Haux, R, C Seggewies, W Baldauf-Sobez, P Kullmann, H Reichert, L Luedecke, and H Seibold. 2003. "Soarian - Workflow Management Applied for Health Care." *Methods of Information in Medicine* 42 (1) (January): 25–36.

Hersh, William R. 2007. "Adding Value to the Electronic Health Record through Secondary Use of Data for Quality Assurance, Research, and Surveillance." *The American Journal of Managed Care* 13 (6 Part 1) (June): 277–8.

Hosmer, David W, and Stanley Lemeshow. 2000. *Applied Logistic Regression*. 2nd ed. John Wiley & Sons, Inc.

Hripcsak, George, Carol Friedman, Philip O Alderson, William DuMouchel, Stephen B Johnson, and Paul D Clayton. 1995. "Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing." *Annals of Internal Medicine* 122 (9) (May 1): 681–688.

Iavindrasana, J, G Cohen, A Depeursinge, H Müller, R Meyer, and A Geissbuhler. 2009. "Clinical Data Mining: A Review." *Yearbook of Medical Informatics* (January): 121–33.

Jaspers, Monique W M, Marian Smeulers, Hester Vermeulen, and Linda W Peute. 2011. "Effects of Clinical Decision-Support Systems on Practitioner Performance and Patient Outcomes: A Synthesis of High-Quality Systematic Review Findings." *Journal of the American Medical Informatics Association : JAMIA* 18 (3) (May 1): 327–34.

Kira, Kenji, and Larry A. Rendell. 1992. "The Feature Selection Problem: Traditional Methods and a New Algorithm" (July 12): 129–134.

Patel, Vimla L, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. 2009. "The Coming of Age of Artificial Intelligence in Medicine." *Artificial Intelligence in Medicine* 46 (1) (May 1): 5–17.

Peng, Hanchuan, Fuhui Long, and Chris Ding. 2005. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8) (August 1): 1226–38.

Press, William H., Saul H. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in "C."* Cambridge University Press.

Rowan, Michael, Thomas Ryan, Francis Hegarty, and Neil O'Hare. 2007. "The Use of Artificial Neural Networks to Stratify the Length of Stay of Cardiac Patients Based on Preoperative and Initial Postoperative Factors." *Artificial Intelligence in Medicine* 40 (3) (July): 211–21.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. 2007. "A Review of Feature Selection Techniques in Bioinformatics." *Bioinformatics (Oxford, England)* 23 (19) (October 1): 2507–17.

Stanfill, Mary H, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. "A Systematic Literature Review of Automated Clinical Coding and Classification Systems." *Journal of the American Medical Informatics Association* 17 (6): 646–51.

Yiming Yang, Jan O. Pedersen. 2014. "A Comparative Study on Feature Selection in Text Categorization." Accessed November 6.

Yu, Lei, and Huan Liu. 2004. "Efficient Feature Selection via Analysis of Relevance and Redundancy." *Journal of Machine Learning Research* 5: 1205–1224.

Modelling Exchanges in a National Blood System

John T. Blake, Dalhousie University & Canadian Blood Services; john.blake@dal.ca

Matthew Hardy, Dalhousie University & Canadian Blood Services; *matthew.hardy@dal.ca*

Canadian Blood Services distributes 850,000 units of red blood cells annually in Canada from ten distribution sites. 5-10% of all units are transhipped between sites to accommodate patient need or balance inventory. In this paper we report on the development of a simulation based method for identifying operational policies to govern transhipments between facilities in a national blood distribution network. We illustrate the use of the model through a set of experiments on the Canadian network. The results of the experiments show that while transportation costs must be traded off against product availability, improvements in operational performance can be identified through the use of such models.

1 Background

Canadian Blood Services is the not-for-profit charitable organization whose mission is to manage the supply of blood and blood products in all parts of Canada outside of the Province of Quebec. Canadian Blood Services produces and distributes approximately 850,000 units of red blood cells (RBCs) annually. These units are distributed through ten sites (nine full sized regional production and distribution facilities and one small distribution-only hub). Each distribution site (DS) services a specific geographic region and is the sole supplier of blood products for hospitals in the area. The volume of red blood cells distributed by the regional distribution sites (DS) ranges from 15,000 units to 350,000 units per annum, with the average DS supplying just over 92,000 units per year. While regions strive to be self-sufficient in blood products, blood is considered to be a national asset and therefore 5 to 10% of blood is transferred between sites. Transfers may be planned to make up for known mismatches between collections and demand or they may happen on an ad-hoc basis to move units with a rare phenotype to meet patient demand or to rebalance inventory across the network. The total cost of site-to-site transfers is in excess of $1M ($CAN) per annum.

1.1 Problem Statement

In this study, we report on a simulation-based method to evaluate site-to-site blood transfers within the Canadian Blood Services network of distribution facilities. The purpose of this study is to evaluate operational policies for inventory transfers between distribution sites to identify practices yielding low levels of shortages without undue transportation costs.

1.2 Literature

Historically, the operational research literature on blood supply chain management has been oriented towards ordering policies for a single hospital or single supplier (Blake and Hardy 2014). See Nahmias (1975) and Pastacos (1984) for early reviews or Beliën & Forcé (2012) for a more recent survey of single-supplier/single-consumer models. Literature on network planning for a blood supply chain is historically less well developed, but recently has become a topic of interest. Simulation methods are particularly common for evaluating network supply chains (Beliën and Forcé 2012). Brodheim and his co-authors describe a number of studies to set inventory levels within a regional blood distribution network under the assumption of a 21-day shelf-life. See, for example, Brodheim and Prastacos (1979). Hesse et al. (1997) describe an application of inventory management techniques to platelets in a system in which a centralized blood bank supplies 35 client hospitals. Katsaliaki and Brailsford (2007) describe the use of a simulation model to evaluate the function of a blood supply chain, but, due to complexity issues, focus on a single-producer, single-consumer system. Yegul (2007) describes the development of a custom model to evaluate inventory policies within a regional network, but does not consider inventory rebalancing. Lang (2010) uses simulation-based optimization to set inventory levels for a system in which transhipment is allowed, but tests on a small problem involving one distribution site and seven hospitals. Blake and Hardy (2014) describe a generic model of a regional blood distribution network and illustrate it‘s use for evaluating the impact of shorter shelf life for red blood cells in a multiple-supplier/multiple-consumer network, but again, do not consider inventory rebalancing. We suggest, therefore, that a model of blood flow between distribution sites in a network of national scope is a novel contribution to the literature.

2 Methods

A simulation approach was adopted to model site-to-site transfers in the Canadian Blood Services network. Once verified and validated, a set of experiments was executed to test different transfer policies. The results were analyzed using classical statistical methods.

A custom built simulation framework was developed in Microsoft VB.Net (Visual Basic). The simulation framework assumes a series of distribution sites (“suppliers”) that each collect, produce and distribute red blood cells to a collection of hospitals (“consumers”) within their catchment area. Supplier sites may exchange red blood cells as either imported or exported units.

Suppliers and consumers are modelled as separate software classes. Each class has a series of properties that together define the state of the object at any given instant and a series of methods that can be called to change or

update the object's state. The supplier and consumer objects are linked together through a simulation control algorithm. This algorithm implements a special case of the next-event, time-advance inventory model in which a fixed set of events are executed sequentially and a single, daily update is made to the simulation clock (Law 2015).

The framework assumes several distribution sites, each of which has its own set of consumer objects (hospitals) that exclusively receive products from that supplier. The supplier object contains methods that simulate the process of collecting, producing, inventorying, aging and distributing blood to consumers, as well as methods for exchanging blood with other suppliers through import/export routines. Each consumer object similarly contains methods for ordering, receiving, inventorying, and aging blood, in addition to methods for simulating patient demand.

At the beginning of each run the system is initialized; model control parameters are read in from an application database and system input data is read in from a transaction database. Supplier and consumer objects are instantiated and assigned a starting inventory, by blood group and type.

Each day, the model steps through a fixed sequence of events at suppliers and consumers. The day begins with a call to advance suppliers' inventory. This ages the stock on hand at each supplier by one day and causes any stock with -1 days of shelf-life remaining to be outdated and to leave the system. A call is then made to supplier objects to have inventory arrive from collections. Each supplier object sample from a day-of-week specific distribution that determines the total number of units that will be collected. Each unit collected is then assigned a blood group and type as well as a remaining shelf-life drawn from empirical distributions specific to that particular DS. Reductions in shelf-life of arriving units are primarily intended to represent delays in the testing process, but are also used to represent mandated reductions in shelf-life when units are irradiated to reduce the risk of graft versus host disease. Units imported from other suppliers may arrive at this time.

Once all incoming inventory is in place at the supplier, the simulation loops through each of the consumer objects and makes a call to advance the inventory. Advancing the inventory at the consumer causes the stock on hand to age by one day. Any units with -1 days of shelf-life remaining are counted as outdated units and exit the system. Each consumer object then determines if an order is required. The consumer object evaluates its inventory position, by blood group and type, and compares it to a threshold level. If the current inventory level is less than the threshold level, the consumer issues an order for additional stock to return the inventory to a target level. Consumer inventory levels are estimated from historical records; hospitals in Canada, though independent actors in the blood supply chain, report their target inventory to the supplier to assist in logistics planning. This value is used in the simulation as the upper inventory level. Inventory order triggers are estimated from historical data by subtracting from the upper inventory level the expected demand between orders at each consumer site.

The process of satisfying consumer requests requires two steps. Each supplier first evaluates, by blood group and type, the total number of units requested from all consumer sites. If there is sufficient stock on hand to meet all requests, no action is taken. However, if stock is insufficient to meet all consumer requests, the supplier will scale consumer demand so that all sites receive the same fraction of their requested amount. Scaling is done on a

prorated basis for each order. For example, if inventory on a particular day is sufficient to meet only 95% of all orders, each consumer order is scaled by 0.95. When consumer demand has been gauged and scaled, if necessary, each consumer object transmits an updated request vector to its supplier, which is then filled exactly. Scaling consumer demand ensures that any shortfalls to stock in a region are experienced equally by all consumers. This assumption mimics the supplier's published operational policy of prorating stock orders in periods of mild shortage. Upon receipt of the request vector, stock counts are decremented at the supplier and incremented at the consumer to simulate order completion.

To simulate demand for product, a call is made to each consumer object in turn to estimate patient requirements. The call generates requests for blood, using a zero-inflated Poisson (ZIP) distribution with a day-of-week specific mean value. A ZIP mixes a distribution degenerate at zero with a Poisson process and models situations with a surfeit of zeroes (Zamani and Ismail 2013). Since patient data was not available, demand was estimated from consumer shipping data, adjusted for hospital reported outdates. Unfortunately, shipping data typically contains an excess of 0 observations, representing days on which orders for a particular blood type were not issued. Accordingly, on a certain fraction of days, no demand is observed in the model. If, however, demand is to be observed, the number of units required is drawn from a Poisson distribution using the mean number of units shipped on that day of week. Once the number of units required has been determined using the ZIP distribution, blood group and type are assigned to demand items via empirical distributions specific to the consumer site. Demand is filled at the consumer FIFO from available units on the shelf. If no unit is available, the consumer site issues a demand for emergency units from its supplier. If no unit is available at the supplier, the demand is considered to be lost and counted as a shortage.

Once all demand from consumer sites has been met, suppliers may export units to other facilities. Imports and exports within the model can be considered as either "standing orders" for fixed amounts of product or "ad-hoc orders" to balance inventory across the network. Standing orders are handled in a manner analogous to regular demand between a supplier and a consumer hospital. Ad-hoc decisions between distribution sites are modelled as an instance of the transportation problem. For each blood group, a search is made across all distribution sites defined in the network. If the inventory on hand at a particular DS is above a certain threshold (for instance greater than 1 day's demand above the average network amount), that amount of inventory is declared "surplus" to need. Similarly, if the inventory on hand at another DS is below a certain threshold (for instance 2 days below the average network amount), that shortage is declared to be "demand". A transportation problem is then set up to minimize distance, subject to constraints that limit the amount of material that can or must be shipped from or to a particular site. The transportation problem is solved by using a Northwest-Corner algorithm to find an initial feasible solution and an implementation of the transportation simplex algorithm to identify the optimal solution. Once an optimal allocation has been obtained, materials are transferred between sites, with an appropriate transport delay. The simulated day ends and statistics are collected, counters are reset, time is advanced, and the daily inventory cycle begins again. See Figure 1 for a flowchart describing the simulation paradigm.

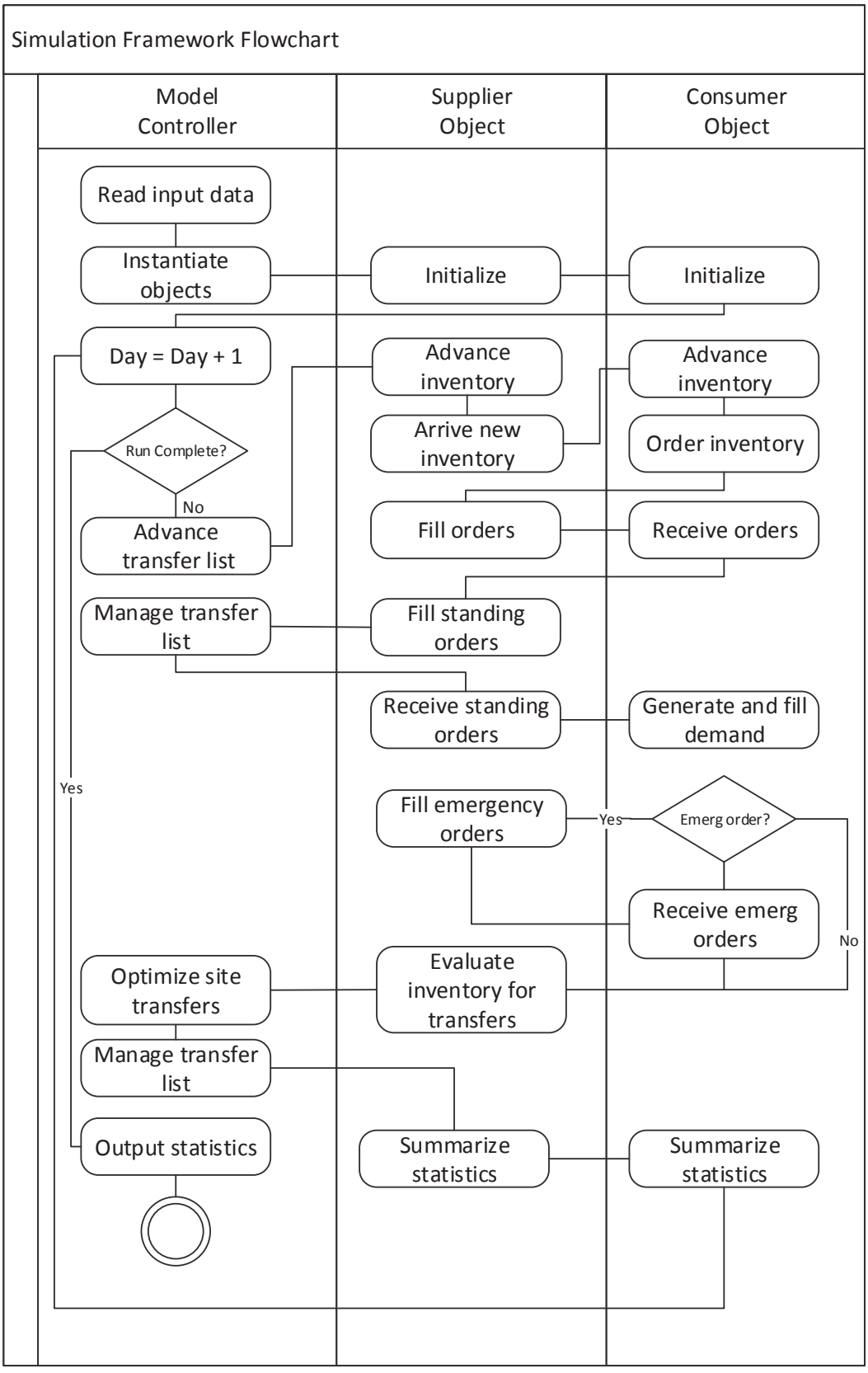


Figure 1- Flowchart describing processing simulation process flow

2.1 Data

Data for this study was derived from transaction level records extracted from Canadian Blood Services' operational database. All RBC units collected, distributed, or disposed between the periods of 18 Feb 11 and 12 May 12 (i.e. fiscal 2011-2012 with a 42 day buffer at the beginning and the end of the period) were provided. In total, the data included slightly more than 1.3 million units. Unit transfers for the period 18 Feb 11 and 12 May 12 were also obtained. This information made it possible to track all units collected, distributed, or disposed of by a Canadian Blood Services distribution site during fiscal 2011/12.

2.2 Verification and Validation

The national model was derived from an existing modelling framework, which was validated and proven to represent regional distribution networks (Blake and Hardy 2014). Thus, the veracity of the underlying conceptual model for the national simulation was assumed to follow from that work. Nevertheless, input measures were verified by ensuring that the amount of inventory collected daily at each DS, by group and type, matched that of the historical record at a 95% prediction interval. Output measures were similarly verified by comparing daily demand at each consumer site against the historical record at a 95% prediction interval. The optimal allocation of surplus units to demand sites was verified by comparing a known example of a transportation problem (Winston 2004) against the model. Once the function of the model was verified, its ability to correctly model imports and exports between sites was validated by comparing model results against historical records.

Table 1: Comparison of model results for imports and exports against historical values

	Imports			**Exports**		
	Simulation		Historical	Simulation		Historical
	Mean	Variance	Mean	Export	Variance	
Site A	56.3	7.40	56.5	0.8	0.03	1.1
Site B	38.8	7.89	35.4	22.9	2.26	21.5
Site C	1.4	0.15	0.7	60.2	1.51	58.6
Site D	2.8	0.86	4.0	37.3	1.30	35.3
Site E	23.9	2.59	23.2	3.5	1.45	3.4
Site F	4.8	0.29	5.3	2.2	0.46	2.2
Site G	31.4	1.53	30.1	4.6	1.68	4.2
Site H	0.7	0.07	0.5	34.5	10.24	34.7
Site I	12.4	0.08	12.7	6.5	0.22	7.4
Total	172.5		168.4	172.5		168.4

The results, shown in **Table 1**, indicate the model produces marginally more imports and exports than seen historically in aggregate. However, when compared on a site-by-site basis, model results were statistically indistinguishable from the historical record at a 95% level. Please note that while there are 10 distribution sites in the

Canadian Blood Services network, only the nine that are full production/distribution sites that engage in site-to-site transfers are included in the national model.

3 Results and Conclusions

A series of experiments was formulated to test the impact of both inventory thresholds and standing orders on network operations. Inventory thresholds (i.e. the level at which a DS would signal that it had surplus inventory or requirements for extra units) were varied and the volume of products shipped between sites according to standing orders was systematically altered. One set of experiments assumed a common set of thresholds for declaring inventory surplus or required, expressed in number of days on hand at each of the distribution sites. Thresholds were varied from 0 to 4 days for import (demand) limits and 0 to -4 in for export (supply) limits; standing orders were also discounted from 0 to 1 in increments of 0.25. The experimental framework consisted of 125 different scenarios, each of which was executed for a total of 10 replications of one year with a 42-day warm-up.

An analysis of variance (ANOVA) was conducted on the results of the experiment. The ANOVA indicated that import thresholds, export thresholds, and reductions to standing orders all significantly affected product availability and transport cost, the two output metrics of greatest interest. In addition, the ANOVA indicated significant two-way interactions between factors. The presence of interactions complicates the search for an ideal policy, since it implies that factors must be considered as a group. Thus, to obtain policy recommendations, a range of relative weightings between a unit shortage and the cost of transporting one unit one kilometer was tested. For each setting, the results of the experimental scenearios were weighted and the scenario with the smallest penalty was identified. The results (**Table 2**), show that three would be considered good, given the range of relative weights of a shortage compared to the cost of transporting a unit of RBC tested. Shortages, in all scenarios, were extremely rare (<1unit/day given an average demand of 2,300 units per day). As might be expected, statistically lower shortage rates were found, when compared to the baseline, when shortages were weighted heavily compared to the cost of transporting a unit. Transport cost, as measured by total distance, showed mixed results. In general, increasing product availability implies a greater transportation cost, as also might be expected. However, with a mid-range weighting of shortages (1.0E+06), model results suggested that a lower shortage rate could be achieved without a statistically significant increase in transport cost, when compared to the base case.

Table 2: Model results comparing optimal scenarios at different relative weights for a shortage. Results that are statistically different from the base line are starred.

Relative Weight of a Shortage	Export Limit	Import Limit	Reduction in Standing Orders	Shortage	Transport Cost
Base case	-	-	-	5.21E-02	1.60E+05
1.00E+08	0	0	0	1.37E-03*	2.49E+05
1.00E+06	0	-1	0.25	5.48E-03*	9.12E+04
1.00E+04	2	-1	0	1.02E-01	5.79E+04

We conclude, therefore, that it is possible to develop a model to represent a national network of blood distribution sites and to use such a model to identify the tradeoffs between product availability and site-to-site transportation costs. Moreover, we demonstrate that it is possible to use the model to identify transfer policies that yield improvements to operational performance parameters without statistically significant increases in costs. We note, however, that only a limited set of scenarios were considered in this set of experiments and thus, while the results in **Table 2**, while interesting, cannot be considered truly optimal in the mathematical sense. We suggest, therefore, that future work could include integrating the model framework with a heuristic optimization controller to explore a larger solution space.

4 References

Beliën, J, and Forcé. 2012. "Supply chain management of blood products: A literature review." *European Journal of Operational Research* 217 (1): 1-16.

Blake, JT, and M Hardy. 2014. "A generic modelling framework to evaluate network blood management policies: The Canadian Blood Services experience." *Operations Research for Healthcare* 3 (3): 116-128.

Brodheim, E, and G Prastacos. 1979. "A regional blood management system with prescheduled deliveries." *Transfusion* 19 (4): 455-462.

Hesse, SM, CR Coullard, MS Daskin, and AP Hurter. 1997. "A case study in platelet inventory management." Edited by GL Curry, B Bidanda and S Jagdale. *Sixth Industrial Engineering Research Conference Proceedings.* Norcross, GA: IIE. 801-806.

Katsaliaki, K, and SC Brailsford. 2007. "Using simulation to improve the blood supply chain." *Journal of the Operational Research Society* 58 (2): 219-227.

Lang, JC. 2010. "Blood bank inventory control with transshipments and substitutions." In *Production and Inventory Management with Substitutions*, by M Beckmann, HG Kunzi, G Fandel and W Trockel, 205-226. Berlin: Springer.

Law, A. 2015. *Simulation Modeling and Analysis, 5th Edition.* New York: McGraw-Hill.

Nahmias, S. 1975. "Optimal ordering policies for perishable inventory." *Operations Research* 23 (4): 735-749.

Prastacos, GP. 1984. "Blood inventory management: An overview of theory and practice." *Management Science* 30 (7): 777-800.

Winston, WL. 2004. *Operations Research: Applications and Algorithms.* Belmont, CA: Brooks/Cole.

Yegul, M. 2007. "Simulation analysis of the blood supply chain and a case study." Master's Thesis, Industrial Engineering, Middle East Technical University.

Zamani, H, and N Ismail. 2013. "Score test for testing zero-inflated Poisson regression against zero-inflated generalized Poisson alternatives." *Journal of Applied Statistics* 40 (9): 2056-2068.

A MACBETH-Choquet Direct Approach to Evaluate Interdependent Health Impacts

Diana F. Lopes, diana.lopes@tecnico.ulisboa.pt, CEG-IST, Centre for Management Studies of Instituto Superior Técnico, Universidade de Lisboa, Portugal

Mónica D. Oliveira, monica.oliveira@tecnico.ulisboa.pt, CEG-IST, Centre for Management Studies of Instituto Superior Técnico, Universidade de Lisboa, Portugal

Carlos A. Bana e Costa, carlosbana@tecnico.ulisboa.pt, CEG-IST, Centre for Management Studies of Instituto Superior Técnico, Universidade de Lisboa, Portugal

Simple additive value models have been used in risk management to evaluate risk sources in multiple dimensions and the MACBETH elicitation technique can be used to build value functions and weighting the dimensions based on qualitative pairwise comparison judgements given by risk managers. However, the simple additive aggregation procedure ignores value interdependencies often detected between evaluation dimensions. To address this issue, several authors in the context of industrial performance evaluation have replaced the linear additive model by the Choquet integral aggregation procedure and combined it with MACBETH for the same purposes. Alternatively, we have recently proposed the "MACBETH-Choquet direct approach". This paper presents and discusses the application of this approach to model interdependencies in the context of evaluating occupational health and safety risks, in the Occupational Health and Safety Unit of the Regional Health Administration of Lisbon and Tagus Valley at Portugal.

1 Introduction

Promoting Occupational Health and Safety (OH&S) is a world-wide challenge, as it affects all sectors and companies, and workers' accidents and poor health in the working place translate into health losses, higher use of health care services, absenteeism and thus into economic and social losses (Hughes et al., 2011). Registers from the International Labour Organisation show that a large number of individuals die per year in the European Union (EU) as a consequence of work-related accidents and occupational diseases, and many workers have their health affected by past accidents. Furthermore, new work-related diseases have been emerging in working places, such as depression. Most of these accidents and diseases can be avoidable, and the first step in preventing them is risk management. Risk management provides decision makers (DMs) with an improved understanding of risks that can threaten individuals and organizations goals (Aven, 2008) and typically requires the evaluation of impacts caused by different sources of risk. Depending on the context (either in OH&S or in other risk manage-

ment contexts), these impacts can be evaluated on a single or in multiple dimensions (Aven, 2008), such as loss of lives, absenteeism, financial impacts and environmental damage. Multicriteria value models can assist in evaluating those impacts (Linkov et al., 2006). Developing such models very often requires the identification and modelling of value interdependencies between impacts, which is a major challenge in Multiple Criteria Decision Analysis (MCDA) literature (Grabisch et al., 2010). Several studies have explored the use of CI operators to this end, with many of these studies having used an extension of MACBETH with CI operators (Clivillé et al., 2007; Merad et al., 2013). As we will briefly describe in Section 2, there are methodological challenges in these methods. In this article we aim at improving existing methods and we report how we have applied an alternative "MACBETH-Choquet direct approach" as a decision aid to model interdependent risk impacts in a real case of evaluating OH&S risks with the Occupational Health and Safety Unit (OHSU) of the Regional Health Administration of Lisbon and Tagus Valley (RHA LVT).

Section 2 of this article briefly reviews key concepts. Section 3 explains how the MACBETH-Choquet direct approach was implemented, while Section 4 presents selected results. Section 5 presents concluding remarks.

2 Review of studies

Proper management of OH&S risks has been recognized as a social concern and as a workers' right by many organizations, including the International Labour Organization and national governments, in line with promoting a healthy and productive labour force (Froneberg, 2005). Risk matrices (RMs) are one of the tools most widely used to evaluate risks in general (Oliveira et al., 2014) and OH&S risks in particular (National Patient Safety Agency, 2008; Administração Regional de Saúde de Lisboa e Vale do Tejo, 2010), as they are easy to handle, demand for limited expertise, have a straightforward interpretation, allow for performing a quick analysis (IEC/FDIS 31010, 2009) and are recommended by international standard rules. Nonetheless, several studies point out that RMs violate theoretical principles that compromise their feasibility and use (Bricknell et al., 2007; Oliveira et al., 2014), with some problems being the misuse of rating scales for impacts (which leads to quantitatively meaningless ratings) and a disregard of the cumulative effects of multiple impacts.

A proper modelling of the effects of multiple impacts on a common scale is required for a proper use of RMs (as well as for other risk evaluation tools), and very often interdependencies between impacts in multiple dimensions exist. In this study we will focus on the evaluation of impacts caused by different types of risks when interdependencies between impacts are observed (and as framed in the context of evaluating the impacts of OH&S risks). A review of studies on this topic has shown that a lot of research has been devoted to the topic of modelling interdependencies. The CI has been used for that purpose in several evaluation contexts, such as: (i) industry (Clivillé et al., 2007); (ii) education (Cardin et al., 2013); and (iii) risk management (Vernadat et al., 2013). Most of these studies have combined the 2-additive CI operator with MACBETH to model interdependencies in real contexts, as it presents a good compromise between complexity and richness in modelling interdependencies (Grabisch et al., 2010).

MACBETH is a MCDA approach that requires only non-numerical judgments about differences in attractiveness between options to help the DMs measure the value of options or the impact value of risks (Bana e Costa et al., 2012). MACBETH allows the building of a meaningful impact rating scale and can take into consideration cumulative effects of multiple impacts. Previous studies have shown that MACBETH provides a simple and transparent approach in modelling complex multidimensional problems with a user-friendly protocol, and hence its wide applicability (Omann, 2003; Ferreira et al., 2014). Although many MACBETH applications use additive models that respect the "difference independence" conditions (working hypothesis), many studies have also shown the case for using MACBETH with the 2-additive CI operator (1) - a generalisation of the linear model - to handle interdependencies:

$$\begin{cases} V_{Ag}(u) = \sum_{i=1}^{m} s_i v_i(x_{iu}) - \frac{1}{2} \sum_{\substack{\{i,j\} \\ i \neq j}} I_{ij} \, | \, v_i(x_{iu}) - v_j(x_{ju}) \, | \\ \sum_{i=1}^{m} s_i = 1 \end{cases} \quad (1)$$

where V_{Ag} represents the aggregated value of an option u, v_i corresponds to the overall value of the option u on the dimension i, considering a specific baseline in the other dimensions; $x_{(i)u}$ represents the impact level of the option u in the dimension i; s_i corresponds to the Shapley parameter of the dimension i; and I_{ij} the interaction parameter between the dimensions i and j – for further details, consult (Lopes et al., 2014a).

Previous studies applying the MACBETH approach with the CI made a key contribution to the modelling of interdependencies, but there is scope for improving these methods. For instance, these studies use questioning protocols that ask for local judgements (i.e., for value judgements in separate dimensions) – whereas, whenever possible, it is preferable to ask for global judgments (i.e., which account for impacts in several dimensions) –, and make a limited use of functionalities of the MACBETH approach (for further details, consult (Lopes et al., 2014a)).

In this study we report how we have implemented and applied the MACBETH-Choquet direct approach, designed by (Lopes et al., 2014a), to model interdependencies of OH&S risk impacts. This study was carried out within the following context: the OHSU is responsible for evaluating OH&S risks for all the individuals working in primary care centres and in administrative offices of the RHA LVT. Following some difficulties felt by the OHSU team in the evaluation of OH&S risk impacts, we have built value-risk matrices (VRMs) (Lopes et al., 2014b)) as an alternative to RMs within the project IRIS - Improving RIsk matriceS using multiple criteria decision analysis (Oliveira et al., 2014). Developing VRMs required converting risk impacts in multiple dimensions into a common scale of value (for further details, consult (Lopes et al., 2014a)), and in this article we only report how the MACBETH-Choquet direct approach was implemented and applied to build such a scale within an interactive learning process with our DMs (which in our case are all members of the OHSU). Up to our knowledge, this is the first study modelling interdependencies with the Choquet Integral and MACBETH in the health care context.

3 Methods

Before testing and modelling interdependencies, the relevant dimensions for appraising risk impacts were structured with our DMs. This led to the value tree depicted in Figure 3.1 that considers three impact dimensions – "employee's health" (EH), "work capability" (WC) and "absenteeism" (AB) – that were operationalised through different descriptors of impact, namely healthy years of life lost (for EH), a five level qualitative scale (for WC) and the number of days/years of absenteeism (for AB).

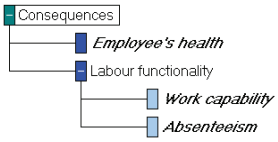
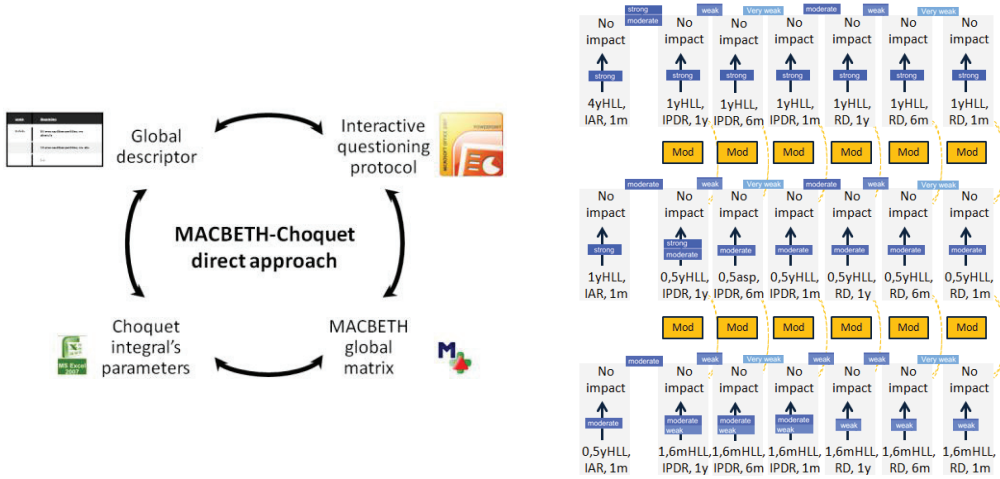


Figure 3.1:
Final value tree built DMs (key dimensions in italic).

We then proceeded in testing whether preference independence conditions were respected. This was done by adopting a structured questioning protocol based on specific questions such as: "what is the attractiveness of reversing the impacts of a risk that implies 15 years (1.6 months) of healthy life lost (HLL), irrecoverable partial disability with return to work and 1 year of absence from work (AW) and another risk that leads to 15 years (1.6 months) of HLL, recoverable disability and 1 year of AW"? Note that both swings depict the case of going from a lower reference level to an upper reference level in the second dimension – work capability –, when different levels of the first dimension are fixed. Our DMs judged the first swing as "strong" and the second one as "weak", which means that the "work capability" is cardinally dependent with "employee's health". Following a similar protocol, we have concluded that "absenteeism" was also cardinally dependent with "employee's health", and that preference independence conditions were not respected, thus being appropriate to apply the MACBETH-Choquet direct approach.

Implementing this approach required the design of a socio-technical process that involved the sequence of interconnected activities depicted in Figure 3.2(a). The social component consisted in various meetings and decision conferences held with our DMs, which enabled the development of a requisite model of risk impacts (with model "requisiteness" as defined by (Phillips, 1984)). On the technical side, we designed the following steps to apply the MACBETH-Choquet direct approach: (1) defining a global descriptor of impacts that considers all plausible combinations of the impact levels of the identified dimensions; (2) adopting a specific questioning protocol supported by the Microsoft Office PowerPoint 2007 (MOP) to compare in sequential and interactive way the desirability of the plausible combinations of impacts that only asks for non-numerical judgments; (3) populating a single MACBETH global matrix that assists in proposing and analysing numerical scales compatible with the judgments given by the DMs; and (4) using of the Microsoft Office Excel 2007 (MOE) for determining the CI parameters from the information given by the MACBETH Decision Support System (DSS).



(a) (b)

Figure 3.2:

(a) Activities developed for implementing and applying the MACBETH-Choquet direct approach at the OHSU (b) and screen of the interactive questioning protocol adopted with the OHSU.

For defining a global descriptor of impacts, feasible combinations of different impact levels across all the three dimensions were defined – in our example, 36 combinations of impacts, varying from the *status quo* corresponding to no impact (0 HLL, null disability and 0 days of AW) to the worst level corresponding to states equivalent to death (34 YHLL, irrecoverable disability (ID) and 18 years of AW) were considered.

Then, using the application of the MACBETH-Choquet based elicitation protocol, the DMs were asked to quantify the relative attractiveness of reverting each set of impacts through a qualitative pairwise comparison questioning mode, demanding for a decision support system. This was done with a MOP application that enabled applying the questioning protocol in a sequential and interactive manner (using the *ActiveX textbox* tool) – see Figure 3.2(b). To illustrate the protocol, the DMs were asked about the attractiveness of fully reversing the following combination of impacts – 34 HLL, ID and 18 years of AW – in the MACBETH categorical scale "null", "very weak", "weak", "moderate", "strong", "very strong" or "extreme". This same type of questioning protocol was applied to all the combinations of risk impacts (as explained in (Bana e Costa et al., 2008)).

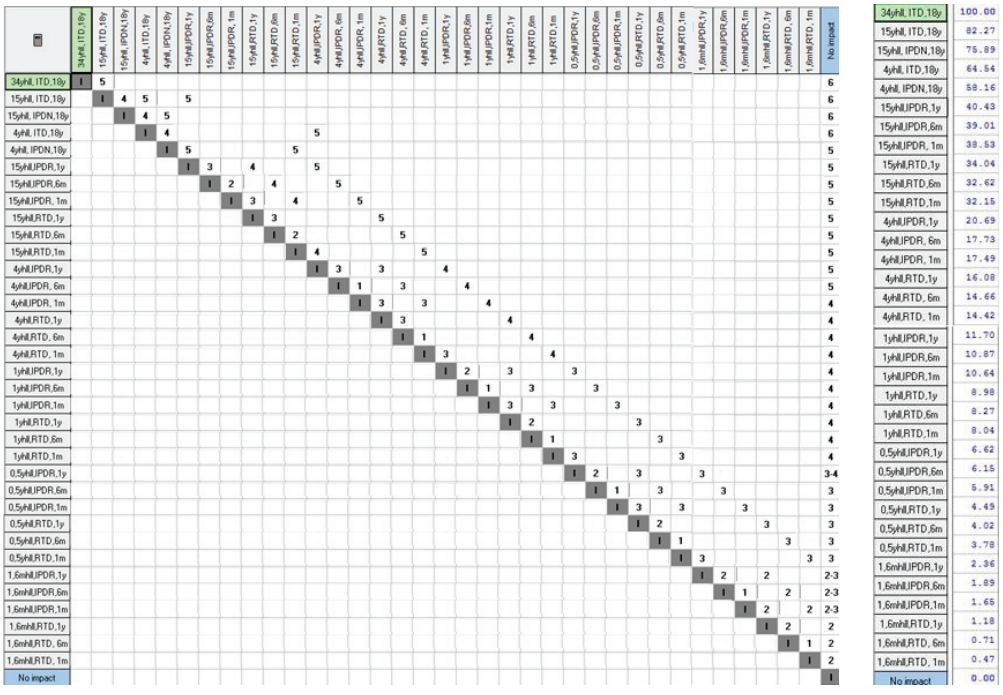
Using those answers, a single MACBETH global matrix with qualitative judgements was populated, validated, and then M-MACBETH assisted in proposing and analysing a numerical scale to be validated by the DMs. MOE was finally used to calculate the CI parameters, and thus the underlying model to evaluate risk impacts.

4 Application results

Figure 4.1(a) presents the MACBETH matrix of qualitative judgements that gathers all the final judgments from the DMs to evaluate distinct combinations of impacts (note that after detection of inconsistencies and group

discussions, some judgments have been revised). Then, M-MACBETH DSS assisted on generating a global (numerical) scale that was finally discussed, adjusted and validated, leading to the numerical scale displayed in Figure 4.1(b). The scale can be read as follows: a risk event that implies a chronic back pain that leads to 4 YHLL, a irrecoverable total disability (ITD) and 18 years of absenteeism corresponds to a score of 64.54, while a risk event that implies a worker's death corresponds to a score of 100. These scores will be used by OHSU to assess risk impacts within a VRM framework.

Based on the global information, the following CI parameters were obtained: the Shapley parameters, $s_{EH}=295/338$, $s_{WC}=32/338$ and $s_{AB}=32/338$, and the interaction parameters, $I_{EH\&WC}=22/169$, $I_{EH\&AB}=5/169$ and $I_{WC\&AB}=I_{EH\&WC\&AB}=0$. These parameters are in line with the expectations of our DMs. The EH dimension is the one that mostly contributes for the aggregated impact value, which matches the DMs preferences. Regarding the interaction parameters, there is a synergism between dimensions EH and WC, and EH and AB, i.e., according to the DMs, the combined impact is valued more than the sum of individual impacts (given by the positive interaction parameters between the dimensions mentioned above). On the other hand, there is no interaction between the three dimensions altogether, neither between the dimensions WC and AB.



(a) (b)

Figure 4.1:

(a) Global MACBETH matrix of judgments; (b) corresponding numerical scale validated by the DM (b).

5 Conclusions

There is scope for developing tools to assist the modelling of interdependent dimensions for decision aid, and this article reports the application of the MACBETH-Choquet direct approach to improve the evaluation of impacts of OH&S risks. The proposed approach asks for global judgments (of combinations of risk impacts) and enables for detecting inconsistent judgments, for modelling cases of differences in opinion or hesitation, promoting more robust results. It also helps overcoming the problems identified in a very popular tool, RMs, by building a meaningful impact rating scale that considers the effects of multiple impacts. For applying that approach to a real evaluation case – which aimed at building a scale to evaluate OH&S risk impacts within the RM framework –, we have developed visual interactive preference modelling system that combined several DSS: MOP assisted in applying the questioning protocol in a sequential and interactive way, MACBETH supported global preference modelling, and MOE enabled the determination of the CI parameters.

The members of the OHSU found it easy to answer to the MACBETH-Choquet questioning protocol and validated the numerical value scale, and reported that developing this scale has helped them to have a better understanding of the impacts of OH&S risks. There is scope for developing further research to assist risk management in general, and the evaluation of OH&S risks in particular – for instance, exploring the compatibility of the proposed approach with other mathematical formulations, such as the multilinear one, and improving DSS tools.

Acknowledgements

This work was funded by National Funds from the Portuguese Public Budget through FCT – *Fundação para a Ciência e a Tecnologia,* within the project PTDC/EGE-GES/119230/2010.

References

Administração Regional de Saúde de Lisboa e Vale do Tejo *Segurança e Saúde no Trabalho: Gestão do risco profissional em estabelecimentos de saúde*. Orientações técnicas nº1, 2010.

Aven, T. *Risk analysis: assessing uncertainties beyond expected values and probabilities*. Wiley, West Sussex, 2008.

Bana e Costa, C.A., De Corte, J.-M. and Vansnick, J.-C. *MACBETH*. International Journal of Information Technology and Decision Making, 11(2), 359-387, 2012.

Bana e Costa, C.A., Lourenço, J.C., Chagas, M.P. and Bana e Costa, J.C. *Development of reusable bid evaluation models for the Portuguese Electric Transmission Company*. Decision Analysis, 5(1), 22-42, 2008.

Bricknell, M. and Moore, G. *Health risk management matrix - a medical planning tool*. J. R. Army Med. Corps, 153(2), 87-90, 2007.

Cardin, M., Corazza, M., Funari, S. and Giove, S. *Building a Global Performance Indicator to Evaluate Academic Activity Using Fuzzy Measures*. *Neural Nets and Surroundings*. Springer Berlin Heidelberg, B. Apolloni, S. Bassis, A. Esposito and F. C. Morabito, 217-225, Smart Innovation, Systems and Technologies, 2013.

Clivillé, V., Berrah, L. and Mauris, G. *Quantitative expression and aggregation of performance measurements based on MACBETH multicriteria method*. International Journal of Production Economics, 105(1), 171-189, 2007.

Ferreira, F., Santos, S.P., Marques, C. and Ferreira, J. *Assessing credit risk of mortgage lending using MACBETH: A methodological framework*. Management Decision, 52(2), 1-36, 2014.

Froneberg, B. *Challenges in occupational safety and health from the global market economy and from demographic change — facts, trends, policy response and actual need for preventive occupational health services in Europe*. Scand. J. Work Environ. Health(1), 23-27, 2005.

Grabisch, M. and Labreuche, C. *A decade of application of the Choquet integral and Sugeno integrals in multicriteria decision-aid*. Annals of Operations Reasearch, 175(1), 247-290, 2010.

Hughes, P. and Ferrett, E. *Introduction to Health and Safety at Work*. Elsevier, 2011.

IEC/FDIS 31010 *Risk management - Risk assessment techniques*, 2009.

Linkov, I., Satterstrom, F.K., Kiker, G., Batchelor, C., Bridges, T. and E., F. *From comparative risk assessment to multi-criteria decision analysis and adaptive management: Recent developments and applications*. Environment International, 32(8), 1072–1093, 2006.

Lopes, D.F., Bana e Costa, C.A., Oliveira, M.D. and Morton, A. *Using MACBETH with the Choquet Integral fundamentals to model interdependencies between elementary concerns in the context of risk management*. *Proceedings of the 3rd International Conference on Operations Research and Enterprise Systems, ICORES 2014*, ESEO, Angers, Loire Valley, France, SCITEPRESS, 116-126 (digital edition), 2014a.

Lopes, D.F., Oliveira, M.O. and Bana e Costa, C.A. *Occupational health and safety: Designing and building with MACBETH a value risk-matrix for evaluating health and safety risks*. Manuscript in preparation, 2014b.

Merad, M., Dechy, N., Serir, L., Grabisch, M. and Marcel, F. *Using a multi-criteria decision aid methodology to implement sustainable development principles within an organization*. European Journal of Operational Research, 224, 603-613, 2013.

National Patient Safety Agency). *A risk matrix for risk managers*. Cited December 2014, Available from http://www.npsa.nhs.uk/nrls/improvingpatientsafety/patient-safety-tools-and-guidance/risk-assessment-guides/risk-matrix-for-risk-managers/, 2008.

Oliveira, M.O., Bana e Costa, C.A. and Lopes, D.F. *Improving risk matrices with MACBETH*. Manuscript submitted for publication, 2014.

Omann, I. *Product service systems and their impacts on sustainable development: A multi-criteria evaluation for austrian companies*. Frontiers, 1-34, 2003.

Phillips, L. *A theory of requisite decision models*. Acta Psychol, 56(1-3), 29-48, 1984.

Vernadat, F., Shah, L., Etienne, A. and Siadat, A. *VR-PMS: a new approach for performance measurement and management of industrial systems*. International Journal of Production Research, 1-19, 2013.

A decomposition approach for the analysis of discrete-time queuing networks with finite buffers

Dr.-Ing. Judith Stoll, Judith.stoll@kit.edu, Institute for Material Handling and Logistics (IFL), Karlsruhe Institute of Technology (KIT)

Abstract: This paper describes an approximation procedure for determining the throughput time distribution of material handling and service systems modeled as queuing networks. We consider finite buffer capacities and general distributed processing times in terms of discrete probability functions. In order to quantify the influence of blocking caused by finite buffers, we present a decomposition approach. Two-server subsystems of the queuing network are analyzed subsequently to obtain the throughput time distribution of the whole queueing system. The quality of the presented approximation procedure is tested against the results of various simulation experiments.

1 Introduction and Problem Description

Health establishments are faced with growing health expenses, while the society tries more and more to reduce health expenses in order to guarantee its social welfare. This leads to the search of efficiency in order to limit the increase of expenses, which is a well known problem in the context of logistics systems. Therefore, some authors started with the performance analysis of support services in health establishments, e.g. the sterilization of medical devices. Starting with a simulation model of a generic sterilization process (see Di Mascolo et al. 2009), analytical methods which were previously used to analyse the material flow in production systems, were applied to analyze the sterilization service in health establishments (see Stoll and Di Mascolo 2013). Discrete time queueing models have been used as a method to achieve a fast and quite accurate way of determining performance figures of service systems. While with classical general queuing models characteristic values are calculated only on the basis of means and variances, in discrete-time modeling all input and output variables are described with discrete probability distributions.

Grassmann and Jain (1989) were the first who analyzed the waiting time distribution in a G|G|1 queuing system in discrete time domain. In addition to their model, several basic elements for network analysis have been treated analytically which can now be combined (Stochastic Finite Elements, see publications of Furmans, Zillus, Schleyer, Matzka/Stoll and Özden between 1996 and 2012). For an overview of the existing discrete-time models see Matzka (2011). The advantage of these models lies in the fact, that they allow the computation of not

only averages but also the distributions of e.g. waiting times. This enables the derivation of quantiles of performance measure, which are often needed for the design of logistics systems.

All the existing discrete time queuing models with general distributions assume infinite buffers. This assumption is not valid for many practical cases. In the sterilization area of health establishments for example, buffers are limited, too. The space for buffering medical devices between the single sterilization steps is limited as there is a given number of racks or boxes to buffer the devices, before the next process step can start (see Stoll and Di Mascolo 2013). If one of these buffers is full, this can cause blocking in upstream process steps. In real material handling and service systems, finite buffers cause blocking situations which can be classified in three main types of blocking: Blocking After Service (BAS), Blocking Before Service (BBS) and Repetitive Service Blocking (RS) (Onvural 1990). For an overview on queuing networks with blocking see Perros and Altiok (1994), Balsamo et al. (2001) and Manitz and Tempelmeier (2012). To our knowledge there is no paper dealing with the analysis of queuing networks with blocking with general distributed service times given in terms of discrete probability functions. Existing models are either dealing with two-parameter approximations or known discrete probability functions, e.g. the binomial distribution. Thus, we will now present a method that enables us to quantify the influence of blocking after service in discrete time queuing networks with general distributed service times.

The paper is organized as follows: First we will give a short introduction to discrete-time modeling in section 2. In section 3 we give an overview of discrete time modeling of the sterilization process in health establishments. In section 4, we present the approximation method for the calculation of the waiting time distribution of blocked queuing systems in discrete time domain. As the presented approach is an approximation we want to give some insights to the quality of this approximation comparing analytical results to simulation results (section 5). In section 6 we give a conclusion and an outlook on further research on this topic.

2 Discrete-time modeling

Analysis in discrete time domain assumes that time is not continuous but discrete. This means, that events are only recorded at discrete moments which are multiples of a constant increment t_{inc}. These events occur, when items are moved or when they change their status, for example by entering a queue, by being served, by merging with a stream of other items or at a split of a stream. In our analysis, events are described by a discrete random variable. When we have given a discrete random variable X, we denote its distribution, which is also called probability mass function (pmf), by

$$P(X = i \cdot t_{inc}) = x_i \quad \forall\, i = 0,1,\dots,i_{max} \qquad (1)$$

As a simplification we reduce this notation to

$$P(X = i) = x_i \quad \forall\, i = 0,1,\dots,i_{max} \qquad (2)$$

When we talk about a distribution in the subsequent sections, we refer to the probability mass function.

3 Sterilization process in health establishments

In a sterilization process, reusable medical devices are re-injected in the process after their use in the operation room. When we integrate the use step, the sterilization process becomes a sterilization loop, with the following steps: use, pre-disinfection (including the transfer from the operating rooms to the sterilization service), rinsing, washing, verification, packing, sterilization, transfer from the sterilization service to the operating rooms, storage, before a new use (Di Mascolo et al. (2006)).

In order to improve the performance of the system, different scenarios for the transfer of medical devices to the sterilization area can be analyzed. In practice, the transport is normally not following a certain rule. This unsteadiness can cause a duration of pre-disinfection that does not lead to the desired effect on the material. Modeling the different possibilities of transport organizations would enable us to compare their performance and especially show the impact that a modification of the transport would have.

In Stoll and Di Mascolo (2013), we used a discrete time queueing network model to analyze some performance figures of a particular sterilization process. In a previous work, Di Mascolo et al. (2006) analyzed a specific health establishment via simulation. We used the same input data to compare our queueing network model to the accordant simulation model.

From the queueing model, we obtained two important performance figures that we compared to the simulation results. One of them is the average duration of the pre-disinfection step. The ideal duration of pre-disinfection, to guarantee an optimal impact of the disinfection liquid to the medical devices, is about 15 minutes. On the other hand, the sojourn time in the liquid should not exceed 50 minutes, because the disinfection product attacks the material, and thus causes a premature ageing. We thus want to know the average pre-disinfection time as well as the percentage of medical devices that stay in the liquid more than 50 minutes. Compared to simulation, our queueing network model obtains quite good results for the average pre-disinfection time (simu: 29.40 min, analysis: 30.50 min) and the percentage of medical devices, that stay in the liquid more than 50 minutes (simu: 92.90 %, analysis: 91.46 %). The deviations are caused by the fact that we had to make some assumptions in the queueing network model.

A second important parameter is the throughput time of medical devices through the sterilization process before they are ready to be used again. From a survey in the Rhone-Alpes region, we know that many health establishments are not able to estimate the duration of their sterilization process (see Di Mascolo et al. (2006)). The average throughput time calculated by our queueing network model can give them an idea of the cycle time of medical devices. This figure also allows us to compare different loading policies for the washers and autoclaves and the influence of the number of parallel machines to the throughput time in discrete time queueing networks with general distributed processing times.

All the existing discrete time queuing models with general distributions we used for the sterilization model assume infinite buffers. This assumption is not valid in many cases. In the sterilization area of health establish-

ments, buffers are limited, too. The space for buffering medical devices between the single sterilization steps is limited as there is a given number of racks or boxes to buffer the devices, before the next process step can start. Thus, it is our intention to get an approach that helps us to quantify the influence of blocking on the throughput time.

4 Decomposition approach

Let us regard a queuing network, consisting of several queuing systems in series. For each node of the network, the service time distribution is given in terms of a general discrete probability function by

$$P(B = \beta) = b_\beta \quad \forall\, \beta = 0,1,\dots,\beta_{max} \qquad (3)$$

If the buffers in front of the servers would be infinite, we could use the methods of Grassmann and Jain (1988 and 1989) to calculate the waiting time distribution for each queuing system and the inter departure time distribution that connects the nodes of the network. The distribution of the number of customers in the system could be calculated according to the method of Furmans and Zillus (1996). As we assume the buffers to be finite, we have to consider blocking of servers if a succeeding buffer is full. This has an influence on the waiting time of jobs in the buffer in front of a blocked server.

In order to consider the influence of finite buffers in discrete-time queuing networks we propose the following decomposition approach. We define subsystems of a queuing network consisting of two queuing stations in series (see figure 1). Starting from the most downstream server, we quantify the influence of blocking to the waiting time at the upstream server(s).

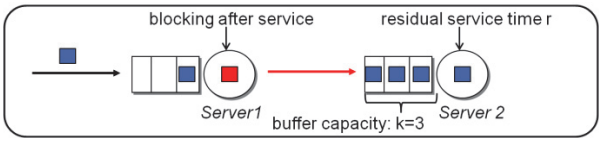


Figure 1:
Two-server subsystem with blocking after service

The first server (upstream server) is modeled as a discrete time G|G|1 queuing system with infinite buffer and the second server (downstream server) is modeled as a G|G|1-k queuing system with a finite buffer capacity k, which means that a maximum of k units can wait in the queue. When the buffer of the second queuing system is full, a material unit that is finished cannot leave server 1 after service. The upstream server is blocked and stops further processing. We analyze the influence of blocking to the sojourn time of a job in the upstream server, and thus the waiting time of jobs in the upstream buffer. We define sojourn time in this context as the time, the job spends in the server, that is service time and blocking time.

We distinguish three cases for the sojourn time distribution of jobs in an upstream server:

1. If the downstream buffer is not full (number of customers in the queuing system N< k+1), when a job of the upstream server is finished and wants to leave the server, there is no blocking and the job enters the downstream buffer. The sojourn time is then equal to the service time of server 1, and the distribution of sojourn time ($\vec{s}$) is equal to the distribution of the service time of server 1 ($\vec{b}_1$): $\vec{s} = \vec{b}_1$.

2. If at the departure of a job, the succeeding buffer is full (N=k+1), the server is blocked and the sojourn time of the job increases by the residual service time of server 2 and the distribution of the sojourn time can be calculated by convoluting the service time distribution of server 1 and residual service time distribution of server 2 ($\vec{r}_2$): $\vec{s} = \vec{b}_1 \otimes \vec{r}_2$.

3. After blocking, the next job in the upstream server and the next job in the downstream server start simultaneously. If the service time β_1 of the job in server 1 is shorter than the service time β_2 of the job in server 2 ($\beta_1 \leq \beta_2$), the upstream server is blocked again and the sojourn time of the job in server 1 is equal to the service time in server 2 ($\vec{s} = \vec{b}_2$). In the opposite case, server 1 is not blocked and the sojourn time distribution in server 1 is equal to his service time distribution ($\vec{s} = \vec{b}_1$).

The three cases build a closed markov chain, where the states are given by the three different cases to calculate the sojourn time of a customer in the upstream server (see figure 2). The transitions between the system states can be interpreted as follows:

If a job was finished without blocking, his sojourn time was $\vec{s} = \vec{b}_1$. The following job also has a sojourn time of $\vec{s} = \vec{b}_1$, if the number of customers N in the succeeding queuing system is smaller than k+1 in his departure moment. If there are k+1 customers in the system, the queue is full and the sojourn time increases to $\vec{s} = \vec{b}_1 \otimes \vec{r}_2$. Note that we calculated the number of customers in the system using the method of Furmans and Zillus (1996) for infinite buffers, and thus have to normalize the probabilities with $P(N \leq k+1)$. If a job was blocked and had a sojourn time of $\vec{s} = \vec{b}_1 \otimes \vec{r}_2$, the following job starts his service simultaneously with the job in server 2. If the service time of the job in server 1 is smaller than the service time in server 2 ($\beta_1 \leq \beta_2$), the sojourn time of server 1 equals the service time of server 2. In the opposite case, the sojourn time is equal to the service time of server 1. When the sojourn time of server 1 equals the service time of server 2, the following job also starts his service simultaneously with the job in server 2 and the same transitions are valid.

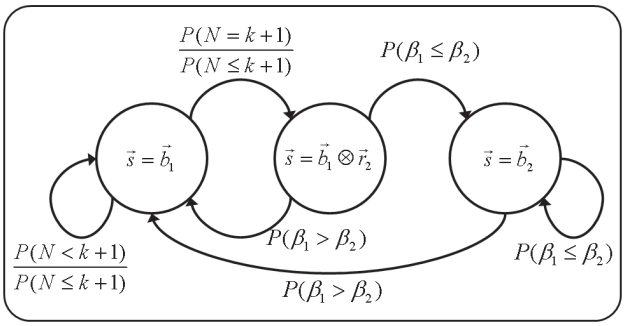


Figure 2:

Markov-chain with sojourn times for succeeding customers

From the above shown markov chain, we get a set of linear equations that leads to the following system state probabilities:

$$\mathrm{P}\left(\vec{\mathrm{s}}=\vec{\mathrm{b}}_1\right)=1-\frac{\mathrm{x}-\mathrm{xy}}{\mathrm{x}+\mathrm{y}}-\frac{\mathrm{x}}{1+\mathrm{x}}\left(1+\frac{\mathrm{x}-\mathrm{xy}}{\mathrm{x}+\mathrm{y}}\right) \tag{4}$$

$$\mathrm{P}\left(\vec{\mathrm{s}}=\vec{\mathrm{b}}_1\otimes\vec{\mathrm{r}}_2\right)=\frac{\mathrm{x}}{1+\mathrm{x}}\left(1+\frac{\mathrm{x}-\mathrm{xy}}{\mathrm{x}+\mathrm{y}}\right) \tag{5}$$

$$\mathrm{P}\left(\vec{\mathrm{s}}=\vec{\mathrm{b}}_2\right)=\frac{\mathrm{x}-\mathrm{xy}}{\mathrm{x}+\mathrm{y}} \tag{6}$$

with $$\mathrm{x}=\frac{\mathrm{P}(\mathrm{N}=\mathrm{k}+1)}{\mathrm{P}(\mathrm{N}\le\mathrm{k}+1)} \tag{7}$$

and $$\mathrm{y}=\mathrm{P}(\beta_1>\beta_2) \tag{8}$$

Using the analytical method of Furmans and Zillus, we can calculate the system state probabilities (system state = number of customers in the system at the arrival of a customer) of a G|G|1 queuing system with infinite buffers. Thus, we can approximate the blocking probability $P(N=k+1)$.

The probability that the service time of server 1 is higher than the service time of server 2 can be calculated according to the following formula:

$$P(\beta_1>\beta_2)=\sum_{i=0}^{\beta_{1,max}}\sum_{j=0}^{i-1} b_{1,i}\,b_{2,j} \tag{9}$$

Knowing the probabilities for the sojourn times of a job in a server in each of the three cases, we can calculate a service time distribution of server 1 as follows:

$$\vec{b}_{mod} = P(\vec{s} = \vec{b}_1) \cdot \vec{b}_1 \qquad (10)$$
$$+P(\vec{s} = \vec{b}_1 \otimes \vec{r}_2) \cdot \vec{b}_1 \otimes \vec{r}_2 + P(\vec{s} = \vec{b}_2) \cdot \vec{b}_2$$

With this modified service time distribution we can calculate the waiting time distribution of queueing system 1 according to the method of Grassmann and Jain (1989). The number of customers in the system can again be calculated by the method of Furmans and Zillus (1996). We then declare server 1 as the downstream server of a new subsystem. These steps are repeated until the first node of the network is reached. Knowing the waiting time distributions and service time distributions of each node, the throughput time distribution of the complete network can be calculated.

5 Analysis of the approximation quality

As the presented approach is an approximation we want to give some insights to the quality of this approximation. For an example with two succeeding servers, we can compare the waiting time distribution calculated with the analytical approach to the distribution obtained by simulation as shown in figure 3.

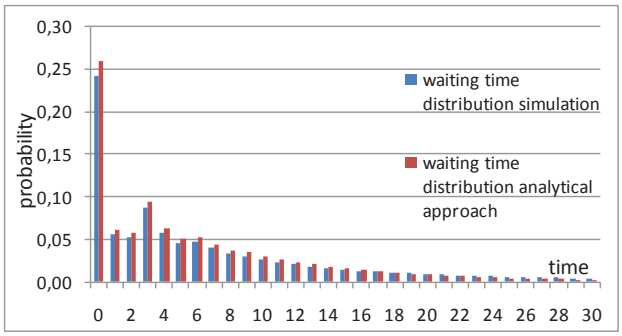


Figure 3:
Waiting time distribution in queuing system 1 obtained by analytical approach and simulation

In order to make the experiments comparable, we calculate the deviation of the mean waiting time of analysis and simulation and show its dependence to the blocking probability. The blocking probability is influenced by the following parameters: number of buffer spaces k, variability of departure process from server 1, variability of the service process of server 2. Table 1 shows the results for a couple of experiments.

blocking prob.	mean waiting time analytical approach	mean waiting time simulation	relative deviation
0.0055%	0.82	0.82	0.28%
0.2705%	0.83	0.82	0.52%
1.7489%	13.76	13.58	1.32%
5.6581%	14.72	14.21	3.58%
12.0475%	6.49	9.52	-31.83%
22.2587%	7.69	12.47	-38.34%
41.1258%	11.52	22.95	-49.83%

Table 1:
approximation quality depending on the blocking probability

We can see that the approximation quality is quite high for low blocking probabilities. The higher the blocking probability gets, the higher are the deviations between analytical results and simulation results. In practice, blocking probabilities of more than 5 % would not be tolerated. The buffer spaces would be increased or the process would be improved to be more stable. Thus, for practical applications, the analytical approach is useful to analyze the waiting time distribution in front of a server with blocking.

6 Conclusion and Outlook

In this paper we presented an analytical approach for the analysis of blocking after service in discrete time queuing networks. For the analysis we build two-server subsystems and approximate the influence of blocking by a modified service time distribution for the blocked server. The approximation quality is good for low blocking probabilities and therefore applicable for practical cases. Our next step will be to apply the new method to our model of the sterilization process in health establishments replacing the model elements with infinite buffers by finite buffer.

7 Acknowledgment

This research is supported by the research project “Analytische Berechnung von Durchlaufzeiten in komplexen Stetigfördersystemen”, funded by the Deutsche Forschungsgemeinschaft (DFG) (reference number FU 273/12-1)

References

Balsamo, S., V. de Nitte Personé and R. Onvural: *Analysis of queueing networks with blocking*. Norwell, MA: Kluwer Academic Publishers, 2001

Di Mascolo, M., A. Gouin, K. Ngo Cong. 2009. *A generic model for the performance evaluation of centralized sterilization services*. Proccedings of the conference on Stochastic Models of Manufacturing and Service Operations, SMMSO 2009: Lecce, Italy.

Furmans, K. and A. Zillus: *Modeling independent production buffers in discrete time queueing networks*. In: Proceedings of CIMAT '96, Grenoble, p. 275–280, 1996

Grassmann, W. K. and J. L. Jain: *Numerical Solutions of the Waiting Time Distribution and Idle Time Distribution of the Arithmetic GI/G/1 Queue*. Operations Research 37 (1), p. 141–150, 1998

Jain, J. L. and W. K. Grassmann: *Numerical solution for the departure process from the GI/G/1 queue.* Computers & OR 15(3), p. 293–296, 1988

Manitz, M. and H. Tempelmeier: *The variance of inter-departure times of output of an assembly line with finite buffers, converging flow of material, and general service times.* OR Spectrum 34, p. 273-291, 2012

Matzka, J.: *Discrete Time Analysis of Multi-Server Queueing Systems in Material Handling and Service.* Dissertation, Karlsruhe Institute of Technology, 2011

Özden, E. and K. Furmans: *Analysis of the discrete-time GX|G[L,K]|1-queue.* In: 24th European Conference on Operational Research, Lisbon 2010

Özden, E.: *Discrete time Analysis of Consolidated Transport Processes.* Dissertation, Karlsruhe Institute of Technology, 2011

Onvural, R. R.: *Survey of closed queueing networks with blocking.* ACM Computing Surveys 22(2), p. 83-121, 1990

Perros, H. G. and T. Altiok: *Queueing networks with blocking.* New York, NY, USA: Oxford University Press, Inc., 1994

Schleyer, M.: *Discrete time analysis of batch processes in material flow systems.* Dissertation, Universität Karlsruhe, Institut für Fördertechnik und Logistiksysteme, 2007

Schleyer, M.: *An analytical method for the calculation of the number of units at the arrival instant in a discrete time G/G/1-queueing system with batch arrivals.* OR Spectrum 34(1), pp 293-310, 2012

Schleyer, M. and K. Furmans: *An analytical method for the calculation of the waiting time distribution of a discrete time G/G/1-queueing system with batch arrivals*. OR Spectrum 29(4), p. 745–763, 2007

Stoll, J. and M. Di Mascolo: *Queueing analysis of the production of sterile medical devices by means of discrete-time queueing models*. Proccedings of the conference on Stochastic Models of Manufacturing and Service Operations, SMMSO 2013: Seeon, Germany.

Does Friendship Matter? An Analysis of Social Ties and Content Relevance in Twitter and Facebook

Christoph Fuchs, fuchsc@in.tum.de, Department of Informatics, TU München
Jan Hauffa, hauffa@in.tum.de, Department of Informatics, TU München
Georg Groh, grohg@in.tum.de, Department of Informatics, TU München

When users try to satisfy information needs, relevance is traditionally defined by metrics based on term or concept distance, link structure of the investigated information collection or previously conducted search sessions and selected results (by the same or other users). Leveraging the information within one's own social network to enrich search results is currently discussed in the concrete forms of Social Media Question Asking (SMQA) and Social Search. Analyzing two large datasets crawled from Twitter ($360,000$ user profiles, 223 million tweets) and Facebook ($25,737$ user profiles, 4.6 million posts from $936,992$ users), our findings suggest that content created by people who are socially close is of higher individual relevance than content created by others. Furthermore, our results indicate that the willingness to help satisfying information needs is higher for users within one's social network.

1 Introduction

Several approaches have been discussed to enhance the search and recommendation process with social aspects (Oeldorf-Hirsch, Hecht, Morris, Teevan, & Gergle, 2014; Lampe, Gray, Fiore, & Ellison, 2014). Traditionally, search results are calculated based on several relevance metrics based on search term distance, link structure, historic information (e.g. for personalization and former relevance judgments) and ontology-based approaches for conceptualization. Taking "social" in Social Search (McDonnell & Shiri, 2011) seriously, relevance must be regarded in a much broader sense, allowing social influence to impact individual relevance judgments like in marketing (Burnkrant & Cousineau, 1975) or in the choice of apps on a mobile phone (Aharony, Pan, Ip, Khayal, & Pentland, 2011). Information items might at first only be relevant because friends are interested in them but also might provide more serendipitous results caused by the *s*ocial relevance of these items in the social groups of a user. Thus they belong to the sphere of the wider unconscious information needs (Groh, Straub, Eicher, & Grob, 2013). A more "social" search engine could allow users to query privacy restricted, non-public information spaces of their friends directly, leading to results with a special social flavor of relevance (due to highly individual knowledge about the information seeker) and a broader information space (due to access to otherwise restricted information). Based on a corpus crawled from Twitter we investigate retweet behavior, considering retweeting a message as a positive relevance judgment. Our findings indicate that users assess tweets of people they directly interacted with as more relevant than messages sent from others. Furthermore, our analysis suggests that people we interacted with before tend to react to our questions faster than others. Analyzing messages posted on Facebook, we show that replies to a question are liked more by the information seeker (and are therefore possibly considered to be more helpful) when the reply is posted from within one's own social network.

2 Research Approach

2.1 Research Questions

The paper focuses on the following research questions:

I. Are relevance judgments on content correlated to the strength of the social relationship between author and recipient of the content?

II. Does social closeness influence the willingness to react to questions in a social media question asking scenario?

A positive relationship in RQ I suggests that search applications could benefit from integrating the content of socially close friends. Affirming RQ II indicates that a distributed social search approach should rely on querying socially close people.

2.2 Dataset

Twitter is one of the major online social networking services with more than 200 million active users by the time we collected our dataset[1]. Users have the possibility to select a tweet and resend it to their own followers (i.e., *retweet* it). In addition, users can send direct (but public) messages to other users using Twitter's @-operator at the beginning of a tweet. We collected a large sub-graph of Twitter on a per-user-basis by means of breadth-first search (Granovetter, 1976), collecting publicly available tweets dated between January 1 and July 26, 2012.

Facebook is one of the world's largest online social networks. Among numerous other things, users can establish friendship edges and post (and reply to) public messages. Users also have the ability to *like* content objects. The Facebook dataset has been retrieved using a crawling procedure based on Metropolis-Hastings Random Walk (cf. Gjoka et al. (Gjoka, Kurant, Butts, & Markopoulou, 2009)).

Tweets/posts ending with a question mark were regarded as questions. Previous research (Teevan, Morris, & Panovich, 2011) analyzing response quality and quantity in SMQA showed that phrasing a question as a single sentence with a question mark improves response quality and quantity. Validity checks on subsets (100 posts/tweets) revealed recall/precision values of 83%/55% (Facebook) and 25%/66% (Twitter)[2]. A question was considered as a "relevant" question only in case one could expect a real answer (e.g., no rhetorical questions). We are not in a position to reliably check to which degree a reply is a valid response to a question - even if the reply is a counter question it could provide information that is considered helpful by the author of the question. A small sanity check of 100 randomly chosen questions from the Facebook dataset and their respective answers confirmed that the answers in general fit the questions. We only collected publicly available data, accessible by any internet user. The data is not published and is only used for legitimate scientific research. The published information derived from the data does not disclose any details about the crawled profiles.

2.3 Evaluation Methods

2.4 RQ I: Correlation of relevance judgments and social connections

Twitter We consider the act of retweeting a tweet as a relevance judgment in favor of the original tweet by the retweeting user and assume that two users are "socially connected" if they have exchanged at least one directed message using Twitter's @-operator (regardless of the direction of the message). *Following* a user is not considered as a form of social connection, since it is one-sided and often motivated by the posted content (and not the respective "real" person). To simplify further analysis, we stick to the following notation:

[1] https://twitter.com/twitter/status/281051652235087872

[2] The lower recall value on Twitter is caused by the heavy usage of hashtags following the question

- M_b^a is the number of directed posts (using Twitter's @-operator) sent from user a to user b,
- RT_b^a is the number of tweets originally sent from user b and retweeted by user a,
- RT^a is the total number of all retweets sent from user a, i.e. $RT^a = \sum_{x \in U} RT_x^a$ with U being the set of all users, and
- TW^a is the number of tweets sent by user a.

RT_b^a and TW^a are also defined for a set of users U, i.e. $RT_U^a = \sum_{u \in U} RT_u^a$ and $TW^U = \sum_{u \in U} TW^u$. The set of retweets posted by a user u may contain tweets which were originally posted by (1) users who are followed by u, (2) users who exchanged at least one direct message with u and (3) users who don't belong to either of the previous two groups. Therefore, we define the following sets of users:

- $Friends(u)$ are users who are followed by u,
- $SocConn(u)$ are users who exchanged at least one direct message with u,
- $Other(u)$ are users who don't belong to either group, i.e. $\overline{Friends(u)} \cap \overline{SocConn(u)}$

Due to Twitter's API limitations, it is not possible to reconstruct the retweet graph: If a user a originally tweets a tweet t, a different user b retweets t as t' and a third user c retweets t' as t'', the tweet t'' is only marked as a retweet of t (but not of t'). Therefore, it's possible that users appear to retweet tweets from strangers (i.e. users not connected via follower edges or direct messages). To quantify retweet ratios, we use the function $R_1^u(x)$ defined as

$$R_1^u(x) := \frac{RT_x^u}{RT_{SocConn(u) \cup Friends(u)}^u}$$

which represents the ratio between tweets originally from users of group x which got retweeted by u and tweets originally from users within u's social contacts and people u follows which also got retweeted by u. A broader indicator, $R_2^u(x)$, also covering tweets from people where no connection exists, is defined as

$$R_2^u(x) := \frac{RT_x^u}{RT_{SocConn(u) \cup Friends(u) \cup Other(u)}^u}$$

Assuming that retweets are distributed equally, it is useful to compare $R_1^u(x)$ and $R_2^u(x)$ to $T_1^u(x)$ and $T_2^u(x)$ reflecting the contribution of the respective group to the overall tweet corpus. $T_1^u(x)$ is defined as

$$T_1^u(x) := \frac{TW^x}{TW^{SocConn(u) \cup Friends(u)}}$$

It represents the ratio of tweets posted by the respective group within the collected corpus, excluding users who are not followed or socially linked. This definition can get extended in analogy to R_2^u and is defined as

$$T_2^u(x) := \frac{TW^x}{TW^{SocConn(u) \cup Friends(u) \cup Other(u)}}$$

While $R_1^u(x)$ and $R_2^u(x)$ represent the contribution of a particular group (socially connected people; friends, i.e. people one follows; strangers) to the set of tweets retweeted by user u, $T_1^u(x)$ and $T_2^u(x)$ represent the proportion of all tweets posted by the respective group within the corpus. If one of the groups is overrepresented within the set of retweets (in comparison with the group's proportion in the full corpus) it could suggest that this group has more relevant content for a user u than other groups. To quantify this overrepresentation, the average ratios $1/|U| \cdot \sum_{u \in U} R_k^u / T_k^u$ are used for $k \in \{1, 2\}$ with U being the set of all available users.

Facebook For our analysis, we rely on the existing friendship network within Facebook and interpret the *likes* of a user as a relevance judgment. We (1) identify the set of questions and (2) analyze the responses for identified questions, i.e.

check whether the response has been posted by a friend of the question asker and check whether the question asker liked the response. A higher *like*-ratio for responses written by friends of the question asker than for other responses could suggest that friends are able to provide more valuable information than strangers. Using an explicit relevance judgment of the asker is in line with previous research on community Q&A (e.g. (Shah & Pomerantz, 2010)), where the answer explicitly chosen by the asker is considered best.

2.5 RQ II: Relation of willingness to help and social closeness

Twitter To assess the willingness to help other users, we analyze tweets containing a question indicated by ending with a question mark. For each of these tweets, we identify the responses and calculate response time and number of messages sent between the user asking the question and the user providing the answer. A negative correlation between response time and the number of exchanged directed messages would suggest that the closer two users are (i.e., the more direct messages they exchanged), the faster they reply to each others' questions. To improve the likelihood of the selected tweets actually forming a relevant question/answer pair, we only considered pairs of tweets posted within a time span of less than three weeks.

3 Results

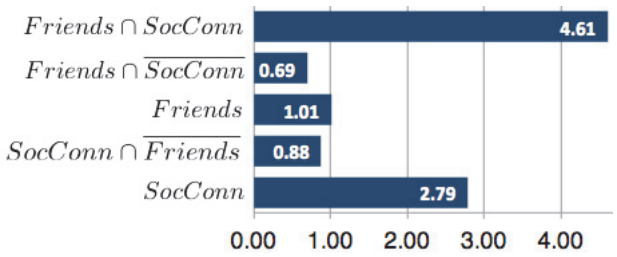
3.1 RQ I: Correlation of relevance judgments and social connections

Twitter The average set of retweets of a user consists of tweets from users in the sets $Friends \cap \overline{SocConn}$ (44.4%), $Friends \cap SocConn$ (26.7%), $Others$ (24.7%) and $SocConn \cap \overline{Friends}$ (4.1%). Table 1b shows the average results for R_1, R_2, T_1 and T_2 for the respective groups. Figure 1a depicts the ratio $\frac{R_1}{T_1}$ which can get interpreted as the degree of overrepresentation of a specific group within the set of retweets. It is noticeable that users retweet tweets from users who they follow and exchange direct messages ($SocConn \cap Friends$) much more often than their contribution to the overall amount of tweets would suggest. Furthermore, users one exchanged messages with, but didn't follow (group $SocConn \cap \overline{Friend}$), got retweeted (relatively) more often than users one only followed (group $Friend \cap \overline{SocConn}$; 0.88 vs. 0.69).

Facebook Out of $82,268$ replies, $73,941$ replies came from friends (thereof $11,144$ were liked by the question asker) and $13,327$ were given from other users (thereof, $1,692$ were liked by the question asker). On average, 15.1% of the answers are liked by the question asker if the author of the response is marked as a *friend* on Facebook – if this is not the case, the question asker likes only 12.7% of the replies. Fitting a linear regression model revealed a significant, but very weak positive correlation ($ASKER\_LIKES = 0.02 \cdot IS\_FRIEND + 0.13$, with $p < 1.021 \cdot 10^{-12}$ but a very low R^2 score of $5.82 \cdot 10^{-4}$).

3.2 RQ II: Relation of willingness to help and social closeness

Twitter We analyzed average and median response time for questions and the social connection between question asker a and responder r (using $max(M_r^a, M_a^r)$). For this part of the paper, we only considered a random sample of $550,680$ replies to questions for performance reasons. A linear regression model explains the response time as $-43.55 \cdot max(M_r^a, M_a^r) + 10,786$ with $p < 2.2 \cdot 10^{-16}$, but does not explain the variance ($R^2 = 6.505 \cdot 10^{-4}$). We estimated the stability of the result by running the same experiment on a smaller dataset ($50,000$ replies), where we got comparable results (-47.691, intercept $8,892$, $p < 1.01 \cdot 10^{-8}$, $R^2 = 6.559 \cdot 10^{-4}$). Given the high number of replies and the low p-value, we don't expect the result to change significantly when analyzing a larger set. The whole collection also



(a) Overrepresentation within the retweet set (R_1/T_1).

Group	R_1	T_1	R_2	T_2
$Friends \cap SocConn$	0.35	0.07	0.27	0.07
$Friends \cap \overline{SocConn}$	0.59	0.85	0.44	0.82
$Friends$	0.94	0.93	0.71	0.89
$SocConn \cap \overline{Friends}$	0.06	0.07	0.04	0.06
$SocConn$	0.41	0.15	0.31	0.13
$Others$	n/a	n/a	0.25	0.05

(b) R_1, T_1, R_2 and T_2, averaged over all users.

Figure 1: Results for RQ1 – users in the intersection of social connection and friendship (i.e. users who are being followed and exchanged at least one directed message) are retweeted 4.6 times more often on average than their overall contribution would suggest.

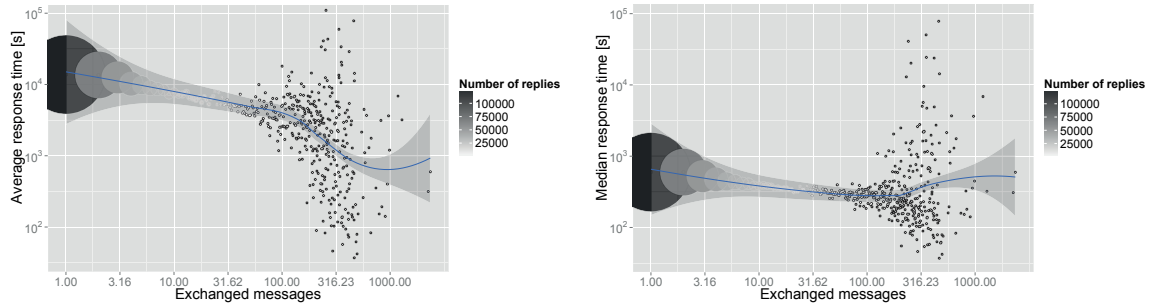
only represents a (larger) subset of all twitter messages and the analyis does not rely on the graph structure, therefore the quality of the result is not weakened. Figure 2a shows the average response time and suggests that the more directed messages two users exchanged, the smaller is the average response time. User pairs who exchanged a direct message for the first time when answering the question under consideration (i.e., exchanged messages equals to 1) have a high average response time of 3.8 hours (13,791 seconds, SD: 73,795) whereas users who have exchanged at least 1 message before have an average response time of 2.5 hours (9,096 seconds, SD: 54,086). In addition, 90% of the question asker / responder pairs have exchanged ≤ 35 messages. While pairs with no previous interaction have a median response time of 10 minutes (598 seconds), pairs who have directly communicated before have a lower median response time of 7 minutes (420 seconds) (see figure 2b).

4 Limitations

The interpretation of Facebook's *like* statement and retweets as relevance judgments is not optimal, since users do not necessarily associate it with a judgment on content quality. One might also doubt whether response time is a valid proxy for the willingness to help others – it could as well be the case that people who reply faster to questions received via Twitter do so because they spend a much larger part of their life online and maintain more and deeper relationships on Twitter. The datasets suffer from high variance, making it difficult to show indications and trends. We currently work towards a more complex modeling approach to further elaborate on this. We only considered direct social relationships, i.e. only a single step within the social graph. In a more sophisticated modeling approach, indirect relationships could also be taken into account.

5 Conclusion & Outlook

In this paper, we analyzed the impact of social relationships (number of directly exchanged messages, existence of a friendship relationship) on content relevance (retweeting a tweet, liking an answer) and willingness to help (approximated using the average response time) using two large datasets derived from Twitter and Facebook. Our results suggest that people we directly interact with are able to provide information that is relevant for us. This can be seen as supporting evidence for the utility of distributed social search, where information spaces from people within one's social network could also get consulted to satisfy information needs. Furthermore, people we interact with seem to reply faster to our questions, laying the foundation of a search concept based on mutual support. We will continue to develop a non-linear model while also integrating larger parts of the social graph.



(a) Average **(b) Median**

Figure 2: Time in seconds between asking the question and posting the response and number of exchanged direct messages between question asker and responding user (logarithmical scale), size and color reflect number of replies.

References

Aharony, N., Pan, W., Ip, C., Khayal, I., & Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, *7*, 643-659.

Burnkrant, R. E., & Cousineau, A. (1975). Normative Social Influence in Buyer Behavior. *Journal of Consumer Research*, *2*(3), 206-215.

Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2009). A Walk in Facebook: Uniform Sampling of Users in Online Social Networks. *arXiv:0906.0060 [cs.SI]*.

Granovetter, M. (1976). Network Sampling: Some First Steps. *The American Journal of Sociology*, *81*(6), 1287-1303.

Groh, G., Straub, F., Eicher, J., & Grob, D. (2013). Geographic Aspects of Tie Strength and Value Of Information in Social Networking. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks.*

Lampe, C., Gray, R., Fiore, A. T., & Ellison, N. (2014). Help is on the Way: Patterns of Responses to Resource Requests on Facebook. In *Proceedings of the 17th Conference on Computer Supported Cooperative Work & Social Computing* (p. 3-15). ACM.

McDonnell, M., & Shiri, A. (2011). Social search: A taxonomy of, and a user-centred approach to, social web search. *Program: electronic library and information systems*, *45*(1), 6-28.

Oeldorf-Hirsch, A., Hecht, B., Morris, M. R., Teevan, J., & Gergle, D. (2014). To Search or to Ask: The Routing of Information Needs Between Traditional Search Engines and Social Networks. In *Proceedings of the 17th Conference on Computer Supported Cooperative Work & Social Computing* (p. 16-27). ACM.

Shah, C., & Pomerantz, J. (2010). Evaluating and Predicting Answer Quality in Community QA. In *Proceedings of the 33rd International ACM SIGIR conference on Research and Development in Information Retrieval.*

Teevan, J., Morris, M. R., & Panovich, K. (2011). Factors Affecting Response Quantity, Quality, and Speed for Questions Asked via Social Network Status Messages. In *Proceedings of the 5th International Conference on Weblogs and Social Media* (p. 630-633). Association for the Advancement of Artificial Intelligence (AAAI).

Combining Crowdfunding and Budget Allocation on Participation Platforms

Claudia Niemeyer, claudia.niemeyer@kit.edu, Karlsruhe Institute of Technology (KIT)

Astrid Hellmanns, hellmanns@fzi.de, Forschungszentrum Informatik (FZI)

Timm Teubner, timm.teubner@kit.edu, Karlsruhe Institute of Technology (KIT)

Christof Weinhardt, weinhardt@kit.edu, Karlsruhe Institute of Technology (KIT)

Participatory budgeting (PB) and civic crowdfunding (CF) are both online participation processes that fund projects of public interest. Our research aims to investigate the advantages of combining the two processes and introduce the first steps of an experimental setting on how to evaluate different institutional budgeting mechanisms combined with crowdfunding.

Keywords: Crowdfunding, participation, budgeting

1 Introduction

Civic participation has been simplified by the use of online platforms. Participatory budgeting and crowdfunding are ever increasing in popularity. With PB, institutions like local governments, enterprises, universities and the like are able to empower their members (i.e., citizens, employees, students, etc.) to participate in decisions on how to spend the institutional budget (Cabannes, 2004). CF, in addition to the process of budget allocation, takes a slightly different perspective as it taps the members' willingness to pay for certain projects at issue (Belleflamme et al., 2014). Naturally, such CF typically entails a decision on how the money should be used- the online platform Kickstarter is a popular example (Ricker, 2011). In this paper, we introduce and compare evaluation categories for these two participation processes and sketch out how to combine the two processes to offer a wider range of financing possibilities in the public sector. Such allocation methods for co-financing projects in institutions yields the advantages of i) empowering the institution's members by involving them in budgeting decisions and ii) tapping their individual willingness to pay and hence increasing financial leeway for project realisation overall. With our work, we seek to provide a structure for evaluating and comparing CF and PB and address the question of how different allocation mechanisms of such co-financing affect the members' behaviour, e.g., in terms of participation, funded amounts, perceived fairness, etc.

This paper is organized as follows: After giving an overview on participatory budgeting and civic crowdfunding in Section 1, we introduce categories to evaluate these platforms in Section 2 and present our research proposition and concluding remarks in Section 3.

1.1 Participatory Budgeting

The participatory budgeting (PB) of Porto Alegre, Brazil, in the late 1980s has set an example on participation worldwide (Herzberg, 2006). From South America, PB has now become even more popular. More than 1500 PB examples can be found at present.[1] These processes differ in numerous aspects. Regarding European examples of PB, a definition has been established (see Sintomer et al., 2010) that contains five criteria: (1) a financial and budgetary dimension, (2) a territorial dimension (compulsory connection to an area), (3) an iterative dimension (excluding one-time voting), (4) a deliberative dimension (PB as a special form of public discourse), and finally (5) commitment (to account for results). In Germany, over 100 local governments included participatory budgets in their financial decision making over the last years (Scherer et al., 2012). Good examples that include an explicit use of ICT are the PB of Berlin-Lichtenberg[2], the city of Potsdam[3], the city of Trier[4], and the city of Köln[5] (see Nitzsche et al. 2012). Although many types of PB exist, the German Basic Law[6] confers in art. 28 (2) the financial autonomy to local authorities. This power of decision thus includes the final word of the local governments. Most PB use discussions and votes for decision making and determining the budget allocation. Muller et al. (2013) introduce a budgeting mechanism similar to CF in an enterprise setting, where IBM equipped all participants with $100 that they can only use to support projects on their website (also see Feldmann, 2014). However, in our work we use the term crowdfunding in its narrow definition, where the individuals are included in the budget provision.

1.2 Crowdfunding

Crowdfunding started as financing model in creativity-based industries (Agrawal et al., 2013) and has quickly become popular with online platforms like Kickstarter as a relatively easy way to access the possibility of start-up funding. CF is as a method of "financing from the crowd" (Moritz et al. 2013, p2), which is initiated by an open call and aims for financial support for a project or company with or without reward (Belleflamme et al. 2010). Usually copious investors engage in the funding on CF websites that act as an intermediary. The platforms as intermediaries are particularly advantageous as they offer a standardized process for investments.

Civic crowdfunding, as CF of projects that are of public interest, has become interesting in the context of social inequality and civic participation (Davies, 2015). Examples can be found on Spaceive, Neighbor.ly and citizinvest.com.

[1] cf. http://www.participatorybudgeting.org/about-participatory-budgeting/where-has-it-worked/
[2] cf. https://www.buergerhaushalt-lichtenberg.de/
[3] cf. https://buergerbeteiligung.potsdam.de/kategorie/buergerhaushalt
[4] cf. https://www.trier-mitgestalten.de/haushalten/4698
[5] cf. https://buergerhaushalt.stadt-koeln.de/2015/
[6] cf. http://www.gesetze-im-internet.de/englisch_gg/

Feldmann et al. (2013) embedded CF inside an organisation and introduced an interaction structure which can be compared to the IBM case (Muller et al. 2013, Muller et al. 2014), but also to the narrow definition of crowdfunding, as soon as the participants invest their own money.

2 Evaluation Categories

With the increasing number of participatory processes on online platforms many questions arise. Besides legal aspects that differ from area to area, it has not completely been investigated how to qualitatively judge these instruments. Beside the requirements on the funding mechanism itself, that we will focus on in Section 3, in this section, we present a number of general quality criteria for participation platforms and civic crowdfunding, followed by a comparison between the two concepts.

Participation and CF platforms can be evaluated in the following categories (cf. Escher, 2013; Kersting, 2014; Scherer et al., 2012), whereas the order proposed is rather random and does not imply importance:

***Achievement of objectives*:** Each platform has its individual aim. This can be the allocation of a budget as desired by the citizens, the realisation of as many projects as possible, the maximisation of the total welfare, the gain of citizens' attention, and many more. This category needs for CF and PB a well-stated aim to be able to be evaluated. This category is closest to the funding mechanism and is therefore further investigated in Section 3.

***Usability*:** The technical and graphical implementation of participation and CF platforms require the same usability demands as other websites. During the conception and implementation of the platform the creation of usage incentivisation and measures to tie users should be especially considered. The usability of a platform is mostly determined by the effectivity and efficiency of the application and the satisfaction of the users.

***Inclusiveness*:** Both CF and PB websites try to reach as many people in the target group as possible and are able to offer a variety of projects. Therefore it is important that participation and CF platforms draw attention from public relations and the connection with other formats and media. Still the motivation behind inclusiveness is different. The democratic nature of PB claims representativity to reach as many citizens as possible and motivate them to participate, whereas CF websites try to maximise the financial capacity. The uniqueness of participation platforms lies in the definition of its target group. Ideally all segments of society and stakeholders should participate to be able to interpret the voting result as a representation of the whole population. The website, however, should also be accessible to underrepresented segments of the society.

***Efficiency*:** Authorities and institutions have great interest in an efficient participation process. The better the process can be integrated in authorities, the easier it is accepted by the employees. The same holds for CF: an easy and efficient process motivates capital acquirers and investors.

***Responsiveness*:** The acceptance (as well as the preparedness of the officials) and an active and numerous participation of citizens is a prerequisite for the success of a civic platform. It can also be measured by the way the

authority integrates itself into the online process and actively participates in the discourse. Intermediaries of CF platforms have a similar responsibility as local authorities. It is important to maintain CF projects to keep the investors' trust and be attractive in the future.

***Commitment*:** Regarding commitment in participatory processes, PB and CF differ, especially in Germany. PB includes citizens in financial decision making, but still has the right to neglect the outcome of the vote[7], so that the PB tend to have a low level of commitment. CF, however, needs a high degree of commitment due to the financial relation of the agents.

***Security*:** When looking at online political opinion making and voting, questions about the safeguard for identity and data security have to be answered by the maintainer of the platform. It is important to sensibly and recognisably weigh between the necessary degree of authentication and identification of the citizen and the risk of being vulnerable to manipulation. The security of personal data in the context of CF faces the same challenge. In addition, even more data of investors as well capital acquirers is needed during the actual payment process, such as bank account numbers. Maintainers have to take over responsibility to defend all participants from data abuse.

Transparency: The traceability of aims and rules is essential for both kinds of participatory platforms, since PB as well as CF platforms claim absolute reliability to encourage users to participate. Therefore, the responsible authorities of PB should bare how they proceed with proposals and voting results. This includes an extensive post-processing. CF depends on the reliability of its projects (and the platform itself), which is promoted by explanations and comments of the responsible agents (e.g., the documentation of project progress).

The list of categories prompts that there is a big intersection in the evaluation of both types of platforms: The categories *Usability*, *Efficiency*, and *Transparency* have predominantly similar requirements, *Security* only differs in the additional personal payment data and *Inclusiveness* in the representativity in the political context.

With regard *Commitment*, PB in Germany is still on an early stage. A higher level of commitment can be reached, if local governments include CF in the process of PB. Additionally, the citizens' willingness to support projects of their interest using their own finances, has not yet been included in PB. Authorities probably do not know their willingness to pay and would therefore be happy to have an adequate instrument to find out. This aspect can be covered by combining participatory platforms with budgeting and crowdfunding techniques.

We start with our first category *Achievement of objectives* that is closely attached to the funding mechanism. When applying our research results on the mechanism on platforms in the future, one will have to check for the other criteria step by step to build a successful participation platform.

[7] cf. http://www.buergerhaushalt.org/sites/default/files/6._Statusbericht_buergerhaushalt.org_.pdf

3 Research Outlook

In this Section we pose the question on how different forms of institutional support mechanisms affect individual decisions in a crowdfunding scenario. For this purpose, we sketch out an experimental blueprint design for evaluating different forms of institutional support mechanisms in the context of crowdfunded projects. Specifically, we address the question of how the design of an allocation mechanism for institutional budgets impacts crowdfunding behaviour in a setting where projects are funded both by the institution and its members. Moreover, we aim at considering total welfare and individual satisfaction. In a 2-by-2 matrix, Table 1 assigns and illustrates the cases when i) the decision of how the budget is used and ii) budget provision is distinguished for individuals and the institution.

		budgeting provision	
		individuals	**institution**
budgeting decision	**individuals**	crowdfunding	participatory budgeting
	institution	taxing, donation	management decision

Table 1: Mechanisms by decision makers and budget providers.

In the following, we briefly sketch out an experimental design suitable to investigate co-funding as outlined above. We assume that an institution, i.e. a community government, a university, or a company, etc. faces several different projects j with cost c_j and needs to decide on how to allocate its budget B across those projects. Budget, however, is limited and may not be sufficient for financing all projects ($B < \sum c_j$). Since the projects, however, directly benefit the members of the institution, i.e., its citizens, students, or employees, they might exhibit a certain willingness to pay for each project too. In some sense, the community members are able to support projects on an internal kickstarter-like platform (Kuppuswamy et al. 2013). In comparison to such "traditional" crowdfunding, we here consider the case in which a central institution is also present and wants to support its members by providing co-funding. The crucial question for practitioners is how to set up the support rules in the most efficient way, i.e. spending its budget optimally from the desired perspective (e.g., total welfare). Let u_{ij} denote the derived utility of subject i if project j is realized. This utility value is zero otherwise. The projects thus have the character of a threshold public good (Zhang et al. 2014). Every subject now has a certain individual budget at his or her disposal, denoted by b_i. We assume that no subject can fund any project entirely on his or her own ($b_i < c_j$ for all i and j) and that also no subject would want to do that ($u_{ij} < cj$ for all i and j). Moreover, we assume that the sum of all individual budgets is not sufficient to fund all projects ($\sum b_i < \sum c_j$). Lastly, from a global perspective, every project is worth realizing, i.e. $\sum u_{ij} > c_j$ for all j.

In the experiment, several subjects are provided with a list of their personal utility and cost values for every project 1 to m. They know that n-1 other subjects face a similar situation, with other utility values, however. These utility values are private information, but may be correlated to some extent among subjects. Subjects then simultaneously decide on how to allocate their budget across the projects. This clearly represents a stark simplification of reality, where typically subjects can observe the progress and fulfilment rate of project funding over the course of days or weeks and can repeatedly adjust their contribution accordingly. We leave this to future research. The individual contributions of subject i to project j are denoted by z_{ij}. If funding for a project is not sufficient, i.e. the threshold is not met ($\sum_{i=1}^{n} z_{ij} < c_j$), the invested amount is transferred back to the subject. Over-investments are refunded proportionally.

Each subject faces the conflict between contributing to the funding of one or more projects which appear beneficial to him or her. At the same time, he or she must consider the possibility of overfunding and that a lower contribution would have yielded a higher payoff ceteris paribus. This, however, holds for all subjects which introduces the risk of not funding a project at all.

This base scenario with no institutional support in place will be compared to a situation where the institution also invests its budget B in several projects following different possible rules. Some mechanisms easily come to mind. The institution could, for instance, i) subsidize each investment by a fraction $\delta > 0$ until the budget is consumed, or ii) close the funding gap of the most popular project which do not have met the threshold. Project popularity may, for instance, be measured in terms of relative or absolute funding. Such different support schemes, assuming a constant institutional budget B, may work with varying degrees of success. These different schemes may then be evaluated by means of i) total welfare (sum of all realized utility values minus money spent), ii) social security/solidarity (utility value for least benefited subject), or iii) the number of realised projects. The resulting payoff Π_i for subject i in the baseline situation is expressed by:

$$\Pi_i = b_i + \sum_{j=1}^{m} (u_{ij} + ((\frac{\sum_{i=1}^{n} z_{ij} - c_j}{\sum_{i=1}^{n} z_{ij}} - 1) * z_{ij}) * \mathrm{I_j}$$

with $b_i \geq \sum_{j=1}^{m} z_{ij}$ and $\mathrm{I_j} = 1 \Leftrightarrow \sum_{i=1}^{n} z_{ij} \geq c_j, \mathrm{I_j} = 0$ otherwise.

Our next steps will be to further investigate which factors might influence the decision behaviour of participants when facing a crowdfunding situation which includes institutional support. We then conduct the experiment including the treatments with different funding mechanisms of crowdfunding with institutional support.

To the best of our knowledge, the combination of an institution's budgeting with crowdfunding processed have not been investigated from an experimental and behavioural perspective. Our research outline intends to address this gap. We anticipate that an institution can most certainly affect its members' willingness to contribute and hence also total welfare if designing support schemes appropriately. In times of narrow public budgets this way of funding appears highly suitable for tapping private capital, helping to ensure high quality public goods and thus yield a positive societal impact.

4 References

Agrawal, A. K./Catalini, C./Goldfarb, A.: *Some Simple Economics of Crowdfunding*. Working Paper. National Bureau of Economic Research, June 2013.

Belleflamme, P./ Lambert, T./ Schwienbacher, A.: *Crowdfunding: An Industrial Organization Perspective*. SSRN Working Paper 2151179.

Belleflamme, P./ Lambert, T./ Schwienbacher, A.: *Crowdfunding: Tapping the right crowd*. Journal of Business Venturing 29, Nr. 5 (September 2014): 585–609.

Cabannes, Y.: *Participatory Budgeting: A Significant Contribution to Participatory Democracy.* Environment and Urbanization 16, Nr. 1 (January 2004): 27–46.

Escher, T.: *Mobilisierung zu politischer Partizipation durch das Internet. Erwartungen, Erkenntnisse und Herausforderungen der Forschung*, in: Analyse und Kritik 02/2013, S. 449-476.

Feldmann, N./ Gimpel, H./ Muller M./ Geyer W.: *IDEA ASSESSMENT VIA ENTERPRISE CROWDFUNDING: AN EMPIRICAL ANALYSIS OF DECISION-MAKING STYLES*. ECIS 2014 Proceedings, Juni 2014.

Herzberg, C.: *Der Bürgerhaushalt von Porto Alegre. Wie partizipative Demokratie zu politisch-administrativen Verbesserungen führen kann*. Auflage: 3., Aufl. Münster u.a.: LIT, 2006.

Kersting, N.: *Online-Beteiligung - Elektronische Partizipation. Qualitätskriterien aus Sicht der Politik*, in: Voss, K. (Hrsg.): Internet und Partizipation, Wiesbaden 2014, S. 53-87.

Kuppuswamy, V./Bayus, B. L.: *Crowdfunding creative ideas. The dynamics of project backers in Kickstarter*, SSRN Electronic Journal 2013.

Moritz, A./Block, J.H.: *Crowdfunding und Crowdinvesting: State-of-the-Art der wissenschaftlichen Literatur (Crowdfunding and Crowdinvesting: A Review of the Literature)*, SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 1. September 2013.

Muller, M./ Geyer, W./ Soule, T./ Daniels, S./ Cheng, L.: *Crowdfunding Inside the Enterprise: Employee-initiatives for Innovation and Collaboration*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 503–12. CHI '13. New York, NY, USA: ACM, 2013.

Muller, M./ Geyer, W./ Soule, T./ Wafer, J.: *Geographical and Organizational Distances in Enterprise Crowdfunding*. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, 778–89. CSCW '14. New York, NY, USA: ACM, 2014.

Nitzsche, P/Pistoia, A./Elsäßer, M.: *Development of an Evaluation Tool for Participative E-Government Services. A Case Study of Electronic Participatory Budgeting Projects in Germany*, in: Administratie si Management Public, 18/2012, S. 6-25.

Ricker, T.: *Kickstarted: How One Company is Revolutionizing Product Development*. The Verge, December 20, 2011 http://www.theverge.com/2011/12/20/2644358/kickstarter-success-product-development-revolution (accessed February 6, 2015).

Scherer, S./Wimmer, M. A.: *Reference Process Model for Participatory Budgeting in Germany*, in: Tambouris, E./Macintosh, A./Sæbø, Ø (Hrsg.): *Electronic Participation*, Lecture Notes in Computer Science 7444, Springer Berlin/Heidelberg, 2012, S. 97-111.

Sintomer, Y/Herzberg, C./Röcke, A: In *Der Bürgerhaushalt in Europa – eine realistische Utopie?*, VS Verlag für Sozialwissenschaften, 2010.

Zhang, N/Datta, S./Kannan, K.N.: *An Analysis of Incentive Structures in Collaborative Economy. An Application to Crowdfunding Platform,* SSRN Scholarly Paper. Rochester, NY: Social Science Research Network 2014.

What is “Industrial Service”? A Discussion Paper

Björn Schmitz, Bjoern.Schmitz@kit.edu, Karlsruhe Institute of Technology

Ralf Gitzel, Ralf.Gitzel@de.abb.com, ABB AG

Hansjörg Fromm, Hansjoerg.Fromm@kit.edu, Karlsruhe Institute of Technology

Thomas Setzer, Thomas.Setzer@kit.edu, Karlsruhe Institute of Technology

Alf Isaksson, Alf.Isaksson@se.abb.com, ABB SE

Abstract - Industrial services are an important source for revenue and growth for its providers. However, neither the term industrial service nor its concrete subareas are unambiguously defined. In this paper we discuss the definition of industrial service and its delineation from related concepts which can serve as a suitable foundation for a research area on *industrial services* for both, academics and industrial researchers. In addition, we will identify future areas of interest in the domain to encourage researchers to advance the field. We hope that the paper triggers fruitful discussions about its subject and increases awareness about specific characteristics of industrial services which have so far not been taken up by the research community.

1 The Importance of Industrial Services

In recent years, more and more manufacturing companies have made the expansion of their service business a core strategy (Vandermerwe and Rada 1988; Quinn, Doorley and Paquette 1990; Wise and Baumgartner 1999; T. S. Baines, et al. 2009). In developed economies *industrial services* often account for more than half of the manufacturing industry‘s profits with annual sales volumes of billions of dollars[1], thus forming an attractive market to operate in (Strähle, Füllemann and Bendig 2012).

Despite their economic importance industrial services are an underresearched topic. There is a lack of standardization and systematic approaches to successfully exploit the service potential are often missing. Moreover, the manufacturing industry is subject to substantial technological changes (e.g. through the emergence of cyber-physical systems, Industry 4.0, etc.). Technology will not only influence the way in which industrial services are delivered and consumed. It will also promote innovation in the development of industrial equipment, vehicles, production processes and entire factories, i.e., it will change the service objects themselves. The lack of methods

1 In Germany, annual sales of industrial (maintenance) services account for approximately 30 billion Euros (Roland Berger Strategy Consultants 2010), compared to approximately 1 trillion dollars annually in the U.S. (Cohen, Agrawal and Agrawal 2006).

and the changing technological environment will result in a variety of economic, organizational and technical challenges which will need to be addressed in the future.

In this paper we aim to contribute to the establishment of a research area on *industrial services* for both, academic and industrial researchers. First, we provide a literate review on definitions of the term industrial service and summarize common characteristics (Section 2.1). Second, we outline different perspectives to look at the subject and develop a matrix which can be used to classify the various types of industrial services (Section 2.2). Third, in order to encourage researchers to advance the field we identify future areas of interest and outline business and technological drivers that will affect the domain of industrial services in the future (Section 2.3). Following these analyses is a short conclusion at the end of the paper.

2 Industrial Services– Today and in the Future

In the following we discuss various definitions of the term industrial service before we outline the fields which industrial services cover and point to future developments in this area.

2.1 Definition of Industrial Service

Today, there is no unambigious definition of what *industrial service* is, even though the term is frequently used in the literature. In order to establish a shared understanding of the domain and the characteristics commonly related to industrial services we will discuss various definitions that have been proposed in the literature. We do not aim to provide an extensive literature review on all possible definitions. Our objective is rather to highlight the diverse nature of industrial services as well as the different perspectives under which they are studied. Based on a review of leading academic journals, conference proceedings and industry publications we found various definitions of the term, each highlighting different aspects of industrial service.

Industry Focused Definitions - Industrial services are defined based on the individual or organization they are marketed to or with regard to the entity providing a service. One of the prevailing views is that industrial services are marketed to industrial clients (Jackson and Cooper 1988) or clients with industrial production (Brax 2005) as opposed to consumer clients. Other authors relax this assumption stating that end-users of industrial services are not restricted to industrial firms (Oliva and Kallenberg 2003). Proposed characterizations have in common that industrial services are defined based on the type of consumer it is delivered to. Contrary to that there are definitions taking a service provider's perspective instead. They regard industrial services as all services developed and provided by industrial suppliers (Reen 2014), by manufacturers of industrial equipment (Brax 2005) or by other companies specializing in services (Brax 2005). Following this line of argumentation there is a number of other definitions relaxing or emphasizing on or the other characteristic of industrial services in this context (Homburg and Garbe 1999; Oliva and Kallenberg 2003; Kowalkowski 2006).

Process Focused Definitions - A second type of definitions emphasize the process view on services following the argumentation that industrial services comprise all services relating to organizational processes. For instance, Kowalkowski (2006) defines industrial services as "processes supporting customers' industrial production processes, so that value for them is created in those processes". Jackson and Cooper (1988) speak of "production services" in this context. Other authors even extend this view and refer to industrial service as "a process of exploiting the competences, knowledge, and technology base of a company's business, manufacturing, and operational processes" (Reen 2014).

Asset Focused Definitions - In academia and industry the term aftersales services is still frequently used as a synonym for industrial services (Homburg and Garbe 1999; Oliva and Kallenberg 2003; Johannsson and Olhager 2004; Paloheimo, Miettinen and Brax 2004; Kowalkowski 2006). This type of definitions follows the rationale that industrial services can be described based on the lifecycle of industrial goods. For instance, Oliva and Kallenberg (2003) regard industrial services as "product- or process-related services required by an end-user over the useful life of a product". Such definitions (indirectly) assume that industrial services are always provided in relation to or in conjunction with industrial goods. Homburg and Garbe (1999), for instance, classify industrial services into pre-purchase, at-purchase and aftersales industrial services. Jackson and Cooper (1988) speak of maintenance, repair and overhaul services, which aim at keeping an organization maintained and operating in this context.

IHIP Focused Definitions - For many years academics have argued about the inherent characteristics that differentiate services from goods. In the course of this debate the so called IHIP criteria (Regan 1963; Zeithaml, Parasuraman and Berry 1985) have evolved which aim at distinguishing services from goods based on four characteristics: intangibility, heterogeneity, inseparability and perishability[2]. Based on the IHIP criteria researchers have proposed to distinguish industrial services from services in general by introducing additional criteria which they consider to be integral parts of industrial services. Examples of such industrial service specific criteria are "specialization" (Jackson and Cooper 1988), "technology" (Jackson and Cooper 1988), "consumption in irregular patterns" (Morris and Fuller 1989) and many others (Jackson and Cooper 1988; Morris and Fuller 1989).

The various definitions illustrate that industrial services are studied from a variety of perspectives each emphasizing particlar characteristic of the concept. Consequently, it is difficult to develop a universally accepted definition of the term. However, we are convinced that a shared understanding of the common characteristics of industrial services is required in order to promote consistent research on the topic despite its interdisciplinary nature. Thus, in the following we want to outline a more holistic perspective on industrial services.

We see industrial services as services relating to industrial products (or rather goods) or industrial systems (a system of / systems of industrial goods). In contrast to consumer goods, industrial goods are used by business or industrial clients as opposed to consumer clients (Jackson and Cooper 1988). Industrial goods or systems may

2 The practical suitability of using IHIP as a means to distinguish services from products has been subject to discussions in academic literature (see e.g. Vargo and Lusch 2004).

range from a single pump to an entire manufacturing plant, from a single truck to an entire fleet of vehicles. From a functional point of view, industrial services go far beyond traditional maintenance, repair and overhaul. Value-added services comprise activities such as condition monitoring, predictive maintenance, advanced diagnostics or asset and fleet management. While some of those services are closely integrated with the industrial good itself (e.g. manufacturer offering remote access control for measurement equipment) other services may only be distantly related to the industrial goods which are perceived to form the core offering of a manufacturing company (e.g. manufacturer of measurement equipment offering tank management)[3]. Besides these changes in the functional dimension, recent developments in service business models and contracts have affected the way in which industrial services are offered. Full-service contracts as well as availability, usage- or performance-based revenue models are gaining momentum. All these different perspectives – the delimitation of industrial and consumer clients, the functional view with traditional and value-added services and the contractual perspective with the evolvement of new business models and service contracts - allow us to study industrial services from various angles and with different methodologies and theoretical approaches.

2.2 Fields of Industrial Services

While there is a series of taxonomies and classifications for services in general (e.g. Schmenner 1986; Silvestro, et al. 1992; Hill 1999; Buzacott 2000) there is relatively little material as to what services are part of the field of industrial services other than the catalogues of service providers (for an exception see Matyas, Rosteck and Sihn 2009 as well as Henkel et al. 2004 (as quoted by Kowalkowski 2006)).

To overcome this problem we have developed a matrix which can be used to classify the various types of industrial services (compare Figure 1). It maps service activities and associated research problems with scientific expertises needed to develop corresponding solutions. In addition, the matrix considers different industrial domains, thus allowing to account for unique technological and business requirements of different industries. The matrix was developed and discussed with more than ten researchers from academia and industry with long-time expertise in industrial service research. In the following we will briefly explain each dimension of the matrix.

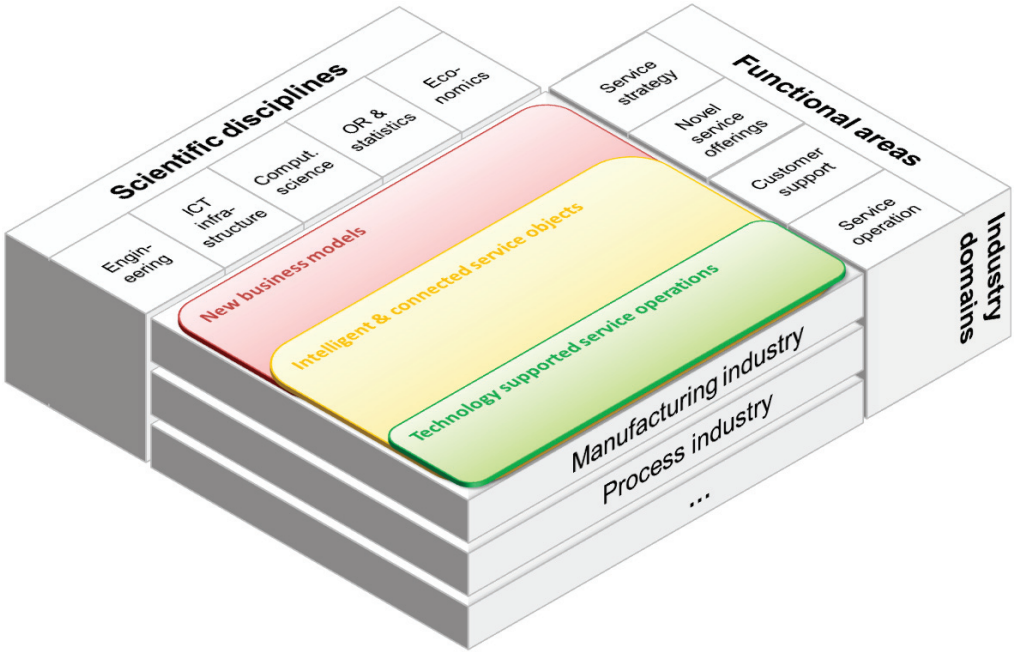
First, there are *functional areas* which represent different types of service activities. Defining a *service strategy* forms an integral part of all subsequent activities and covers strategic planning at different levels. Considering the market situation outlined in Section 1, companies need to develop *novel service offerings* which are technologically advanced, customer-oriented and proactive in order to remain competitive. C*ustomer support* is mainly concerned with providing information and supporting the customer with required knowledge to operate equipment or services efficiently. Finally, the *service operation* constitutes the functional back bone of the service business. It covers activities from organizing, scheduling, executing and controlling the service delivery process.

3 Similar discussions have been taken up in research on *product service systems* which are integrated offerings of goods and services (see T. S. Baines, et al. 2007). Although we are not able to discuss product service systems in detail in this work we want emphasize that both domains are closely linked to each other. Following this line of thought Meier, Roy and Seliger (2010) state that understanding the „[m]eaning of industrial service is a necessary prerequisite to understand the nature of an Industrial Product-Service System".

The second dimension of the cube describes *scientific disciplines* which loosely define the academic backgrounds required for the advancement of industrial services and associated research problems. The main contributions come from economics, operations research and statistics, computer science, more hardware-oriented information technologies to various engineering disciplines.

In addition, we consider a third important dimension which are different *industry domains* such as manufacturing industry, process industry and utilities. They reflect the different economic, organizational and technical requirements which need to be considered in order to implement industrial services effectively.

Figure 1 Fields of industrial service



2.3 Industrial Services of the Future

Having proposed a matrix to classify different types of industrial services we aim to identify future areas of interest in the domain by discussing how business and technological drivers are going to affect industrial services in the future.

There is currently a world-wide drive towards strengthening the industrial base in developed economies (see e.g. Popkin and Kobe 2010; European Commission 2012a; European Commission 2012b; Ministry of Industrial Renewal France 2013). In Europe, promoting *industrial leadership* is one of the focal points of research and innovation initiatives to achieve the strategic goals of the agenda Europe 2020. While the majority of research projects currently aim at the development of *technologies and (product) innovations* (see e.g. European Commission 2012a), complementary research and innovation in industrial services are needed to maintain technological leadership of the European industry and to achieve sustainable growth of the European economy in

a globalized world. In fact, all technological innovations promoted through European and national research initiatives such as *Industry 4.0* (Federal Ministry of Eduction and Research and Federal Ministry of Economics and Technology 2012)*, Factories of the Future* (European Commission 2013a), *Sustainable Process Industries* (European Commission 2013b), *The New Industrial France* (Ministry of Industrial Renewal France 2013) and many others require an equal advancement in research and innovation on industrial services to maintain and extend industrial leadership.

Holistically, the calls mentioned above paint a picture of a future that can be used to extrapolate the necessary service research. There are business drivers and technology drivers that cause dramatic changes in the way industrial services are offered and delivered (compare Figure 1).

On the business side, we see changing customer expectations and the demand for *new business models* (Kindström and Kowalkowski 2009; Sakao, Sandström and Matzen 2009; Ostrom, et al. 2010; Meier, Roy and Seliger 2010; Reen 2014). Due to the ever increasing complexity of industrial products and services, customers are willing to buy more and higher value services from the supplier than they did in the past. On the other hand, customers want to avoid the risk of unpredictable service costs (Neely 2008; Meier, Roy and Seliger 2010) and more and more customers go over to requesting service offerings which guarantee performance or outcome via performance-based or full-service contracts (see e.g. Kim, Cohen and Netessine 2007; Hypko, Tilebein and Gleich 2010; Huber and Spinler 2014).

Advances in technology are causing changes in almost every aspect of industrial services. The *service objects*, industrial products, are becoming more and more *intelligent and connected*. They can sense and monitor their condition, they can predict and analyze failures and they can configure, manage, and heal themselves. They are connected via networks (e.g. the Internet) with other systems or service centers. They know which part needs to be replaced and can order the required spare part by themselves. They have become cyber-physical systems, i.e., integrated computational and physical processes (Lee 2006; Lee 2008). This allows easy access and identification for service personnel and the possibility to do remote diagnosis and maintenance. As a result new or advanced types of services become possible, which require active development to realize their full potential.

Regardless of how industrial production systems do or do not change, the *service operation* itself can significantly be *improved and supported by technology*. Service technicians are nowadays equipped with smart phones or hand-held devices that can automatically guide them to the point of service and can give them any information necessary to perform the service task. Augmented reality devices (e.g. eyeglasses) can display information (labels, instructions) in the field of view of the technician to support his or her work. Intelligent diagnosis systems can assist service personnel by quickly and automatically analyzing symptoms of a failure or malfunction and identifying probable causes that can explain these symptoms. With new technological advances around the corner, there will be new applications in industrial service as well. (Aleksy, Stieger and Vollmar 2009; Aleksy and Rissanen 2012; Tesfay, et al. 2013)

All these developments underpin that industrial services will play an important role in industry and academia for the years to come. Exploiting its potential will require an interdisciplinary approach which takes account of the diverse nature of industrial services.

3 Conclusion

In this paper, we have presented a definition and taxonomy of industrial services. To take account of business and technological trends we have outlined future developments in the field which will affect the way in which industrial services will be developed, delivered and consumed in the future. We aim to inspire new research in this domain and we hope that this paper will contribute to constructive debates about the characteristics, boundaries and the future of industrial services.

References

Aleksy, Markus, and Mikko J Rissanen. "Utilizing wearable computing in industrial service applications." *Journal of Ambient Intelligence and Human Computing, 5(4)*, April 2012: 443-454.

Aleksy, Markus, Bernd Stieger, and Gerhard Vollmar. "Case Study on Utilizing Mobile Applications in Industrial Field Service." *IEEE Conference on Commerce and Enterprise Computing (CEC 09).* Vienna, Austria, 2009. 333-336.

Baines, T S, et al. "State-of-the-art in product service-system." *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 221(10)*, October 2007: 1543-1552.

Baines, Tim S, Howard W Lightfoot, Ornella Benedettini, and John M Kay. "The servitization of manufacturing: A review of literature and reflection on future challenges." *Journal of Manufacturing Technology Management, 20(5)*, 2009: 547-567.

Brax, Saara. "A manufacturer becoming service provider - challenges and a paradox." *Managing Service Quality, 15(2)*, 2005: 142-155.

Buzacott, John A. "Services system structure." *International Journal of Production Economics, 68(1)*, October 2000: 15-27.

Cohen, Morris A, Narendra Agrawal, and Vipul Agrawal. "Winning in the Aftermarket." *Harvard Business Review, 84(5)*, May 2006: 129-138.

European Commission. *Horizon 2020 - Call for Factories of the Future.* 2013. http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/calls/h2020-fof-2015.html accessed February 10th, 2014.

—. *Horizon 2020 - Call for SPIRE - Sustainable Process Industries.* 2013. (http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/calls/h2020-spire-2015.html accessed February 10th, 2014.

—. *Horizon 2020 - Industrial Leadership.* 2012. http://ec.europa.eu/programmes/horizon2020/en/h2020-section/industrial-leadership accessed March 20th, 2014.

—. "A stronger European industry for growth and economic recovery." *Industrial policy communication update. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions.582 final*, October 10th, 2012.

Federal Ministry of Eduction and Research, and Federal Ministry of Economics and Technology. *Project of the Future: Industry 4.0.* 2012. http://www.bmbf.de/en/19955.php accessed January 15th, 2014.

Henkel , Carsten B, Oliver B Bendig, Tobias Caspari, and Nihad Hasagic. "Industrial Services Strategies: The quest for faster growth and higher margins." Consultancy Report, Monitor Group, 2004.

Hill, Peter. „Tangibles, Intangibles and Services: A New Taxonomy for the Classification of Output.“ *The Canadian Journal of Economics, 32(2)*, April 1999: 426-446.

Homburg, Christian, and Bernd Garbe. "Towards an Improved Understanding of Industrial Services: Quality Dimensions and Their Impact on Buyer-Seller Relationships." *Journal of Business-to-Business Marketing, 6(2)*, 1999: 39-71.

Huber, Sebastian, and Stefan Spinler. "Pricing of Full-Service Repair Contracts with Learning, Optimized Maintenance, and Information Asymmetry." *Decision Sciences, 45(4)*, August 2014: 791-815.

Hypko, Phillipp, Meike Tilebein, and Ronald Gleich. "Clarifying the concept of performance-based contracting in manufacturing industries: A." *Journal of Service Management, 21(5)*, 2010: 625-655.

Jackson, Ralph W, and Philip D Cooper. "Unique Aspects of Marketing Industrial Services." *Industrial Marketing Management, 17(2)*, May 1988: 111-118.

Johannsson, Pontus, and Jan Olhager. "Industrial service profiling: Matching service offerings and processes." *International Journal of Production Economics, 89(3)*, June 2004: 309-320.

Kim, Sang-Hyun, Morris A Cohen, and Serguei Netessine. "Performance Contracting in After-Sales Service Supply Chains." *Management Science, 53(12)*, December 2007: 1843-1858.

Kindström, Daniel, and Christian Kowalkowski. "Development of industrial service offerings: a process framework." *Journal of Service Management, 20(2)*, 2009: 156-172.

Kowalkowski, Christian. „Enhancing the industrial service offering: New requirements on content and processes.“ *Dissertation from the International Graduate School of Management and Industrial Engineering.* Linköping, Sweden: Linköping University, Institute of Technology, 2006.

Lee, Edward A. „Cyber Physical Systems: Design Challenges.“ *Technical Report No. UCB/EECS-2008-8.* Berkeley, California: Electrical Engineering and Computer Sciences Department, University of California at Berkeley, January 2008.

—. "Cyber-Physical Systems - Are Computing Foundations Adequate?" *Position Paper for National Science Foundation Workshop on Cyber-Physical-System: Research Motivations, Techniques and Roadmap October 16-17,2006, Austin, Texas, USA.* Berkeley, California: Electrical Engineering and Computer Sciences Department, University of California at Berkeley, October 2006.

Matyas, Kurt, Armin Rosteck, and Wilfried Sihn. "Industrial Services - Corporate Practice and Future Needs for Action in Companies and in Applied Research." *Proceedings of the 42nd CIRP Conference on Manufacturing Systems, Grenoble, France*, June 2009.

Meier, H, R Roy, and G Seliger. "Industrial Product-Service Systems - IPS²." *CIRP Annals - Manufacturing Technology, 59(2)*, 2010: 607-627.

Ministry of Industrial Renewal France. *The new face of the industrial France.* 2013. http://www.redressement-productif.gouv.fr/files/nouvelle_france_industrielle_english.pdf accessed March 25, 2014.

Morris, Michael H, and Donald A Fuller. "Pricing an Industrial Service." *Industrial Marketing Management, 18(2)*, May 1989: 139-146.

Neely, Andy. "Exploring the financial consequences of the servitization of manufacturing." *Operations Management Research, 1(2)*, December 2008: 103-118.

Oliva, Rogelio, and Robert Kallenberg. "Managing the transition from products to services." *International Journal of Service Industry Management, 14(2)*, 2003: 160-172.

Ostrom, Amy L, et al. "Moving Forward and Making a Difference: Research Priorities for the Science of Service." *Journal of Service Research, 13(1)*, February 2010: 4-36.

Paloheimo, Kaija-Stiina, Ilkka Miettinen, and Saara Brax. "Customer Oriented Industrial Services." *Report Series.* Espoo: Helsinki University of Technology, BIT Research Centre, 2004.

Popkin, Joel, and Kathryn Kobe. "Manufacturing Resurgence - A Must for U.S. Prosperity." *Study for the National Association of Manufacturers and the NAM Council of Manufacturing Associations.* Washington, Washington DC: Joel Popkin and Company, January 2010.

Quinn, James Brian, Thomas L Doorley, and Penny C Paquette. "Beyond Products: Services-Based Strategy." *Harvard Business Review, 68(2)*, March - April 1990: 58-67.

Reen, Natalia. „The Pricing of Industrial Services.“ *Dissertation.* Turku, Finland: Åbo Akademi University, 2014.

Regan, William J. "The Service Revolution." *Journal of Marketing, 27(3)*, July 1963: 57-62.

Roland Berger Strategy Consultants. „INDUSTRIESERVICES IN DEUTSCHLAND - Status Quo und zukünftige Entwicklungen.“ *Market Report.* Munich, April 2010.

Sakao, Tomohiko, Gunilla Ölundh Sandström, and Detlef Matzen. "Framing research for service orientation of manufacturers through PSS approaches." *Journal of Manufacturing Technology Management, 20(5)*, 2009: 754-778.

Schmenner, Roger W. "How can services businesses survive and prosper." *MIT SLOAN Management Review, 27(3)*, Spring 1986: 21-32.

Silvestro, Rhian, Lin Fitzgerald, Robert Johnston, and Christopher Voss. "Towards a Classification of Service Processes." *International Journal of Service Industry Management, 3(3)*, 1992: 62-75.

Strähle, Oliver, Michael Füllemann, and Oliver Bendig. "Service now! Time to wake up the sleeping giant - How services can boost long-term growth with attractive returns in industrial goods business." *Consultancy Report.* Munich/Zurich: Bain & Company Inc., 2012.

Tesfay, Welderufael B, Markus Aleksy, Karl Andersson, and Marko Lehtola. "Mobile Computing Application for Industrial Field Service Engineering: A Case for ABB Service Engineers." *The 7th IEEE LCN Workshop on User Mobility and Vehicular Networks (ON-MOVE 2013).* Sydney, Australia, 2013. 188-193.

Vandermerwe, Sandra, and Juan Rada. "Servitization of Business: Adding Value by Adding Services." *European Management Journal, 6(4)*, Winter 1988: 314-324.

Vargo, Stephen L, and Robert F Lusch. "The Four Service Marketing Myths - Remnants of a Goods-Based, Manufacturing Model." *Journal of Service Research, 6(4)*, May 2004: 324-335.

Wise, Richard, and Peter Baumgartner. "Go Downstream: The New Profit Imperative in Manufacturing." *Harvard Business Review, 77(5)*, September - October 1999: 133-141.

Zeithaml, Valarie A, A Parasuraman, and Leonard L Berry. "Problems and Strategies in Services Marketing." *Journal of Marketing, 49(2)*, Spring 1985: 33-46.

Total service experience as a function of service experiences in service systems

Ronny Schueritz, ronny.schueritz@kit.edu, KIT

Service firms act as part of one or more service systems for the purpose of co-creating value. Customers interact with multiple service firms during a customer journey, thus have several service encounters with different service firms. As one encounter influences following encounters, as well as the overall customer experience, the provided service quality of one service firm might impact the perceived service quality of another service firm and the overall service quality. Therefore, it is not sufficient for a service firm to look exclusively at their own service encounters, but rather at the full customer journey from a service system perspective. In other words, service providers need to focus on optimizing the whole service system in order to ensure a positive total customer experience. This paper proposes a model that offers an holistic view on the impact factors of the total customer experience. It illustrates the relationship between service systems, service encounters and customer experience and thus gives guidance to researchers and practitioners on how to optimize the total customer experience.

1 Introduction

The most common customer related metric used by managers is the customer satisfaction, as it is known to be connected to repurchases (GUPTA & ZEITHAML, 2007). It is even referred to as the "one number you need to grow" in order to increase profits (REICHHELD, 2003).

Since customer satisfaction is very transactionally focused and does not show the full picture, service design favors to look at the customer experience (VOSS, ROTH, & CHASE, 2008). Even if both measures differ in their definition, an increase lead to higher customer retention and loyalty (CARUANA, 2002).

What if factors that are influencing the customer experience are not on the agenda of researchers and managers? If they do not understand all factors that are influencing the customer experience, they potentially fail to improve it and thus miss the opportunity to maximize profits.

Through a number of publications in recent years a service and service system science has been established. So far researchers have not sufficiently connected the research streams of service systems and customer experience. However, service providers are in fact service systems that act as part of a bigger service systems (MAGLIO, VARGO, CASWELL, & SPOHRER, 2009). If customers start to interact with these systems they will potentially interact with different providers. Thus it needs to be analyzed if interacting with a service system can influence the customer experience and eventually if customer experience can be measured on a service system level.

For that purpose the existing literature around the relevant topics is reviewed in the first part of the paper and the established understanding is presented. By proposing a holistic model on the topic we will then connect existing research in order to create a service system perspective on customer experience.

2 Related work

This section illustrates the understanding of research towards customer satisfaction and how it is created in the service encounter through the perception of service quality. This is followed by the concept of customer experience that looks, opposing to customer satisfaction, more holistically at the topic. At the end, we briefly introduce the concept of service systems.

2.1 Customer satisfaction, service quality and service encounter

It is generally accepted that customer satisfaction is the result of a subjective comparison between expected and perceived performance of a product or a service (LEWIS & BOOMS, 1983; PARASURAMAN, ZEITHAML, & BERRY, 1985, 1988; OH, 1999). In services, expectations are the considerations made by the customer of what might happen in the service encounter. If providers fail to achieve these expectations, the customer is dissatisfied (OLIVER, 1997). Customer retention, which can be linked to customer satisfaction, is highly important for service firms as it shows the likelihood of a customer to repurchase from a specific service provider based on a positive attitude towards this provider (GREMLER & BROWN, 1996).

Customer satisfaction of a service is influenced by a variety of factors such as service quality, price, situational and personal factors. In pure services (education, financial services, etc.), the service quality is the most crucial factor, it represents the customers perception of reliability, assurance, responsiveness, empathy and tangibles (PARASURAMAN et al., 1988). Customers have certain expectations towards the service which they compare with the perceived service, resulting in the perceived service quality.(GRÖNROOS, 1984)

This customer perception of a service forms at the 'moment of truth', the service encounter. In the encounter the customer actually interacts with the service provider and experiences the service quality (SHOSTACK, 1985). Service encounters can differ in nature. There are generally three types of service encounters: remote encounters, telephone encounters and face-to-face encounters. A customer can experience several service encounters with a mix of these types with one service firm (SHOSTACK, 1985; NORMANN, 2000), *for example an airline customer books a flight online, may calls the airline hotline, at the airport they check-in and drop off their luggage at the desk, later on board they order a drink, and so on.*

Research shows that customers perceive each service encounter with a different level of service quality. The airline customer may perceive the booking and checking with a high service quality but is unsatisfied with the service on board. Furthermore, the perceived quality of different service encounters has a different impact on the overall service quality of the service experience. Hence, the overall customer satisfaction with the service provided by a service firm is a function of the perceived service quality in the service encounters(WOODSIDE, FREY, & DALY, 1989; HANSAN & DANAHER, 1999). Therefore the airline customer may still be very satisfied with the airline as weighing of importance among different encounters in a service journey.

2.2 Customer experience

Concepts around service quality and service encounters are well established in research. Nevertheless they have been criticized in the last years for their sole focus on the transaction between customer and service firm, where as service design looks at the concepts of customer experience and customer journey from a less static perspective (VOSS et al., 2008). The customer experience describes the whole journey a customer is taking and therefore possibly starts way before the actual service encounters and goes on afterwards (BERRY, CARBONE, & HAECKEL, 2002). The created customer experience is a subjective response of the customer towards multiple interactions with one service firm (GENTILE, SPILLER, & NOCCI, 2007; MEYER & SCHWAGER, 2007). Therefore, the customer experience can include marketing communications (BRAKUS, SCHMITT, & ZARANTONELLO, 2009), word of mouth (KWORTNIK & ROSS, 2007) or

the experience approaching the service firm (GILMORE & PINE, 2002). Customer experience is consequently the " total experience, including the search, purchase, consumption, and after-sale phases" (VERHOEF et al., 2009). Referring to retail, Verhoef further describes that this includes also elements outside of the control of the service firm. *In the example of the airline customer, the customer experience therefore also includes among other things the experience of finding the airline online (search, purchase), the time at the airport and the flight (consumption).* Some models are proposed to illustrate the customer experience and the influence factors. They suggest that a customer holistically asses its journey with the service provider. So far empirical research around customer satisfaction has only focused in assessing parts of it in isolation (VOSS et al., 2008; GREWAL, LEVY, & KUMAR, 2009; VERHOEF et al., 2009; PAYNE, STPRBACKA, & FROW, 2007).
Eventually Lemke suggests a connection to these theories in order to make customer experience more measurable by including the service quality aspect and proposes a model that represents the customer experience quality (LEMKE, CLARK, & WILSON, 2010).

2.3 Service systems

The goods-dominant-logic represents "the neoclassical economics research tradition"(HUNT, 2000) which is focused on the production and the output of goods and services. In this logic value is created by a producer and destructed by a consumer. Further, services are inferior and only seen as an intangible good or as an add-on for goods. Service-dominat logic (SD-logic) is based on the view that service is the fundamental basic for economic exchange in which competences (knowledge and skills) are applied to benefit another party and eventually co-create value(VARGO & LUSCH, 2004). Further SD-logic rejects the view of a producer and consumer (VARGO, 2007). In a concept where service is exchanged for service and and value is co-created the entities are referred to as *service systems* (MAGLIO et al., 2009). A service system is a "...dynamic value-cocreation configuration of resources, including people, organizations, shared information (language, laws, measures, methods), and technology, all connected internally and externally to other service systems by value propositions" (MAGLIO et al., 2009). Service systems can be present in all kinds of representations such as people, businesses, government, etc that engage in *service interactions* with each other in order to co-create value (MAGLIO et al., 2009). *Referring to the airline example that means the airline is a service system itself but is also part of a wider service system consisting of a variety of partners such as the airports (departure and arrival) with all its players (security, ground control, shops,etc.) and so on.*

3 Total service experience

The measurements of customer satisfaction are classically based on the perceived service quality in the service encounters with the service firm (WOODSIDE et al., 1989). Research in service design has expanded this very transaction focused theory and looks at the customer experience in a more holistic way (VOSS et al., 2008; GREWAL et al., 2009; VERHOEF et al., 2009; PAYNE et al., 2007). Authors point out that they understand that there are other external elements outside of the control of the service firm that influence the customer experience, but so far they do not look at it from a system perspective. However, nowadays customers interact with a service system consisting of several players as service providers have created a whole network of contractors and partners in order to fulfill their value proposition. In a customer journey, customers often experience several service encounters in which they interact with multiple service providers and service systems. *The airline customer may use an online search engine to find the right flight, uses and online check in service and uses services at the airport. All of the these service encounters take place directly with the airline but might be performed with the support of contractors and partners. In some cases the customer might not even notice that he does not interact with the airline itself.* Todays models around customer experience do not cover this fact, but additional service

providers and systems add factors influencing the customer experience that need to be considered in order to understand how customer experience is created. The proposed model illustrates the relationship of these factors.
The proposed model in figure 1 shows a series of **service experiences** (E_n) with various service firms. A service experience is created through the combination of different factors such as the customers perception of a certain service quality, price, external factors and others. The current mood of the customer can heavily impact the service experience and even be affected by a service experience itself (PRICE, ARNOULD, & DEIBLER, 1995). This good or bad mood will then spill over and affect following service experiences to a certain extent (y_n) (PRICE, ARNOULD, & TIERNEY, 1989). Woodside proved in his research that each service encounter is valued differently by the customer and is by that affecting the total customer satisfaction with a different impact (WOODSIDE et al., 1989). In our proposed model the series of service experiences (E_n) are as well affecting the **total service experience** (E_T) to a certain extent (x_n). This means that total service experience is the sum of weighted single service experiences or in mathematical terms: $E_T = \sum(x_n * E_n)$. This becomes very clear when using the airline customer as an example again.*The customer might have a good experience with the check in at the airport but have a very unpleasant contact with the security control which might causes a bad mood. The very adequate service on board is not perceived as satisfying anymore and the total service experience is possibly only negative.*

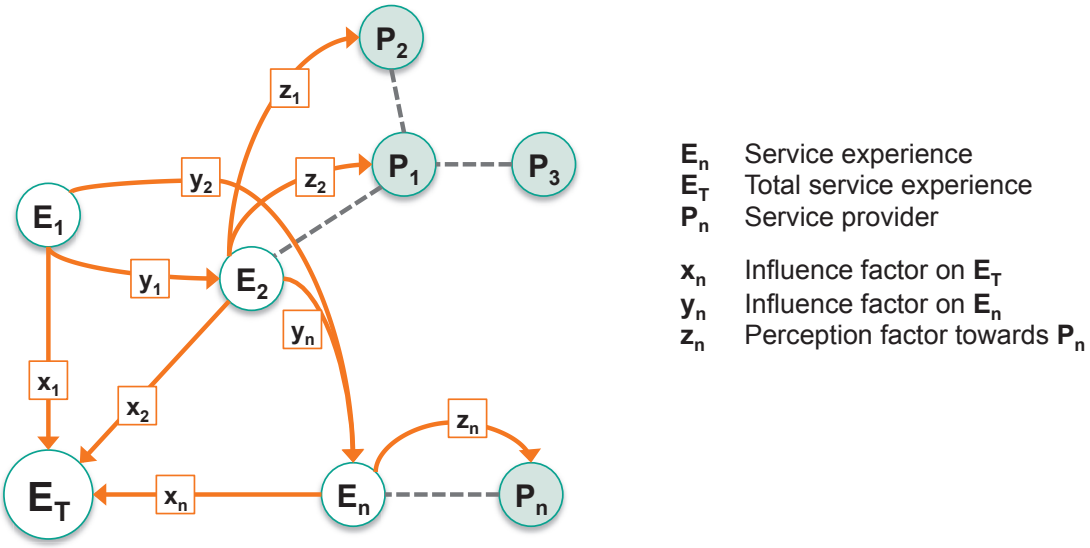


Figure 1: Total service experience model

As Kwan and Hottum illustrate in their research the service provider may rely on the integration of other providers' services or even the service encounter might not be facilitated by the service provider itself rather by a contracted partner. (KWAN & HOTTUM, 2014) In the later case the service provider has a service level agreement with their customers but has less control over the actual service encounter, as it is fulfilled by a contracted partner. In any case there are several service providers (P_n) involved in one service experience. *The airline wants to transport their customers and the customers luggage from A to B, but parts of that proposition get fulfilled by contractors at the airport and not the airline itself. If the airport is losing the luggage, the customer might blame the airline for the loss and connect the bad service experience with the airline, not the airport.* No matter the explicit setup, it is important who the customer perceives as the service provider and who they affiliate the experience with. The degree to which a customer is connecting the experience with a specific service provider is expressed by the factor (z_n).

A series of service experiences is created by a variety of service firms (e.g. airport and airline). These service providers might not even be directly connected to each other or have a contracted obligation to each other (e.g. the airline and a taxi or a bar at the airport), but they are connected through the customer journey. Hence, service providers are formally or informally connected in bigger more abstract service systems through their specific value proposition in the customer journey.
Using this model as fundamental basis of understanding service experience in service systems, some implications are emerging:

- Service providers, while being service systems itself, are embedded in higher-level service systems (with partners and contractors) that are connected to other service systems via a customer that interacts along a sequence of service experiences.
- As customers have service experiences created by other service providers and all service experiences of a customer journey influence the total service experience, each service provider might be impacted by the service quality of another service provider or system.
- Service providers need to understand the whole customer journey from a system perspective in order to manage alliances, partnerships and value propositions accordingly.

4 Conclusion and future research

The concepts around customer experience and customer journey provide a promising basis to enhance the service of a service provider as they have a more holistic view than customer satisfaction. Service providers and researchers still need to change the way they understand customer experience. Instead of designing the customer experience from one service providers view, it is important to view and optimize the entire service system. Customers are interacting with different service systems through their customer journey and by that service systems are impacted by each others performance even if they are not directly connected. Therefore it needs a service system view to the topic of customer experience, introduced in this paper as the total service experience. In this paper we have contributed a model that provides a service system perspective on the customer experience and introduces the concept of total service experience which is the sum of single service experiences. This model will further provide orientation and basis for future research. It does not claim to show the complete picture of influence factors, but shows how total service experience is created in service systems. It still needs to be answered how to measure total service experience in a realistic and practical way. Further it needs to be defined where the customer journey starts and where it ends in service systems and what factors ultimately influence the service experience.

References

BERRY, L. L., CARBONE, L. P., & HAECKEL, S. H. (2002). Managing the Total Customer Experience Recognizing the Clues. *MIT Sloan Management Review*, *43*(3), 85-89.

BRAKUS, J., SCHMITT, B., & ZARANTONELLO, L. (2009). Brand experience: what is it? How is it measured? Does it affect loyalty? *Journal of marketing*, *73*(3), 52–68.

CARUANA, A. (2002). Service loyalty. *European Journal of Marketing*, *36*(07/08), 811-828.

GENTILE, C., SPILLER, N., & NOCCI, G. (2007). How to Sustain the Customer Experience:. *European Management Journal*, *25*(5), 395–410.

GILMORE, J. H., & PINE, B. J. (2002). Customer experience places: the new offering frontier. *Strategy & Leadership*, *30*(4), 4-11.

GREMLER, D., & BROWN, S. (1996). Service loyalty: its nature, importance and implications. In B. S. J. R. Edvardsson B. & E. Scheuing (Eds.), *Advancing service quality: a global perspective* (p. 170-180). International Service Quality Acossiation.

GREWAL, D., LEVY, M., & KUMAR, V. (2009). Customer Experience Management in Retailing: An Organizing Framework. *Journal of Retailing*, *85*(1), 1–14.

GRÖNROOS, C. (1984). A Service Quality Model and its Marketing Implications. *European Journal of Marketing*, *18*(4), 36–44.

GUPTA, S., & ZEITHAML, V. (2007). Customer metrics and their impact on financial performance. *Marketing Science (forthcoming)*.

HANSAN, D., & DANAHER, P. (1999). Inconsistent performance during the service encounter. *Journal of Service Research*.

HUNT, S. (2000). *A general theory of competition: resources, competences, productivity and economic growth*. Sage Publications.

KWAN, S., & HOTTUM, P. (2014). Maintaining consistent customer experience in service system networks. *Service Science*, *6*(02), 136-147.

KWORTNIK, R. J., & ROSS, W. T. (2007). The role of positive emotions in experiential decisions. *International Journal of Research in Marketing*, *24*(4), 324–335.

LEMKE, F., CLARK, M., & WILSON, H. (2010). Customer experience quality: an exploration in business and consumer contexts using repertory grid technique. *Journal of the Academy of Marketing Science*, *39*(6), 846–869.

LEWIS, R., & BOOMS, B. (1983). The marketing aspects of service quality. In S. G. Berry L.L. & G. Upah (Eds.), *Emerging perspectives in service marketing* (Vol. 36). American Marketing Association.

MAGLIO, P., VARGO, S., CASWELL, N., & SPOHRER, J. (2009). The service system is the basic abstraction of service science. *Information Systems and e-Business Management*, *7*(04), 395-406.

MEYER, C., & SCHWAGER, A. (2007). Understanding customer experience. *Harvard business review*, *85*(2).

NORMANN, R. (2000). *Service management: Strategy and leadership in the service business* (3rd ed.). John Wiley and Sons.

OH, H. (1999). Service quality, customer satisfaction, and customer value: A holistic perspective. *International Journal of Hospitality Management*, *18*, 67-82.

OLIVER, R. L. (1997). *Satisfaction: A behavioral perspective on the consumer*. McGrawe-Hill.

PARASURAMAN, A., ZEITHAML, V., & BERRY, L. (1985). A conceptual model of service quality and its implication for future research. *Journal of Marketing*, *49*, 41-50.

PARASURAMAN, A., ZEITHAML, V., & BERRY, L. (1988). Servqual: a multiple-item scale for measuring consumer perceptions of service quality. *Journal of retailing*, *64*, 33-49.

PAYNE, A. F., STPRBACKA, K., & FROW, P. (2007). Managing the co-creation of value. *Journal of the Academy of Marketing Science*, *36*(1), 83–96.

PRICE, L., ARNOULD, E., & DEIBLER, S. (1995). Consumers emotional responses to service encounters. *International Journal of Service Industry Management*, *6*(03), 34-63.

PRICE, L., ARNOULD, E., & TIERNEY, A. (1989). Going to extrems: managing service encounters and assessing provider performance. *Journal of Marketing*, *59*, 83-97.

REICHHELD, F. F. (2003). *The one number you need to grow.* Retrieved 09.01.2014, from `https://hbr.org/2003/12/the-one-number-you-need-to-grow`

SHOSTACK, G. (1985). Planing the service encounter. In J. CZEPIEL, M. Solomon, & C. SURPRENANT (Eds.), *The service encounter* (p. 243-54). Lexington Books.

VARGO, S. (2007). On a theory of markets and marketing: From positively normative to normatively positive. *Australasian Marketing Journal*, *15*(01), 53-60.

VARGO, S., & LUSCH, R. (2004). Evolving to a new dominant logic for marketing. *Journal of Marketing*, *68*(01), 1-17.

VERHOEF, P. C., LEMON, K. N., PARASURAMAN, A., ROGGEVEEN, A., TSIROS, M., & SCHLESINGER, L. A. (2009). Customer Experience Creation: Determinants, Dynamics and Management Strategies. *Journal of Retailing*, *85*(1), 31–41.

VOSS, C., ROTH, A. V., & CHASE, R. B. (2008). Experience, Service Operations Strategy, and Services as Destinations: Foundations and Exploratory Investigation. *Production and Operations Management*, *17*(3), 247–266.

WOODSIDE, A., FREY, L., & DALY, R. (1989). Linking service quality, customer satisfaction, and behavioral intention. *Journal of Healthcare Marketing*, *9*(03), 5-17.

Conversion Centered Personalization – a Data-Driven Approach to Service Design

Dirk Ducar, dirk.ducar@de.ibm.com, IBM Deutschland

Jella Pfeiffer, jella.pfeiffer@kit.edu, Institute of Information Systems and Marketing, Karlsruhe Institute of Technology

We propose a data-driven approach to fuel the creative process in service design. It enables the identification of deficits and potentials in the ways services can be accessed and completed. This approach relies on a user centric perspective and is based on a synopsis of user clusters and three different types of conversions. In particular, we stress the importance of focusing service design not only on improving how a service is used (micro-conversions) and completed (pull-out-conversions) but also on how the service is accessed (link-up-conversions). The overall goal of our approach is to improve the user experience and thereby increase conversion rates. This is achieved by widening and adjusting the array of elements that can be used in personalization according to the usage patterns of different user clusters.

1 Introduction

Enterprises with complex high value products, such as car manufacturers, run websites that incorporate a lot of different services around their brand and products. Product catalogs, product configurators, information and news sections, entertainment and sweepstakes or feedback forms are prominent examples. Such websites offer users different ways of accessing, using and completing these services. Consequently, personalization of websites should also comprise these dimensions of how users interact with the offered service. First, the design of different ways of using and completing a service enables the personalization of services themselves. Second, the design of alternative access ways to services allows to personalize the way users may start a service.

Monitoring and optimizing of design and personalization features usually focuses the successful completion of services (Kohavi and Parekh 2003). This web analytics approach is very useful for evaluating the functionality of individual services and appropriateness of contents of personalized elements. It is of limited use when the goal is supporting an overall orchestration and a further development of services, access ways and personalized elements. This is because the focus lies mainly on the functionality of services and not on the different ways they are accessed and used by different groups of users.

Our concept of **Conversion Centered Personalization** tackles the complexity of this enhancement task by shifting the perspective towards the user. We propose a data-driven approach that enables the identification of deficits and potentials in what access ways to services are provided and what options of using and completing a service are offered. In order to achieve this goal, we take a closer look at different types of of conversions that can be used to monitor and enable the user journey from service to service. This approach results in a widened array of personizable elements that meet the needs and intentions of actual user groups.

2 Conceptual Framework

2.1 Enabling the User Journey

Website personalization uses datamining techniques and/or business rules in order to create a bespoke user experience based on actual user data. In most cases, strategic reasoning is combined with mathematical algorithms in one way or another (Markellou, Rigou & Sirmakessis 2005). The general setup of personalization techniques reflects the marketing strategy followed. Traditional online marketing research approaches stress the importance of defining goals of customer journeys and focus on the completion of steps that have to be taken on the way (Heinrich & Flocke 2014). The idea of one distinctive sales funnel with minor variations is common sense around online marketers. This includes the notion that there is only one logical sequence of service use – the funnel. Sales funnel management is trying to jockey the customer through this funnel. Obviously this type of strategy has strong implications for the design and personalization of services.

However, actual website users do have more complex motivations. For example, many users of car manufacturers web sites do not plan or cannot even afford to buy in the nearer future but still do have a strong interest in the brand and its products for very different reasons. To be successful, the marketing strategy underlying website personalization needs to acknowledge these different users' attitudes. To follow a user centric approach has immediate consequences to the perception of the user journey and strong implications for the design and personalization of services and access ways.

Putting this into perspective, we suggest a shift in marketing strategy that ties together classical **sales funnel management** with the idea of **behavior driven targeting**. While sales funnel management is trying to **activate** the user, behavior driven targeting is **comforting** them by simply offering the experience an individual user will most likely respond to. This can include promoting website usage that might be considered counterproductive from a sales funnel perspective.

2.2 Rethinking Conversions

In prior research the concept of conversions is used to describe a hierarchy of marketing goals (Hassler 2012) or to focus on intermediate steps that need to be performed in order to complete services (Kohavi and Parekh 2003).

These authors use an idea of conversions that aims at constituting a sales funnel or at optimizing the way a user is led through a service, respectively. Into a different vein, we want to establish a notion of conversions that can be used in the field of diversification and personalization of access ways to and completion modes of services. Following Kohavi and Parekh (2003) it is based on the idea of defining all steps that perform or execute a service as **micro-conversions**. Following a functional perspective, we introduce two further kinds of conversions, namely **link-up-conversions** and **pull-out-conversions**. Link-up-conversions are actions that invoke a service. The successful completion of one of the services on a website is regarded a pull-out-conversion.

Website personalization usually includes specific **calls for action** to individual users. These calls are suggesting different ways of using services and work as shortcuts and gateways to start services like preconfigurations, lists of suggested products, prefilled forms or complete services. A user that follows such a call for action executes a link-up- or a pull-out-conversion.

In order to get a clearer picture of the different ways in which web site usage can be personalized, we suggest a categorization of calls for actions and associated conversions that locates them somewhere on a continuous scale between those concepts of comforting and activating introduced above. From a methodological standpoint this is quite a challenge but it is a necessary step in translating the marketing strategy formulated above into a framework for an overall analysis.

Figure 2.2 below shows a row of 'calls for actions' that provide access to (personalized) services and will thus lead to link-up-conversions. They get more and more serious from a users perspective and can be ranked in an ordinal scale of categories like passive engagement, active involvement and real commitment.

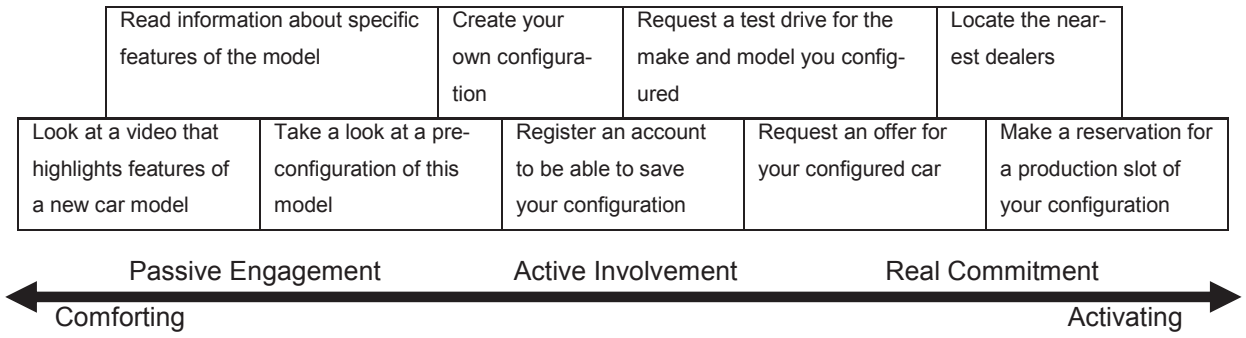


Figure 2.2

Categorization example of calls for action as in the car manufacturing industry.

Calls for action connetcted to alternative pull-out-conversions can be situated on the same scale between comforting and activating. Examples for alternative calls to complete a configuration service are to save the configuration, to share it in a social network or to buy the product.

Different design components available on a web site can be categorized in a similar way. The call to register an account can be embedded into a comforting video that advertises the content provided to registered users. Another, more activating way of linking-up to the registration service would be to pop up a registration form once the user

enters the web site. The user specific application of design components is another way websites are being personalized.

The different ways such link-up- and pull-out-conversions are orchestrated by personalization of contents and design components that do reflect the marketing strategy in charge. Instead of trying to make the user follow a predefinded pathway by using more and more obtrusive calls for action and desing components, we suggest a more flexible and user centric approach to personalization and service design that enables different usage patterns and user journeys. It includes supporting backwards mobility in a suggested sequence of services and alternative ways of combining and completing services. We propagate a design approach that endorses a multitude of different pathways and a personalization that suggest users individual routes though a network of services. A notion that is less focused on overall marketing goals but on the needs and intentions of the user. This approach is based on the idea that meeting the customers' preferences results in improved long term results concerning conversions.

From this widened orchestration perspective it becomes clear that a pull-out-conversion might flag a change over to another service and thereby have double functionality: it can both be a link-up as well as a pull-out conversion. Furthermore, it should become clear that some pull-out-conversions actually represent what is typically classified as conversion: the successful completion of the service resulting in a lead like an actual sale or order request. Our notion of pull-out-conversion however is broader in that is also includes conversions like saving a product configuration or even just clicking on a teaser that starts one but ends another service.

3 A Data-Driven Approach to Service Design

The conceptual framework outlined above enables a data-driven approach to the design of services and affiliated conversions. Where other authors advise to design and personalize services according to the presumptive needs of predefined target audiences (Kramer, Noronha & Vergo 2000), we want to use the behavioral data of real user groups in order to generate experiences that meet their needs. This can be achieved by looking at who uses the conversions and how – and who does not and why. Thus, we suggest a synopsis of user clusters and conversions that enables the detection of deficits and potentials in service design.

3.1 User Clusters

To improve website design and thereby enable an improved personalized user experience it is crucial to better understand user segments and user journeys. Cluster analysis can enable marketing strategists to spot different types of life cycles and their specific stages. In our approach, we use clusters to evaluate the offering of conversions on a website.

Ongoing analysis of actual user data shows that there are natural groups of users with very similar journeys. They differ in intensity of use and general service usage patterns. Intensity of use can be measured by variables as active time on the page, numbers of sessions, time between sessions, duration of sessions, number of pages viewed per

session, user interactions per session and so on. General interest variables look at numbers as proportion of time spent using different services on the page, number of different products customer has shown interest in, interest proportions for first second and third product a customer is interested in and so forth.

The analysis of user clusters reveals that different groups can plausibly get described with labels as bouncers, owners, shoppers or fans with different levels of activity. Obviously those clusters need far more scrutiny to fully understand the nature of the user groups they constitute. For the purpose of outlining how they can be used to help optimize the orchestration of services and calls for actions on a website the current level of cluster detail is, however, sufficient.

3.2 Clusters, Conversions and the Design Process

A synopsis of user clusters and conversions can provide real explanations for refusal-rates for link-up-conversions and dropout rates of services – explanations that reach beyond simple usability issues. A deep understanding of clusters allows deductive reasoning about why some users never use (certain) link-up-conversions or about why some users frequently use but never complete specific services. Such insight should be valuable input to the design of further developments for a web site. First results provide evidence that there are both positive and negative relationships between user clusters and different conversions. This is not limited to the content of calls to action but extends to the way they are presented to the customer.

Reducing this to practice, it is plausible to supposse thatsome users are less inclined to follow calls for actions that demand a high level of engagement. Especially when suggested link-up-conversions do not really meet users needs from a content perspective. For example, a car enthusiast with no buying intention might refuse to register to a manufacturer's webpage via a pop up window. This sort of pushy design components will only appeal to users with a high propensity to engage with a certain service. Where recent buyers, eager to participate in the customer care program, will frankly follow this call for action, auto fans will be more likely to engage in a micro game on the page and might end up registering in order to be able to save some reward they just gained. A link to the product configurator can be embedded into an emotional video for a first time visitor, or a table comparing technical details of different cars for users that seems to be quite close to taking a final buying decision.

This also works for pull-out-conversions. A user that refuses to buy a product at the end of a configuration process might be invited to finalize the configuration service by adding the configuration to a wishlist. This example shows that the fact that certain user clusters do not complete services may give valuable input for designing alternative pull-out-conversions that determine a meaningful endpoint and not an unfortunate breakup.

Based on the presumption that all relevant user clusters will execute personalized link-up- and pull-out-conversions if they meet their intentions and needs, deficits and potentials can be spotted whenever certain groups hardly execute any conversions.

4 Conclusions

As a conclusion we suggest the following roadmap for a data-driven service design project:

1. Identify user clusters based on their general usage patterns and their intensity of using the website.
2. Describe cluster profiles and potentially run a survey in order to fully understand users' attitudes and intentions.
3. Make a comprehensive list of link-up- and pull-out-conversions and generate a taxonomy that classifies them in terms of what level of engagement they demand.
4. Visualize a web of conversions that allow a personalized orchestration of services. And look for holes in the grid where different types of connections are missing.
5. Look for relationships betwenn clusters and the use of specific services and conversions in order to answer the following questions:
    a. Is there a lack of cluster specific link-up- and pull-out-conversions at different levels of engagement?
    b. Are there specific uses of a service that are not suggested by corresponding link-up- or pull-out-conversions?
    c. Are there link-up-conversions that do not perform according to their marketing purpose?
6. Enable a design process for further development of services and their personalization based on your findings.
7. Find out if the designed components create the suggested conversions in an experimental setting.
8. Use findings and actual tracking data for personalization of link-up- and pull-out-conversions.

References

Hassler, M. Web Analytics. MITP, 2012.

Heinrich, H. Flocke, L. (2014): Customer-Journey-Analyse-Ein neuer Ansatz zur Optimierung des (Online-) Marketing-Mix. In: Heinrich, H. (Eds.): Digitales Dialogmarketing 2014, pp 825-855.

Kohavi, R., Parekh, R.:Ten Supplementary Analyses to Improve E-commerce Web Sites. In: Proceedings of the Fifth WEBKDD Workshop, pp: 29–36.

Kramer, J., Noronha, S. & Vergo, J. (2000): A user-centered design approach to personalization, Communications of the ACM, Vol. 43 No. 8, pp: 45-8.

Markellou, P., Rigou, M., & Sirmakessis, S. (2005): Web personalization for e-marketing intelligence. In S. Krishnamurthy (Ed.): Contemporary research in e-marketing: Volume 1. Hershey. PA. Idea Group.

Enhancing Interoperability of Web APIs with LAV Views

Maria-Esther Vidal, mvidal@ldc.usb.ve, Universidad Simón Bolívar, Venezuela
Simón Castillo, scastillo@ldc.usb.ve, Universidad Simón Bolívar, Venezuela

Existing Open Data initiatives have fostered the publication of a large number of datasets as well as the development of Web APIs to manage these data. Nevertheless, because data providers usually do not follow standards to publish both data and Web APIs, different interoperability problems may arise when they are integrated to meet a given request. We illustrate the benefits of using the Local-As-View approach to define and integrate Web APIs. Additionally, we report on theoretical and empirical results that show the size of the space of alternative Web API compositions that answer a user request. The reported results suggest that LAV is expressive and appropriate to model catalogs of Web APIs, even in the case, Web APIs are replicated and frequently change.

1 Introduction

Open Data initiatives have stimulated the publication of both data and Web APIs. Currently, more than 1,008,591 datasets have been published in the International Open Government Data Search (IOGDS) catalog[1], while more than 11,910 Web APIs are available at the *programmableweb*[2] catalog. Scientific organizations as NASA[3], and Web platforms like Facebook, also offer Web APIs for developers. However, because data providers are autonomous and usually do not follow any publication standard, syntactical and semantic heterogeneity conflicts may exist across both open datasets and Web APIs (Bülthoff & Maleshkova, 2014). For instance, only 75 out of the 11,910 *programmableweb* Web APIs are using existing vocabularies or schemas. Moreover, Web APIs may constantly change, e.g., Facebook and LinkedIn. Additionally, interface and output data can be heterogeneous: for example, FourSquare, Google Places, and Yelp all provide Web APIs that output data of commercial places, but each API requires different input, and the output contains duplicates. In this paper, we describe a framework that is grounded on the Local-As-View (LAV) approach (Levy, Mendelzon, Sagiv, & Srivastava, 1995). This LAV-based framework facilitates the definition and integration of dissimilar data sources, and naturally contributes to enhance interoperability among Web APIs.

The LAV approach has been defined in the context of the mediator-wrapper architecture to provide a flexible and uniform schema to dissimilar data sources whose functionality may frequently change. Using the LAV approach, Web APIs are described using views in terms of concepts from ontologies; views correspond to conjunctive rules where input and output restrictions are represented as binding restrictions of the rules. Likewise, the problem of Web API composition is cast into the problem of LAV query rewriting where user requests are represented as conjunctive queries in the ontology concepts. Although the problem of query rewriting in the LAV approach has shown to be NP-complete (Levy et al., 1995), several approaches have been defined to efficiently enumerate the space of rewritings, e.g., MCDSAT (Arvelo, Bonet, & Vidal, 2006), GQR (Konstantinidis & Ambite, 2011), MiniCon (Halevy, 2001), and to evaluate SPARQL queries against LAV views (Montoya, Ibáñez, Skaf-Molli, Molli, & Vidal, 2014). These approaches can provide the basis for an efficient and scalable implementation of the LAV Web API composition problem.

We illustrate how the space of Web API compositions grows as the number of replicated Web APIs and the input and output restrictions increase. Additionally, we conducted an evaluation and report on the impact that the number of Web

[1] http://logd.tw.rpi.edu/iogds_analytics_2
[2] http://www.programmableweb.com/developers
[3] http://open.nasa.gov/developer/

APIs has on the Web API compositions that can answer a given user request. All these results provide insights about the expressiveness power and the appropriateness of the LAV approach for Web API description and composition.

The rest of the paper is as follows. We describe the mediator-wrapper architecture and existing approaches that solve the LAV problem of query rewriting. Next, we present the LAV approach as a model to describe Web APIs, and report the results of our evaluation. Then, we conclude and finish with a discussion.

2 Related Work

Typically, Web APIs are characterized by: *i*) different answer formats; *ii*) possible incomplete content; *iii*) interfaces that usually impose some input and output restrictions; *iv*) possible redundancy between other Web APIs; and *v*) autonomicity (Bülthoff & Maleshkova, 2014). In the literature, the mediator-wrapper architecture has successfully used to solve the problem of providing access to data from heterogeneous sources like Web APIs (Wiederhold, 1992). In this architecture, a uniform global schema is exported; thus differences in interfaces, schemas, and contents are hidden from the users. Sources are folded by the *wrappers*, which solve mismatches between source local schemas and the global schema. Mediators, on the other hand, are able to receive queries against the global schema and translate them into the local schemas; they are also responsible for gathering the responses produced by the contacted wrappers. Given the characteristics of the Web APIs, we can rely on this architecture to provide a uniform view of their functionalities and response format, e.g., XML files, CSV files, relational or RDF data. The Local-As-View (LAV) (Levy et al., 1995) is a well-known approach to map global and local schemas in the mediator-wrapper architecture; it relies on views to define these mappings, where a view corresponds to a conjunctive query on the predicates that define the concepts of the global schema. LAV views can be easily adapted to changes in the sources. Additionally, input and output restrictions as well as content description can be naturally represented in the views. Thus, as we will illustrate in this paper, the LAV approach provides the basis to represent Web APIs and scale up to the frequent changes that these data sources may suffer.

Furthermore, user requests that require the solution of the composition of Web APIs problem are modeled as conjunctive queries over the concepts in the global schema, in a way, that this problem is cast as the problem of rewriting a query in terms of a set of views, i.e., the Query Rewriting Problem (QRP) (Levy et al., 1995). Several approaches have been defined to efficiently enumerate the LAV query rewritings; for example, MCDSAT (Arvelo et al., 2006), GQR (Konstantinidis & Ambite, 2011), and MiniCon (Halevy, 2001). Recently, in the context of the Semantic Web, Izquierdo et al. (Izquierdo, Vidal, & Bonet, 2010) propose a logic-based solution to efficiently select the services that best meet a user requirement. This solution also adopts the LAV approach to represent services, while user requests are expressed as conjunctive queries on these concepts. Additionally, users can describe their preferences, which are used to rank the rewritings. We rely on these approaches to support our proposal of using the LAV approach as an expressive formalism to describe Web APIs. Likewise, Montoya et al. (Montoya et al., 2014) have proposed SemLAV, an approach able to answer conjunctive queries against views defined using the LAV approach. SemLAV computes a ranked set of relevant views which are materialized by wrappers. Each time a new view is fully materialized, the original query is executed to deliver results as fast as possible. A demonstration of SemLAV published at (Folz, Montoya, Skaf-Molli, Molli, & Vidal, 2014) clearly shows how the LAV approach can be used to integrate data from the Deep Web and the Linked Data.

Finally, to bridge the gap between Open Data sources and Web APIs, different approaches have been proposed (Pedrinaci & Domingue, 2010; Speiser & Harth, 2011; Taheriyan, Knoblock, Szekely, & Ambite, 2012). These approaches attempt at describing data sources or Web APIs in a way that they can be discovered and integrated into existing federations. Particularly, Taheriyan et al. (Taheriyan et al., 2012) propose Karma, a system to semi-automatically generate source descriptions. Using Karma, contents of structured data sources and Web APIs are described in terms of a given ontology using LAV views. Karma has been used to generate source descriptions of a variety of open datasets and Web APIs; for

instance, recently, data from the Smithsonian American Art Museum has been described as LAV views and made available through a SPARQL endpoint (Szekely et al., 2013). Furthermore, Harth et al. (Harth, Knoblock, Stadtmüller, Studer, & Szekely, 2013) describe an application that exploiting source descriptions provided by Karma and the Data-Fu Linked Data engine (Stadtmüller, Speiser, Harth, & Studer, 2013) is able to integrate on-the-fly, static and dynamic data provided by Web APIs. Based on the state-of-the-art approaches, there are evidence that the LAV approach can be used to describe Web APIs. In the next sections, we will elaborate a bit more this assumption.

3 The Local As View Approach as A Service Description Model

3.1 The LAV Service Description Model

We illustrate the expressiveness power of the LAV approach using the travel domain presented by Izquierdo et al. (Izquierdo et al., 2010). Consider an ontology that contains information about flight and train trips between cities and information about which cities are in Europe. The ontology is comprised of the predicates: *trip*, *flight*, *train*, and *europeanCity*. The first predicate relates cities (x, y) if there is a direct trip either by plane or train between them. The *flight* predicate relates (x, y, t) whenever there is a direct flight from x to y operated by airline t, and similarly, for *train*. The predicate *europeanCity* indicates if a given city is or not an European city. The ontology axioms capture two subsumption relations:

$$Axiom1 : \textit{flight}(x, y, t) \sqsubseteq \textit{trip}(x, y).\ Axiom2 : \textit{train}(x, y, t) \sqsubseteq \textit{trip}(x, y).$$

For the Web API, we assume the following description of the corresponding Web APIs:

- *reg-flight*(x, y) relates two European cities that are connected by a direct flight,
- *from-ka*(x) tells if there is a flight from Karlsruhe to x with the airline operator AB,
- *reg-train*(x, y) relates European cities that are connected by a direct train.

Based on the ontology concepts, the services are described using the following LAV views:

$$\begin{aligned} \textit{reg-flight}(x, y) &:- \textit{flight}(x, y, t),\ \textit{europeanCity}(x),\ \textit{europeanCity}(y)\ . \\ \textit{from-ka}(x) &:- \textit{flight}(\text{KA}, x, \text{AB})\ . \\ \textit{reg-train}(x, y) &:- \textit{train}(x, y, t),\ \textit{europeanCity}(x),\ \textit{europeanCity}(y)\ . \end{aligned}$$

Formally, a catalog of Web APIs (*WC*) corresponds to a triple *WC*=$\langle O, W, M\rangle$ where O is an ontology, W is a set of Web APIs, and M is a set of LAV mappings. Thus, the properties of the data produced by a Web API are expressed with the LAV paradigm in terms of mappings that describe this API in terms of concepts in the domain ontology(Ullman, 2000). Each mapping corresponds to a *conjunctive query* on the predicates that represent the concepts in the ontology. For example, the rule *reg-flight*(x, y) is defined in terms of three predicates *flight*(x, y, t), *europeanCity*(x), and *europeanCity*(y). The variable x corresponds to the origin European city of the direct flight, while y models the destination city; the variable t takes values of airline operators. The conjunction of these three predicates indicates that the output of the Web API is a pair (x,y) where both variables are instantiated with European cities that are connected by a directed flight. Moreover, if a new Web API pops up, a new view is added to the catalog. Additionally, if a Web API changes, only its corresponding view needs to be modified. For instance, suppose the service *reg-train* is extended, and now it retrieves train trips between cities around the world; then, the corresponding view is redefined as follows:

$$\textit{reg-train}(x, y)\ :- \textit{train}(x, y, t)\ .$$

Finally, LAV can also describe input and output restrictions of a Web API, i.e., the binding restrictions of the API. Suppose

that $S(x, y)$ that returns information about flights originating at a given European city, then $S(x, y)$ can be described as:

$$S(\$x, y) \ :- \textit{flight}(x, y, t), \textit{europeanCity}(x) .$$

where the symbol '$' denotes that x is an input attribute of S. The output of the API S denoted by the variable y, will be instantiated with values of cities to which exist a direct flight from the European city represented by the variable x.

3.2 The Service Composition Problem

A user request is expressed as a conjunctive query Q on the predicates that represent the concepts of the ontology O. Consider a user that needs to select the Web APIs able to produce one-way trips between two European cities with at most one stop. The following conjunctive query represents this request:

$$Q(x, y) \ :- \textit{europeanCity}(x), \textit{europeanCity}(y), \textit{trip}(x, y) .$$

Any rewriting of the ontology predicates in terms of the Web APIs defined by the predicates in the query, implements the request. In total this query has *eight* rewritings; some of the valid rewritings are the following:

$$I(x, y) \ :- \textit{reg-flight}(x, y), \textit{reg-flight}(y, x) .$$

The execution of $I(x, y)$ returns pairs of European cities x and y, such that, there is a direct flight from x to y, and from y to x. Further, the next combinations of Web APIs also implement that query $Q(x, y)$:

$$I'(x, y) \ :- \textit{reg-flight}(x, u), \textit{reg-flight}(u, y), \textit{reg-flight}(y, x) .$$

$$I''(\mathrm{KA}, y) \ :- \textit{from-ka}(u), \textit{reg-flight}(u, y), \textit{reg-flight}(y, \mathrm{KA}) .$$

$$I'''(x, y) \ :- \textit{reg-flight}(x, \mathrm{KA}), \textit{from-ka}(y), \textit{reg-flight}(y, x) .$$

$$I''''(x, \mathrm{KA}) \ :- \textit{reg-flight}(x, u), \textit{reg-flight}(u, \mathrm{KA}), \textit{from-ka}(x) .$$

3.3 Complexity Issues

Although describing Web APIs using the LAV paradigm increases the expressiveness and facilitates the adaptation of descriptions to changes in the Web APIs, it is important to take into account the complexity that characterizes the query rewriting problem. First, given a conjunctive query Q with n predicates and a catalog *WC* with m Web API descriptions, the decision problem of determining if a rewriting that combines n Web APIs in *WC* is a rewriting of Q is an NP-complete problem (Levy et al., 1995). Additionally, the space size of different rewritings can be impacted by the input and output restrictions of the Web APIs. For example, suppose there is a Web API *reg-flight2* which produces the same results than *reg-flight* but it requires that the origin city of the fly to be bound, i.e., there is an input restriction on this Web API:

$$\textit{reg-flight2}(\$x, y) \ :- \textit{flight}(x, y, t), \textit{europeanCity}(x), \textit{europeanCity}(y) .$$

According to this restriction, the following rules correspond to valid rewritings of query Q, i.e., instead of having one rewriting, there are 2^{n-1} equivalent rewritings whenever no bindings are giving in the query:

$$I_0(x, y) \ :- \textit{reg-flight}(x, u), \textit{reg-flight}(u, y), \textit{reg-flight}(y, x) .$$

$$I_1(x, y) \ :- \textit{reg-flight}(x, u), \textit{reg-flight2}(\$u, y), \textit{reg-flight}(y, x) .$$

$$I_2(x, y) \ :- \textit{reg-flight}(x, u), \textit{reg-flight}(u, y), \textit{reg-flight2}(\$y, x) .$$

$$I_4(x, y) \ :- \textit{reg-flight}(x, u), \textit{reg-flight2}(\$u, y), \textit{reg-flight2}(\$y, x) .$$

In general, suppose a Web API S' has k equivalent Web APIs, each one has different input and output restrictions, and bindings can be received in the query; then the following rewriting will have k^n equivalent query rewritings:

$$I'(\overline{x}) \ :- S(\overline{x_1}) , S(\overline{x_2}) , \ldots , S(\overline{x_n})$$

where $\overline{x_i}$ represents the variables of the *i-th* call of the Web API S, and $\overline{x}$ is the list of variables in the rewriting.
Although existing approaches can efficiently address the problem of query rewriting, none of them considers input and output restrictions. Given the impact that these restrictions have on the space of compositions, existing techniques need to be extended to be able to generate spaces of compositions that respect these restrictions.

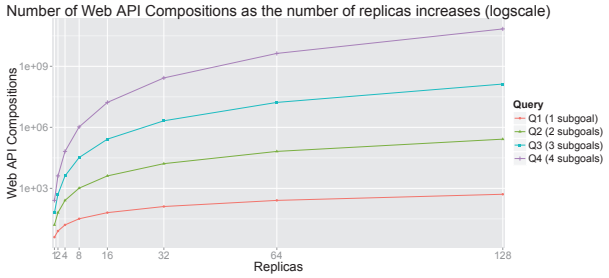
4 Evaluation

We ran a simple test to analyze the size of the space of rewritings; MCDSAT (Arvelo et al., 2006) was used to count the number of valid rewritings. This evaluation provides insights of the size of the spaces of Web APIs compositions that implement the studied requests. For simplicity, we just consider the following query:

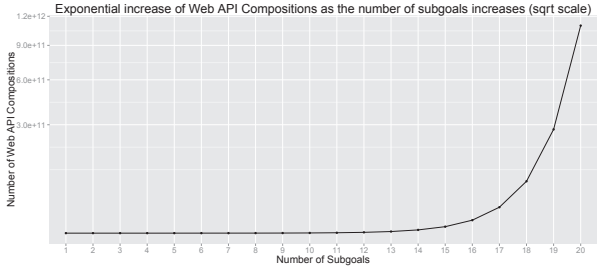
$$Q'(x_1, x_2, x_3, x_4, \ldots, x_n) \ :- \ flight(x_1, x_2, t_1)\,, flight(x_2, x_3, t_2)\,, flight(x_3, x_4, t_3)\,, \ldots\,, flight(x_{n-1}, x_n, t_{n-1})\,.$$

We assume that the following four Web APIs are available. Our research questions are: *i*) *"How is the space of Web API compositions impacted whenever new replicas of these four APIs pop up?"*, and *ii*) *" What is the impact of the size of a user request on the space of Web API compositions?"*.

$$\begin{aligned}
\textit{reg-flight}(x, y) \ &:- \textit{flight}(x, y, t),\ \textit{europeanCity}(x),\ \textit{europeanCity}(y)\,. \\
\textit{reg-flight3}(x, y) \ &:- \textit{flight}(x, y, t),\ \textit{europeanCity}(x)\,. \\
\textit{reg-flight4}(x, y) \ &:- \textit{flight}(x, y, t),\ \textit{europeanCity}(y)\,. \\
\textit{int-flight}(x, y) \ &:- \textit{flight}(x, y, t)\,.
\end{aligned}$$



(a) Number of Web API compositions for query Q' as the number of replicated Web APIs increases (log-scale)



(b) Number of Web API compositions for query Q' as the number of Predicates increases increases (sqrt scale)

Figure 1: Number of Web API Compositions for Query Q'

Figures 1(a) and (b) report on the number of Web API compositions of Q' as the number of Web API replicas and the size of the query increase. In Figure 1(a), we consider that query Q' has up to *four* predicates and the number of *replicas* varies from 1 to 128; we can observe that the space grows smoothly, i.e., the space complexity is *polynomial* in the number of *replicas*. Complementary, Figure 1(b) shows the size of the Web API composition space when the number of *replicas* is 1, and the number of *query predicates* ranges from 1 to 20. In this case, the size of the space is *exponentially* in the size of the query, and the number of compositions is $12 \times e^{24}$ when the query has 20 sub-goals. Because the *complexity* of the query rewriting problem is affected by the *size* of the space of rewritings, these results suggest that the LAV approach is more *suitable* for *large* catalogs where Web APIs are duplicated and *frequently* appear. On the other hand, even in presence of *small* catalogs, the space of Web APIs composition can be extremely *large* depending on the *size* of the user request. Similar *behavior* will be observed whenever the input and output restrictions are considered.

5 Conclusions

We propose the description of Web APIs as LAV views in terms of concepts in domain ontologies. Using the LAV approach, user requests are expressed as conjunctive queries on the concepts in domain ontologies, and the Web API compositions that implement these requests correspond to valid query rewritings of the corresponding queries. This formulation allows us to exploit the properties of state-of-the-art query rewriting approaches to provide an efficient solution. Initial results suggest that the approach can naturally represent real-world problems. We are currently working on the extension of MCDSAT (Arvelo et al., 2006) in order to provide a scalable solution to catalogs of replicated Web APIs and with different input and output restrictions. In the future, we plan to make available a Web API composer able to rank the compositions and produce first those that produce more answers for a request.

References

Arvelo, Y., Bonet, B., & Vidal, M.-E. (2006). Compilation of query-rewriting problems into tractable fragments of propositional logic. In *Aaai* (pp. 225–230).

Bülthoff, F., & Maleshkova, M. (2014). Restful or restless - current state of today's top web apis. In *The semantic web: ESWC 2014 satellite events.*

Folz, P., Montoya, G., Skaf-Molli, H., Molli, P., & Vidal, M. (2014). Semlav: Querying deep web and linked open data with SPARQL. In *The semantic web: ESWC 2014 satellite events.*

Halevy, A. Y. (2001, December). Answering queries using views: A survey. *The VLDB Journal*, *10*(4).

Harth, A., Knoblock, C. A., Stadtmüller, S., Studer, R., & Szekely, P. A. (2013). On-the-fly integration of static and dynamic sources. In *International workshop on consuming linked data (cold).*

Izquierdo, D., Vidal, M., & Bonet, B. (2010). An expressive and efficient solution to the service selection problem. In *The semantic web - ISWC 2010.*

Konstantinidis, G., & Ambite, J. L. (2011). Scalable query rewriting: a graph-based approach. In T. K. Sellis, R. J. Miller, A. Kementsietsidis, & Y. Velegrakis (Eds.), *Sigmod conference* (p. 97-108). ACM.

Levy, A. Y., Mendelzon, A. O., Sagiv, Y., & Srivastava, D. (1995). Answering queries using views. In *Proceedings of the fourteenth acm sigact-sigmod-sigart symposium on principles of database systems* (p. 95-104). ACM.

Montoya, G., Ibáñez, L. D., Skaf-Molli, H., Molli, P., & Vidal, M.-E. (2014). SemLAV: Local-As-View Mediation for SPARQL. *Transactions on Large-Scale Data- and Knowledge-Centered Systems XIII, LNCS, Vol. 8420*.

Pedrinaci, C., & Domingue, J. (2010). Toward the next wave of services: Linked services for the web of data. *J. UCS*, *16*(13), 1694-1719.

Speiser, S., & Harth, A. (2011). Integrating linked data and services with linked data services. In *Extended semantic web conference (eswc)* (p. 170-184).

Stadtmüller, S., Speiser, S., Harth, A., & Studer, R. (2013). Data-fu: a language and an interpreter for interaction with read/write linked data. In *The international world wide web conference (www)* (p. 1225-1236).

Szekely, P. A., Knoblock, C. A., Yang, F., Zhu, X., Fink, E. E., Allen, R., et al. (2013). Connecting the smithsonian american art museum to the linked data cloud. In *Extended semantic web conference (eswc)* (p. 593-607).

Taheriyan, M., Knoblock, C. A., Szekely, P. A., & Ambite, J. L. (2012). Rapidly integrating services into the linked data cloud. In *International semantic web conference (iswc)* (p. 559-574).

Ullman, J. D. (2000). Information integration using logical views. *Theoretical Computer Science*, *239*(2), 189-210.

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *IEEE Computer*, *25*(3), 38-49.

VISION PAPER

Bottom-up Web APIs with self-descriptive responses

Ruben Verborgh, Erik Mannnens, Rik Van de Walle
{firstname.lastname}@ugent.be – Ghent University – iMinds, Belgium

The success or failure of Semantic Web services is non-measurable: many different formats exist, none of them standardized, and few to no services actually use them. Instead of trying to retrofit Web APIs to our models, building APIs in a different way makes them usable by generic clients. This paper argues why we should create Web APIs out of reusable building blocks whose functionality is self-descriptive through hypermedia controls. The non-functional aspects of such APIs can be measured on the server and client side, bringing us to a more scientific vision of agents on the Web.

1 The failed promise of Semantic Web services

The initial Semantic Web vision talks quite expressively about how *intelligent agents* will improve our lives by doing things on the Web for us (Berners-Lee, Hendler, & Lassila, 2001). Unfortunately, after roughly 15 years of Semantic Web research, nothing even vaguely resembling the envisioned agents has come out of our research community. An important reason seems the apparent failure of Semantic Web services (Pedrinaci, Domingue, & Sheth, 2011) to gain any traction or usage. Even if there were agents, they would not have any services at their disposal. Companies such as Apple and Google have released software that seemingly acts as a remarkably clever agent; however, each binding with a Web service appears hand-made and is thus still far away from agents that dynamically interact with automatically discovered services. It seems that one of the only successes of the Semantic Web is—paradoxically—intelligent *servers*, with SPARQL endpoints (Feigenbaum, Williams, Clark, & Torres, 2013) providing ad-hoc queryable semantic knowledge. Sadly, even that can hardly be called a success: we found out the hard way that such omnipotent Web services result in an unacceptably low availability (Buil-Aranda, Hogan, Umbrich, & Vandenbussche, 2013), making them one of the likely contributors to the rather negative external perception of Semantic Web technologies. After all, if the average "intelligent" server is unavailable for more than 1.5 days each month, how much remains for the intelligent agents?
Somewhere, something went horribly wrong. The main culprit might well be the way we have been building Web services so far. Oddly enough, "Web" services have surprisingly little to do with the Web. The Web, at its core, is a distributed hypermedia system (Fielding, 2000), a collection of three separate inventions that closely connect together:

- Resources on the Web are identified by a URL.
- Through the HTTP protocol, we can retrieve a representation of a resource via its URL.
- Representations in a hypermedia format such as HTML contain the URLs of related resources.

This recursive mechanism, characteristically driven by hypermedia, sets the Web apart from all other information systems. Traditional Semantic Web services, on the other hand, use the Web's core mechanisms in unrelated ways:

- A URL identifies an endpoint (i.e., a purely technical artefact of a process or implementation, instead of information).
- An HTTP message serves as an envelope to call procedures on that endpoint.
- Hyperlinks are not relevant at all; URLs are hard-coded, sometimes in separate description documents.

Indeed, such services treat the Web simply as a black box to perform remote-procedure calling (RPC); the usage of HTTP and URLs is merely an artefact to have things working through TCP port 80. The same functionality can be (and has been) recreated with, for instance, the SMTP e-mail protocol (Cunnings, Fell, & Kulchenko, 2001). Therefore, traditional Web services are by no means native Web citizens at all, sharing none of their principles with the rest of the Web.

Web-compliance by itself could be seen as simply a matter of technical purity, but the opposite is true. If we create such an artificial world within the Web, it is quite meaningless to talk about Semantic *Web* agents. After all, to what extent do they belong to the field of Web research if they simply execute pre-programmed remote procedures over TCP?

2 The false hope of service descriptions

From their inception, Semantic Web services had been associated with verbose descriptions. With Web services in essence being firewall-friendly abstraction layers over RPC, a mechanism to describe the *semantics* of such procedures was necessary. Both OWL-S and WSMO claimed technical superiority (Lara, Roman, Polleres, & Fensel, 2004), but none of the proposals ever achieved W3C standardisation, let alone adoption or usage. Furthermore, the manual work to generate such descriptions by far outweighs their benefits, given the non-existence of agents that could use them.

In an effort to reduce the verbosity surrounding Web services, the service landscape moved towards services with a smaller payload, abandoning the XML domination that had reigned so far. Instead of implementing a protocol such as SOAP *on top* of HTTP, those Web APIs directly operate through the HTTP protocol. This gave rise to a new generation of service descriptions such as Linked Open Services (Krummenacher, Norton, & Marte, 2010), and Linked Data Services (Speiser & Harth, 2011). Such descriptions accept the heterogeneity of HTTP interfaces and employ so-called *lifting* and *lowering* to convert between the non-RDF responses of the interfaces to the RDF model and vice-versa. The descriptions themselves rely on graph patterns to capture input and output patterns, indicating that an RPC way of thinking (a method call) still underpins their design. While this is not technological burden, it does not bring us closer to a vision of agents on the *Web*.

To realize this, we have to look at the Web's architectural properties and the exact point where it currently fails for machines. Contrary to what current practice seems to imply, machines have no inherent need for RPC interfaces; they are certainly capable of consuming hypermedia interfaces as offered by the Web. What they cannot do at present, is parsing the natural language found in the majority of Web documents. Hence, machines need information in a machine-interpretable model such as RDF, but this does not warrant an entirely different interface. Instead, through the content negotiation feature of HTTP, the same conceptual HTTP resources and URLS can be shared between machine and non-machine interfaces, with only the representation having a different format (Verborgh et al., 2015). Clients consume such hypermedia APIS by navigating the links inside of their representations. However, since links only allow to look ahead one step at a time, automated clients have difficulties performing multi-step tasks. To mitigate this, functional hypermedia API description formats such as RESTdesc (Verborgh et al., 2012) explain a hypermedia API's resources by detailing the affordances offered by its links. Data-Fu (Stadtmüller, Speiser, Harth, & Studer, 2013) follows the example of RESTdesc by using rules and the HTTP vocabulary to deal with remote functionality. However, its rules live on the client side, so it is unclear to what extent it can withstand change on the server side—something that is a frequent practice on the Web.

There are three major issues with all of the above approaches, regardless of whether they deal with RPC or hypermedia:

1. Creating descriptions or rules remains manual work, which will realistically not be undertaken given the lack of a single standard and actually implemented use cases.
2. They all rely on the premise that *reasoning* will fix all remaining gaps. If there are ontological differences between the desired goal and/or different services or descriptions used, such issues are implicitly assumed to be trivial and entirely solvable by existing reasoning technologies. As a result, pressing problems remain unresolved.
3. No scientific methodology exists to compare the different description techniques. Most of the aforementioned technologies have claimed to be better than one another, but such claims lack scientific evidence. No metrics have been defined, and existing evaluations (e.g., Stadtmüller et al., 2013) examine the performance of a resulting system, but fail to give a quantifiable measure of the appropriateness of any Web interface as such. Yet the Semantic Web research community needs quantifiable results upon which we can build (Bernstein & Noy, 2014).

3 Fostering reusability through a self-descriptive bottom-up approach

Lacking better measurements, the Web API community has been heading the same quantity-over-quality course that has characterized the first years of the Linked Data initiative. An often-quoted fact in Web API papers and articles is the ever increasing number of Web APIs (Figure 1), which is supposed to be an indicator of the ecosystem's excellent health. However, as Linked Data researchers have become painfully aware, quantity only loosely correlates with quality or usefulness. Perhaps for Web APIs, the correlation between quantity and utility could even be negative. Few other communities would pride themselves on the existence of more than 12.000 different micro-protocols to achieve essentially the same thing: communicating between clients and servers over HTTP. Of course, each application has its own domain and domain-specific vocabulary, but does that also warrant an entirely different way of exposing this, especially when we have RDF as a uniform data model? Each different API currently requires a different client, given the lack of a uniform API description format to explain the API's response structure and functionality. Clearly, this approach to Web APIs is a dead end.

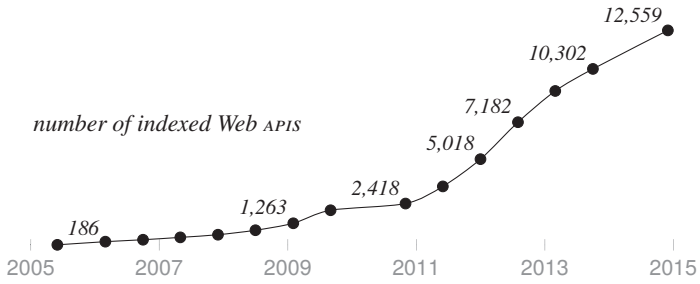


Figure 1: The increasing number of Web APIs is often named an indicator of their success, while the overgrowth of such *custom* micro-protocols is unnecessary—and detrimental to the development of *generic* Web API clients. *(data: programmableweb.com)*

In order for machines to use information autonomously, it has to be composed out of pieces they can recognize and interpret. The RDF model achieves this by identifying each of the triple components by reusable IRIs, which have a meaning beyond the scope that mentions them. Furthermore, the Linked Data principles mandate the use of HTTPURLs, which turn these components into affordances toward relevant information. For instance, given the following RDF triple:

```
<http://dbpedia.org/resource/Bill_Clinton> <http://xmlns.com/foaf/0.1/knows>
    <http://dbpedia.org/resource/Al_Gore>.
```

the knowledge of the `foaf:knows` predicate is sufficient for a machine to determine that this relation is symmetric, and that `dbpedia:Bill_Clinton` and `dbpedia:Al_Gore` are instances of `foaf:Person`—even though it might have never encountered any of those IRIs before. Furthermore, should the `foaf:knows` property be unfamiliar, its IRI can be dereferenced to find this information expressed in ontological predicates. Knowledge of these predicates in turn allows an interpretation of `foaf:knows` and hence the aforementioned derivation. We herein recognize two characteristics in particular:

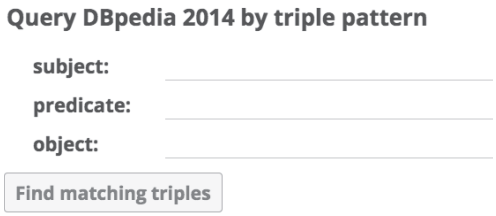
- The information is structured in a **bottom-up** way: machines interpret a larger unit of information through its pieces instead of interpreting the pieces through the whole (while humans are capable of doing both simultaneously).
- Each piece in the unit is **self-descriptive**: anything needed to interpret a piece is contained within itself, with its IRI acting as both an identifier and a direct handle towards additional interpretation mechanisms. No external resource is required beforehand, given the knowledge of a limited set of basic concepts.

This sharply contrasts with current practice for Web APIs. Machines are assumed to interpret each API operation in its entirety, as such smaller pieces do not exist, and API descriptions—if present—are external documents that must be collected and interpreted before consumption is possible. While this does not imply the inviability of such an approach, it raises serious doubt as to whether that is the most effective strategy towards automated Web API consumption by generic clients.

3.1 Making APIs self-descriptive

In order to answer the question of what bottom-up, self-descriptive Web APIs should look like, we can find inspiration in how the human Web works. Self-descriptiveness exists on two levels: on the one hand, humans understand natural language, so written texts can guide them through a webpage. On the other hand, they interact with *hypermedia controls* such as links and form elements provided through the `<a>` and `<input>` HTML tags, which avoids the need to read a separate document to find out how to craft HTTP requests against a particular server.

The first level of self-descriptiveness is provided by RDF: instead of describing content in natural languages, machines retrieve a machine-interpretable representation of the response—similar to how French-speaking people would receive a French-language representation. Analogously, we can also apply the same strategy of hypermedia controls to address the second level. Instead of providing the HTTP controls in HTML, we need a hypermedia ontology for RDF such as the Hydra Core Vocabulary (Lanthaler, 2013). Figure 2 shows how an HTML form can be represented in RDF, preserving the exact same semantics. Note in particular how each of the English field labels corresponds to an RDF property, which machines can apply like any other RDF construct. Furthermore, the semantics of "searching parts in a larger set" is expressed using the `hydra:search` predicate, so that an automated client can interpret the consequences of its actions.




```
<http://fragments.dbpedia.org/2014/en#dataset> hydra:search [
    hydra:template "http://fragments.dbpedia.org/2014/en{?s,p,o}";
    hydra:mapping
        [ hydra:variable "s"; hydra:property rdf:subject ],
        [ hydra:variable "p"; hydra:property rdf:predicate ],
        [ hydra:variable "o"; hydra:property rdf:object ]
    ].
```

Figure 2: An HTML form can be expressed in an RDF equivalent through the Hydra Core Vocabulary.

3.2 Building APIs from the bottom up

While humans adapt much better to unknown conditions than machines, we often also resort to recognizable patterns in unfamiliar environments such as new Web pages. They include frequently reused types of forms, such as search or contact forms, and usability conventions such as underlined hyperlinks. To appreciate the strength of such factors, consider how we have become accustomed to the fairly recent trend of clicking the main logo on any subpage to navigate to the website's homepage (Cappel & Huang, 2007).

Given the Web API ecosystem's total lack of reuse on the macro level, it is not surprising that reuse is lacking on the micro level as well. For instance, despite being highly similar in intent, each social media site offers a radically different API to post an updated status message—even though the human counterparts in HTML look and behave fairly similar. This is unintuitive, as machines actually need *more* structure than humans, precisely because they experience increasing difficulty when adapting to new situations.

The need for reusable building blocks can also be addressed by hypermedia controls, given that they are carefully designed in a progressive way. Again, the structure of webpages can inspire us here: we should provide smaller groups of hypermedia controls than can be reused in different combinations, with a preference of giving more hypermedia controls to users (and certainly not less). For instance, a common practice on current Web APIs is to have a dedicated entry point, which exposes significantly more functionality than other resources. However, this contrasts with websites: while there is indeed an index page (often located at `/`), usually *all* of the pages carry navigational controls that provide access to the entire website. In that sense, no page is more special than any other. Another issue is that people often omit "trivial" metadata which they would include for humans: how many result pages there are, how these results can be filtered and sorted, and so on. If we want to see intelligent generic clients, server have to *enable* such intelligent behavior by providing the necessary affordances instead of making use-case-specific assumptions about their necessity.

4 Transforming Web API engineering into measurable research

Such self-describing, bottom-up APIs bring us to the question on how a particular approach can be evaluated from a research perspective. In other words, we need to define objectively quantifiable metrics such that two or more Web APIs used for the same task can be compared in a reproducible manner. If such metrics do not exist, as has been the case in the majority of Web API papers and articles so far, no matter how good the engineering contributions are, there is no way for others to analyze or improve existing solutions (Bernstein & Noy, 2014).

As applied in our previous work on a self-describing, bottom-up Web API (Verborgh et al., 2014), we propose to measure Web APIs as follows. First, since *backward compatibility* (defined here as the ability of a client to interact with future versions of an API) cannot be quantified generically, we require full backward compatibility if a modified version of a Web API is proposed. Furthermore, the amount of out-of-band knowledge needed to consume the same part of an API cannot increase. Should these constraints not be satisfied, the new and old APIs must be considered different and any improvements are not characteristics of the original API. Second, APIs cannot be evaluated in isolation, as their non-functional characteristics depend on a particular use case. As such, when discussing the results of API evaluations, *external validity* should be thoroughly argued. Finally, the measurements must take into account quantifiable parameters, such as:

- The **average number of HTTP requests** it takes for a client to solve (a particular task of) a use case.
- The **average response size**.
- The **average response time** of the server.
- The **amount of processor and memory usage** on the client and server in function of parameters such as the concurrent number of clients and the complexity of use case tasks.
- The **cache hit rate** of responses.
- The **total completion time** of (particular tasks of) a use case.

Backward compatibility is an especially important criterion to avoid artificial manipulation of the above measurements. For instance, removing hypermedia controls from responses would certainly improve the average response size and, consequently, also the response time of the server. Yet this would mean an increase in out-of-band knowledge and/or breakage of previous clients. In order to progress toward an objectively verifiable and improvable scientific discipline, the utilization of such idiosyncratic shortcuts must be ruled out.

Figure 2 depicts the hypermedia form of a real-world hypermedia API that provides domain-agnostic access to triples. Following the principles above, we have analyzed the concrete use case of SPARQL query execution through this triple pattern fragments API (Verborgh et al., 2014) and the SPARQL protocol (Feigenbaum et al., 2013). Backwards compatibility is ensured by self-descriptiveness and the API's bottom-up structure. For instance, the current API only allows lookups with exact matching of triple components. Should the API be extended with full-text search, an additional hypermedia control would be added to facilitate this. That would guarantee compatibility with existing clients (since the triple-pattern "building block" continues to exist), while at the same time, clients that can interpret the full-text control (the new "building block") can realize a faster execution time for several queries. In this particular case, we have shown that the choice of interface influences the trade-offs between some of the parameters listed above: SPARQL query answering through a triple pattern fragments interface requires significantly more HTTP requests compared to when the server offers a SPARQL interface, but responses are more cacheable and thus withstand a growing number of clients much better.

The example indicates how self-describing, bottom-up APIs allow a gradual evolution, with different clients using different parts of the interface—just like humans do on the Web. Reusing such building blocks ensures that similar use cases can be tackled with recurring strategies. Furthermore, given the evaluation strategy introduced above, the effectiveness or fitness of a Web API for a certain use case can be quantified. This enables a more meaningful debate in the Web API community, beyond a hollow description format discussion that only pushes the vision of autonomous agents on the Web further away.

References

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The Semantic Web. *Scientific American*, *284*(5), 34–43. Retrieved from http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Bernstein, A., & Noy, N. (2014). *Is this really science? The Semantic Webber's guide to evaluating research contributions* (Tech. Rep. No. IFI-2014.02). Department of Informatics, University of Zurich. Retrieved from https://www.merlin.uzh.ch/contributionDocument/download/6915

Buil-Aranda, C., Hogan, A., Umbrich, J., & Vandenbussche, P.-Y. (2013, November). SPARQL Web-querying infrastructure: Ready for action? In *Proceedings of the 12th international semantic web conference.* Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-41338-4_18

Cappel, J. J., & Huang, Z. (2007). A usability analysis of company websites. *Journal of Computer Information Systems*, *48*(1), 117–123.

Cunnings, R., Fell, S., & Kulchenko, P. (2001, November). *SMTP transport binding for SOAP 1.1* (Tech. Rep.). Retrieved from http://www.pocketsoap.com/specs/smtpbinding/

Feigenbaum, L., Williams, G. T., Clark, K. G., & Torres, E. (2013, March 21). *SPARQL 1.1 protocol* (Recommendation). W3C. Retrieved from http://www.w3.org/TR/sparql11-protocol/

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures* (Doctoral dissertation, University of California). Retrieved from http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Krummenacher, R., Norton, B., & Marte, A. (2010). Towards Linked Open Services and processes. In *Future internet symposium* (Vol. 6369, pp. 68–77). Springer. doi: 10.1007/978-3-642-15877-3_8

Lanthaler, M. (2013). Creating 3rd generation Web APIs with Hydra. In *Proceedings of the 22nd international conference on world wide web companion* (pp. 35–38). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from http://dl.acm.org/citation.cfm?id=2487788.2487799

Lara, R., Roman, D., Polleres, A., & Fensel, D. (2004). A conceptual comparison of WSMO and OWL-S. In L.-J. Zhang & M. Jeckle (Eds.), *Web services* (Vol. 3250, pp. 254–269). Springer.

Pedrinaci, C., Domingue, J., & Sheth, A. (2011). Semantic Web services. In J. Domingue, D. Fensel, & J. Hendler (Eds.), *Handbook of Semantic Web technologies* (pp. 977–1035). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-92913-0_22 doi: 10.1007/978-3-540-92913-0_22

Speiser, S., & Harth, A. (2011). Integrating Linked Data and services with Linked Data Services. In *The Semantic Web: Research and applications* (Vol. 6643, pp. 170–184). Springer. doi: 10.1007/978-3-642-21034-1_12

Stadtmüller, S., Speiser, S., Harth, A., & Studer, R. (2013). Data-Fu: a language and an interpreter for interaction with read/write linked data. In *Proceedings of the 22nd international conference on world wide web* (pp. 1225–1236). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from http://dl.acm.org/citation.cfm?id=2488388.2488495

Verborgh, R., Hartig, O., De Meester, B., Haesendonck, G., De Vocht, L., Vander Sande, M., … Van de Walle, R. (2014, October). Querying datasets on the Web with high availability. In P. Mika et al. (Eds.), *Proceedings of the 13th international semantic web conference* (Vol. 8796, pp. 180–196). Springer. Retrieved from http://linkeddatafragments.org/publications/iswc2014.pdf doi: 10.1007/978-3-319-11964-9_12

Verborgh, R., Steiner, T., Van Deursen, D., Coppens, S., Gabarró Vallés, J., & Van de Walle, R. (2012, April). Functional descriptions as the bridge between hypermedia APIs and the Semantic Web. In *Proceedings of the third international workshop on RESTful design* (pp. 33–40). Retrieved from http://ws-rest.org/2012/proc/a5-9-verborgh.pdf

Verborgh, R., van Hooland, S., Cope, A. S., Chan, S., Mannens, E., & Van de Walle, R. (2015, March). The fallacy of the multi-API culture: Conceptual and practical benefits of representational state transfer (REST). *Journal of Documentation*, *71*(2). Retrieved from http://freeyourmetadata.org/publications/named-entity-recognition.pdf

Towards Pervasive Web API-based Systems

Felix Leif Keppmann, felix.leif.keppmann@kit.edu, Karlsruhe Institute of Technology
Maria Maleshkova, maria.maleshkova@kit.edu, Karlsruhe Institute of Technology

Recent technology developments in the area of services on the Web are marked by the proliferation of Web applications and APIs. This development, accompanied by the growing use of sensors and mobile devices raises the issue of having to consider not only static but also dynamic data accessible via Web APIs. New implications on the communication in such networks can be derived, i.e., the active role of pulling or pushing data is no longer exclusively assigned to specific roles, e.g. services, or clients. Furthermore, the establishing of data flow between devices in such a network may be initiated by any of the participating devices. In this paper we present a general approach for the modelling of a service-based systems, which takes into consideration the providing/consuming of dynamic data sources. In particular, we develop a formal model for capturing the communication between Web APIs and client application, by including dynamic data producing services and data consuming client applications. We complement our model with a decision function for optimising the pull/push communication direction and optimise the amount of redundant transferred data (i.e. data that is pushed but cannot be processed or data that is pulled but is not yet updated). The presented work lays the foundation for creating intelligent Web API-based systems and client application, which can automatically optimise the data exchange.

1 Introduction

Web services provide means for the development of open distributed systems, based on decoupled components, by overcoming heterogeneity and enabling the publishing and consuming of data and functionalities between applications. Recently the world around services on the Web, thus far limited to "classical" Web services based on Simple Object Access Protocol (SOAP) and Web Services Description Language (WSDL), has been enriched by the proliferation of REST services, when their Web Application Programming Interfaces (APIs) conform to the Representational State Transfer (REST) architectural principles (Richardson & Ruby, 2007). These Web APIs are characterised by their relative simplicity and their natural suitability for the Web, relying directly on the use of Uniform Resource Identifier (URI), for resource identification and Hypertext Transfer Protocol (HTTP) for service interaction and message transmission.

At the same time, emerging complex and distributed systems composed of nodes with heterogeneous hard- and software characteristics face new challenges. For example, the wider use of sensors and mobile devices raises the issue of having to consider not only static but also dynamic data. However, traditional client-server based architectures are not optimal for supporting the consumption and provision of static and dynamic data alike. Furthermore, devices seen as nodes in a network, cannot be directly identified as data consumers or provides based solely on their role as server or client application. This has implications on the overall communication between the nodes. As a result, mapping the data flow in a distributed complex scenario to the Web pull-push-based communication becomes a challenge. While both, push and pull, enable the data flow between nodes, one may be less efficient in terms of transmitting redundant data, with impact on the overall system.

In this work we address some of the issues appearing in these scenarios. In particular, we provide: 1) a model to capture a network of data producing and consuming nodes with their relevant properties, 2) a decision function to optimise the communication between services and clients and thus determine, which nodes in the network should be actively initiating the communication.

2 Motivation

In order to establish a more specific notion of the problem we use the scenario of a house monitoring system. In this example different sensors serve as services, e.g. as Web API, and provide their sensor data on the network. Client applications monitor the sensors and allow end users to access and visualise the sensor data. Due to the nature of the sensors, the rate, with which their services register and provide new data, i.e. the sensor's update frequencies, differs. Similarly, the characteristics of the monitoring clients differ in their update rates, for instance in the type of visualisation (e.g., highly frequent display or rarely updated web-based map). Some of the services and clients are connected by a data flow, thus the client requires data from the service to provide a certain functionality.

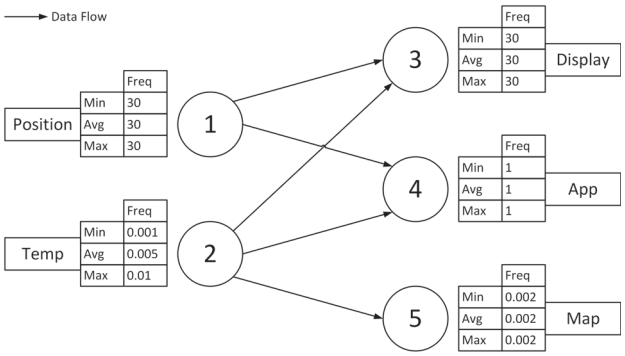


Figure 1: House Monitoring System Scenario

We introduce our monitoring system scenario by looking at the communication and data flow in terms of a network. This network incorporates services, which can produce or simply provide access to data, and clients, which consume data. A service may be perceived as static (e.g., a blueprint of a house) or as dynamic (e.g., a video or a temperature sensor). In addition, a service has an update frequency (over a certain time span), which may be zero in the case of static data, or several times per seconds, in case of dynamic data. Clients can consume data from a data source very rarely (e.g., only once to create a map) or up to several times in short intervals (e.g., multiple times per second to update a visualisation). Thus clients are also characterised by a specific update frequency, depending on the functionality that they provide.

The result is a network of communicating services and clients as nodes, characterised by their update frequencies, and the connections between the nodes, based on data production and consumption, visualised in Figure 1. In case of a data flow a connection between a service and a client is placed. Furthermore, some nodes expose data at a constant update frequency, while others, such as the temperature sensor, expose a minimum, maximum and average update frequency.

We use the scenario of a house monitoring system as the basis for deriving requirements for a formal model that is capable of capturing services and client applications in terms of a data-driven communication network. In a scenario, as the one described above, services and clients communicate to transfer data from a service to a client. Thereby, two basic communication directions can appear – the client requests data from the service, i.e. pull, or the service actively sends data to the client, i.e. push. In both cases the data, which is transferred in the messages, may be redundant. On the one hand, a message contains the same data if a client requests data from a service, which has not been updated since the last request. On the other hand, data is discarded if a service sends data to a client, which at that time cannot be processed by the client. How often the data of services is updated and how often clients are able to process data are represented by their update frequencies.

Based on the scenario and the described characteristics of the network nodes, we can derive requirements for defining a model that supports the optimisation of the data-driven communication between data services and clients. In particular, the model should enable: 1) the minimisation of redundant transferred data contained in messages. Moreover, it should 2) respect the flow of data in the network of services and clients, and 3) optimise the direction of active communication between nodes in the network (services that actively provide data, i.e. push, vs. clients that actively request data, i.e. pull). Since they express the capabilities to provide and request data, 4) the frequencies of services and clients must be taken into account.

3 Communication Model

In this section we introduce our communication model, which captures all nodes, the data flow in the network, as well as the different update frequencies of services and clients. Based on both, the data flow and the frequencies, a decision function is introduced, which support the optimisation of the pull-push directions in the network. The optimised pull-push directions result in minimising the volume of redundant data that is being transferred, therefore, reducing the overall data exchange volumes. In the following we first elaborate on important preliminaries and definitions, second we introduce the communication model and finally, we describe the decision function.

The following preliminaries apply, in order to keep the function and the model of the basic approach simple. Nodes in the network may play the role of a service, client, or both service and client and expose one frequency, i.e. data consumption and/or data provisioning update frequency. This frequency is, in case of a service, the update frequency of new data provided on the network. In case of a client, it is the frequency, with which the client requires data coming from a service. Nodes acting as a service and a client at the same time have only one frequency for both.

The following definitions apply. The number of events per unit time is called frequency. In this context we use the unit Hertz (Hz) for frequency, defined as number of events (e.g. new data provided by a service) per second $f = \frac{k}{t}$. While k is the number of events and t is the time in seconds. It can be measured by counting the number of occurrences of an event for a given period of time. All nodes in a network are numbered by $1, \ldots, n$ with $n \in \mathbb{N}$. Each node in the network may play the role of a service, denoted by S, or client, denoted by C. A node in a specific role is denoted by S_i or C_j with $1 \leq i, j \leq n, n \in \mathbb{N}$. Each node in the network exposes its frequency with a minimum, average and maximum. A constant frequency is denoted as $f_i = f_i^{min} = f_i^{avg} = f_i^{max}$ with $1 \leq i \leq n, n \in \mathbb{N}$. A variable frequency it is denoted as $f_i^{min} < f_i^{avg} < f_i^{max}; 1 \leq i \leq n, n \in \mathbb{N}$. For convenience and readability, the frequencies of nodes with a particular role are denoted as f_{S_i} and $f_{S_i}^{min,avg,max}$ or f_{C_j} and $f_{C_j}^{min,avg,max}$ with $1 \leq i, j \leq n, n \in \mathbb{N}$.

3.1 Model

The model consist of a data flow graph D and the minimal N, average G and maximal X frequencies of all involved network nodes. It allows to determine a communication graph C, representing the nodes in the network which actively communicate. Data flow and communication graph combined make a point, on which nodes pull or push data in the network to establish the data flow. A decision function, described in Section 3.2 determines the communication graph C based on D, N, G and X.

The data flow is represented by a directed unweighted graph and is encoded in the adjacency matrix D. Nodes playing the role of a service are indexed by m, of a client by n and the size $m \times n, m = n$ of the adjacency matrix is determined by the number of nodes in the network. Each connection $d_{m,n}$ between a service and a client is encoded by 1 in the data flow adjacency matrix, in direction from service to client. The diagonal entries are set to 0 to avoid loops from a node to itself in the graph. Each entry in the matrix specifies the direction, in which data is transferred

$$D_{m,n} = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & 0 & \cdots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \cdots & 0 \end{pmatrix}$$

$$d_{m,n} \in \{0,1\}, \quad m = n, \quad m, n \in \mathbb{N} \tag{1}$$

$$N_n = \begin{pmatrix} f_1^{min} \\ f_2^{min} \\ \vdots \\ f_n^{min} \end{pmatrix} G_n = \begin{pmatrix} f_1^{avg} \\ f_2^{avg} \\ \vdots \\ f_n^{avg} \end{pmatrix} X_n = \begin{pmatrix} f_1^{max} \\ f_2^{max} \\ \vdots \\ f_n^{max} \end{pmatrix}$$

$$f_n^{min,avg,max} \in \mathbb{R}, n \in \mathbb{N} \tag{2}$$

$$C_{m,n} = \begin{pmatrix} 0 & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & 0 & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & 0 \end{pmatrix}$$

$$c_{m,n} \in \mathbb{R}, \quad m = n, \quad m, n \in \mathbb{N} \tag{3}$$

between a particular service and client.

The minimum, average and maximum of the frequencies of all nodes in the network are encoded as vectors. They are indexed by the index of the respective node and the size of the vector is determined by the number of nodes.

The determined communication graph is encoded in a similar way as the data flow graph by a directed finite simple graph with weights, encoded in the adjacency matrix C. In contrast to the data flow adjacency matrix, the range of the communication adjacency matrix is real numbers. Each entry in the matrix specifies the frequency and direction, in which a particular service has to push, or a particular client has to pull data in order to establish the data flow encoded in the data flow adjacency matrix.

3.2 Decision Function

The entries of the communication adjacency matrix C are derived from D, N, G and X. Each entry $c_{i,j}$ of the matrix is determined by the decision function.

We distinguish four cases of the function, which differ in the constant or inconstant frequencies of services and clients. By default $c_{i,j} = 0$ is set. In the first case, both the service and the client have constant frequencies. In the second case, the service has an inconstant frequency and the client has a constant frequency. The border cases $f_{S_i}^{max} = f_{C_j}$ and $f_{S_i}^{min} = f_{C_j}$ are included in the cases $f_{S_i}^{max} \leq f_{C_j}$ and $f_{S_i}^{min} \geq f_{C_j}$ correspondingly. A minimal frequency $f_{S_i}^{min}$ of the service S_i that equals the constant frequency f_{C_j} of the client C_j means in average a higher frequency of the service compared to the client, which has at least the frequency of the client. The second border case is handled analogically. In the third case, the service has a constant frequency and the client has an inconstant frequency. Finally, in the fourth case both service and client have inconstant frequencies.

In summary, the decision function determines, which node, given two nodes that are exchanging data (i.e. a services and a client), has to be the active one. Combined with the communication model, the decision function supports the formal capturing of a network of nodes, based on data-driven communication, and prescribes the optimal set of pushing and pulling nodes, in order to minimise the transfer of redundant data.

$$
\begin{aligned}
&< D, N, G, X, C > \\
&D \in \{0,1\}^{n \times n}; \quad N, G, X \in \mathbb{R}^n; \quad C \in \mathbb{R}^{n \times n} \\
&i, j = 1, \ldots, n; \quad n \in \mathbb{N}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
c_{i,j} = &\begin{cases} f_{S_i} & \text{if } d_{i,j} = 1 \text{ and } f_{S_i} \leq f_{C_j} \\ f_{C_i} & \text{if } d_{j,i} = 1 \text{ and } f_{S_j} > f_{C_i} \end{cases} \\
&\text{if } f_{S_i}^{min} = f_{S_i}^{max} \text{ and } f_{C_j}^{min} = f_{C_j}^{max} \\
&\text{with } f_{S_i} = f_{S_i}^{min} = f_{S_i}^{avg} = f_{S_i}^{max} \\
&\text{and } f_{C_j} = f_{C_j}^{min} = f_{C_j}^{avg} = f_{C_j}^{max}
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
c_{i,j} = &\begin{cases} f_{S_i}^{avg} & \text{if } d_{i,j} = 1 \text{ and } f_{S_i}^{max} \leq f_{C_j} \\ f_{S_i}^{avg} & \text{if } d_{i,j} = 1 \text{ and } f_{S_i}^{min} < f_{S_i}^{avg} \leq f_{C_j} < f_{S_i}^{max} \\ f_{C_i} & \text{if } d_{j,i} = 1 \text{ and } f_{S_j}^{min} < f_{C_i} < f_{S_j}^{avg} < f_{S_j}^{max} \\ f_{C_i} & \text{if } d_{j,i} = 1 \text{ and } f_{S_j}^{min} \geq f_{C_i} \end{cases} \\
&\text{if } f_{S_i}^{min} < f_{S_i}^{avg} < f_{S_i}^{max} \text{ and } f_{C_j}^{min} = f_{C_j}^{max} \\
&\text{with } f_{C_j} = f_{C_j}^{min} = f_{C_j}^{avg} = f_{C_j}^{max}
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
c_{i,j} = &\begin{cases} f_{S_i} & \text{if } d_{i,j} = 1 \text{ and } f_{S_i} \leq f_{C_j}^{min} \\ f_{S_i} & \text{if } d_{i,j} = 1 \text{ and } f_{C_j}^{min} < f_{S_i} \leq f_{C_j}^{avg} < f_{C_j}^{max} \\ f_{C_i}^{avg} & \text{if } d_{j,i} = 1 \text{ and } f_{C_i}^{min} < f_{C_i}^{avg} < f_{S_j} < f_{C_i}^{max} \\ f_{C_i}^{avg} & \text{if } d_{j,i} = 1 \text{ and } f_{S_j} \geq f_{C_i}^{max} \end{cases} \\
&\text{if } f_{S_i}^{min} = f_{S_i}^{max} \text{ and } f_{C_j}^{min} < f_{C_j}^{avg} < f_{C_j}^{max} \\
&\text{and } f_{C_j} = f_{C_j}^{min} = f_{C_j}^{avg} = f_{C_j}^{max}
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
c_{i,j} = &\begin{cases} f_{S_i}^{avg} & \text{if } d_{i,j} = 1 \text{ and } f_{S_i}^{max} \leq f_{C_j}^{min} \\ f_{S_i}^{avg} & \text{if } d_{i,j} = 1 \text{ and } f_{S_i}^{avg} \leq f_{C_j}^{avg} \\ & \quad \text{and } f_{S_i}^{max} > f_{C_j}^{min} \wedge f_{S_i}^{min} < f_{C_j}^{max} \\ f_{C_i}^{avg} & \text{if } d_{j,i} = 1 \text{ and } f_{S_j}^{avg} > f_{C_i}^{avg} \\ & \quad \text{and } f_{S_j}^{max} > f_{C_i}^{min} \wedge f_{S_j}^{min} < f_{C_i}^{max} \\ f_{C_i}^{avg} & \text{if } d_{j,i} = 1 \text{ and } f_{S_j}^{min} \geq f_{C_i}^{max} \end{cases} \\
&\text{if } f_{S_i}^{min} < f_{S_i}^{avg} < f_{S_i}^{max} \\
&\text{and } f_{C_j}^{min} < f_{C_j}^{avg} < f_{C_j}^{max}
\end{aligned} \tag{8}
$$

4 Evaluation

In this section we provide two sets of preliminary evaluation results. First, we apply the communication model and the decision function on actual nodes, with specific frequencies, and taking into consideration available connections. Second, we check the conformity of the model to the derived requirements. An in-depth experimental evaluation is part of future work.

We apply the model on the motivation scenario s (see Section 2) and, subsequently, calculate the communication matrix. We construct the data flow matrix D_s of scenario s based on the description of the scenario and as shown by the arrows for data flow in Figure 2 (solid arrows). Each non-zero entry in the matrix represents one data flow between nodes in the scenario (e.g., from node 1 as service (S_1) to node 3 as client (C_3)).

The frequencies of participating nodes in the network are given by the specification of the devices and applications, also shown in Figure 2. We construct all three vectors for minimal N_s, average G_s and maximal X_s frequencies, apply the decision function on the given arguments and derive the communication matrix C_s.

The derived set of active nodes within the network are visualised in Figure 2 as dashed arrows. For example, for nodes 1 and 4, 1 is the service and 4 is the client (solid arrow), however, based on the update frequencies, 4 should be the active node and request or pull data from 1.

We also evaluate the communication model and the decision function based on its conformity to the requirements, which were derived in Section 2. The decision function is based on optimising (minimising) the redundancy in terms of transferred data, thus minimising the volume of exchanged data. It uses the update frequencies in order to determine, which node should be the active in terms of producing or consuming the data, thus reducing the data, which was not updated for pull or cannot be processed for push (Requirement 1). Furthermore, the model and the decision function do not require that the function of a node is reassigned (services remain services and clients remain clients) or that the data flow direction is changed (Requirement 2), they simply determine, which nodes should be active (Requirement 3), based on taking the update frequencies into consideration (Requirement 4).

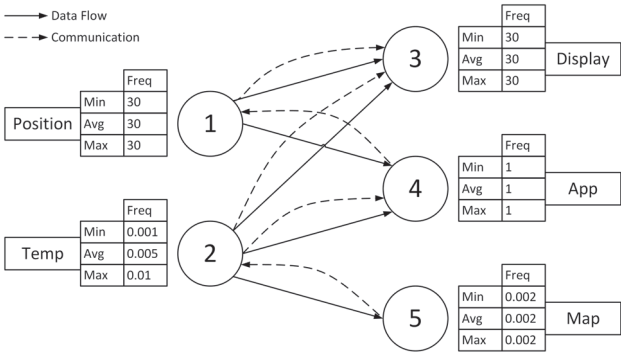


Figure 2: House Monitoring System Scenario

$$D_s = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{array} \begin{array}{c} \begin{array}{ccccc} N & C_1 & C_2 & C_3 & C_4 & C_5 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{9}$$

$$N_s = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cc} \text{N} & \text{Hz} \end{array} \\ \begin{pmatrix} 30 \\ 0.001 \\ 30 \\ 1 \\ 0.002 \end{pmatrix} \end{array} \quad G_s = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cc} \text{N} & \text{Hz} \end{array} \\ \begin{pmatrix} 30 \\ 0.005 \\ 30 \\ 1 \\ 0.002 \end{pmatrix} \end{array} \quad X_s = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cc} \text{N} & \text{Hz} \end{array} \\ \begin{pmatrix} 30 \\ 0.01 \\ 30 \\ 1 \\ 0.002 \end{pmatrix} \end{array} \tag{10}$$

$$C_s = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} N & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 0 & 30 & 0 & 0 \\ 0 & 0 & 0.005 & 0.005 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0.002 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{11}$$

5 Related Work

Related work has already been conducted in different domains, including economics and business administration, or, taken as excerpt, the field of sensors and sensor networks. The authors in (Cheng, Perillo, & Heinzelman, 2008) discuss different deployment strategies to avoid the decrease of sensor network lifetimes caused by high energy consumption of specific important network nodes. They propose a general model to compare these strategies under certain restrictions as well as calculate the costs caused by the deployment. For the scenario of a sensor network deployed over an area for surveillance, (Mhatre, Rosenberg, Kofman, Mazumdar, & Shroff, 2005) propose a cost model approach to determine an optimal distribution of sensing nodes and cluster head nodes. For large wireless ad-hoc networks (Liu, Huang, & Zhang, 2004) propose a combination of pull- and pushed-based strategies to optimise the routing for specific information needs. Thereby, the query frequencies are take into account. Compared to our work these approaches focus more on including factors specific to the deployment of sensors, e.g. power consumption, wireless strength or equipment costs, while we focus more on the optimisation of general issues in a network of data producing and consuming nodes.

6 Conclusion

Current developments of the Web are marked by the growing adoption and use of Web APIs. This trend, in combination with the wider use of sensors and mobile devices, raises new unaddressed challenges. In particular, there is the need for a general solution that enables the modelling of service-based systems, which is also capable of handling static as well as dynamic data exposed via Web APIs. In this paper we have presented a modelling approach that captures Web APIs and client applications, as data providers and consumers, which are characterised by their update frequencies. We use the formal model as the basis for applying a decision function for automatically determining, which communicating party needs to be active, by sending or requesting the data, in order to optimise the communication in terms of minimal data redundancy. The presented work lays the foundation for creating intelligent Web API-based systems and client application, which can automatically optimise the data exchange. As part of future work, we aim to integrate further influential factors. Currently, the proposed model allows only one type of frequency per network node. However, a node can actually serve as a service and client at the same time and would have an input and an output frequency, which may be equal or, e.g. in an aggregating node, differ. Similarly, latency and bandwidth are not considered in the current model. Latency would influence both the communication model and, especially, the decision function, which would have to take into account not only frequencies but also the possible latencies. The addition of bandwidth limitations would help to optimise data-driven communication. On the one hand, some factors could exclude others (i.e. bandwidth limitations overrule update frequencies). On the other hand, in combination with latency, the model could show the trade-off between lower latency and higher bandwidth.

References

Cheng, Z., Perillo, M., & Heinzelman, W. (2008). General Network Lifetime and Cost Models for Evaluating Sensor Network Deployment Strategies. *IEEE Transactions on Mobile Computing*, *7*(4), 484-497.

Liu, X., Huang, Q., & Zhang, Y. (2004). Combs, Needles, Haystacks: Balancing Push and Pull for Discovery in Large-Scale Sensor Networks. In *Proceedings of the Conference on Embedded Networked Sensor Systems.*

Mhatre, V., Rosenberg, C., Kofman, D., Mazumdar, R., & Shroff, N. (2005). A Minimum Cost Heterogeneous Sensor Network with a Lifetime Constraint. *IEEE Transactions on Mobile Computing*, *4*(1), 4-15.

Richardson, L., & Ruby, S. (2007). *RESTful Web Services*. O'Reilly.

FBWatch: Extracting, Analyzing and Visualizing Public Facebook Profiles

Lukas Brückner, lukas@lukas-brueckner.de, Kyto GmbH

Simon Caton, Simon.Caton@ncirl.ie, National College of Ireland

Margeret Hall, hall@kit.edu, Karlsruhe Service Research Institute

An ever-increasing volume of social media data facilitates studies into behavior patterns, consumption habits, and B2B exchanges, so called Big Data. Whilst many tools exist for platforms such as Twitter, there is a noticeable absence of tools for Facebook-based studies that are both scalable and accessible to social scientists. In this paper, we present FBWatch, an open source web application providing the core functionality to fetch public Facebook profiles en masse in their entirety and analyse relationships between profiles both online and offline. We argue that FBWatch is a robust interface for social researchers and business analysts to identify analyze and visualize relationships, discourse and interactions between public Facebook entities and their audiences.

1 Big Data Challenges in the Social Sciences

The vision of a Social Observatory is a low latency method for the observation and measurement of social indicators. It is a computer-mediated research method at the intersection of computer science and the social sciences. The term Social Observatory is used in its original context (Lasswell 1967; Hackenberg 1970); the framework is the archetypal formalization of interdisciplinary approaches in computational social science. The essence of a Social Observatory is characterized by (Lasswell 1967) as follows:

"The computer revolution has suddenly removed age-old limitations on the processing of information [...] But the social sciences are data starved [...] One reason for it is reluctance to commit funds to long-term projects; another [...] is the hope for achieving quick success by 'new theoretical breakthroughs' [...] It is as though we were astronomers who were supposed to draw celestial designs and to neglect our telescopes. The social sciences have been denied social observatories and told to get on with dreams"

This is also in line with the approach of the American National Science Foundation's call for a network of Social Observatories:

"Needed is a new national framework, or platform, for social, behavioral and economic research that is both scalable and flexible; that permits new questions to be addressed; that allows for rapid response and adaptation to local shocks […]; and that facilitates understanding local manifestations of national phenomena such as economic volatility."

Today, the notion of a Social Observatory lends itself towards social media platforms, as digital mediators of social exchange, discourse and representation. This, as demonstrated by the COSMOS project (Burnap et al. 2014), becomes especially valuable when combined with government data streams. However, empowering social scientists to access data from social media platforms (even in the singular) is non-trivial.

Figure 1 illustrates a general architecture of a modern Social Observatory entailing three processes; namely 1) Data Acquisition; 2) Data Analysis; and 3) Interpretation. Whilst it is apparent that a Social Observatory captures multiple sources of data, currently few scientific papers or services report this ability in a way easily replicable by social scientists (Cioffi-Revilla 2014). This is despite prevalent availability of Application Programming Interfaces (APIs), and an almost endless supply of papers and studies that focus on specific platforms (Russell 2013).

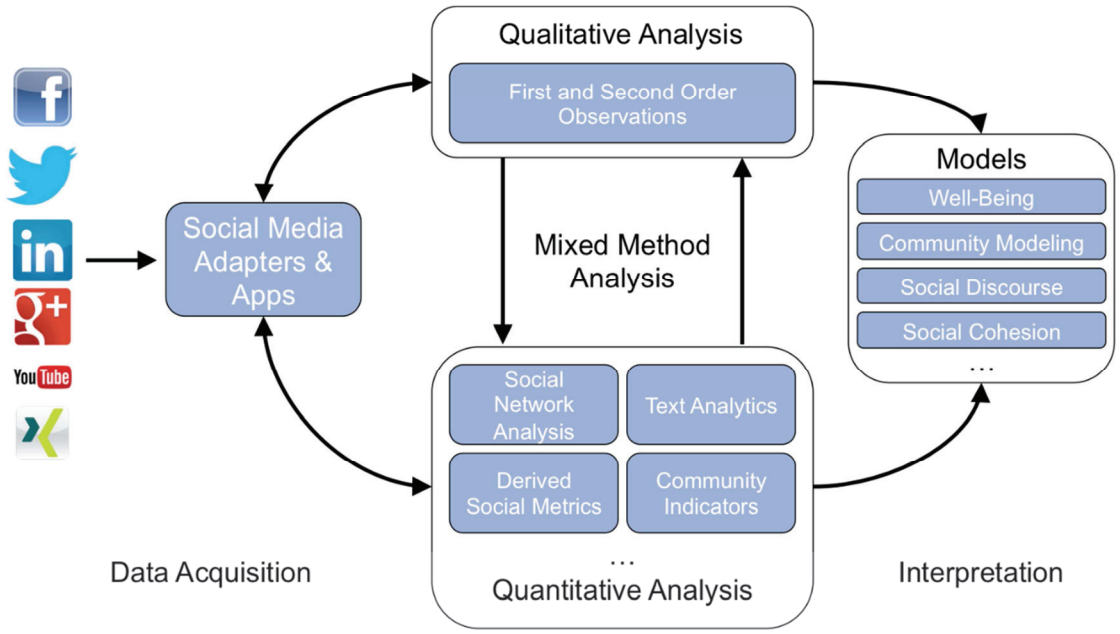


Figure 1. A General Architecture for a Social Observatory

Data Acquisition is well supported by most social media platforms via REST or streaming APIs, which are underpinned by lightweight data interchange formats like JSON. User authentication and access authorization is handled by technologies such as OAuth. There are also an ever-increasing number of software libraries available, reducing the implementation effort to extract data.

The challenges instead lie in data *volume, velocity, and variety*, access rights, and cross-platform differences in curating data. The big data aspects of social media data are well known: producing 2,200 Tweets (at around 58kb each) per second, Twitter is a clear demonstrator of data volume and velocity. Variety is best shown using a Facebook post as an example: version 1 of Facebook's Graph API contained at least 15 categories for a user post and this discounts other social actions like tagging, commenting, poking etc., as well as the diverse content range of a Facebook user's profile. Lastly, the method of data curation is not without its ambivalence. Twitter data

curation tends to be proactive; by accessing future Tweets that fulfil a specific set of user-driven attributes (e.g., hashtags or geolocation). Facebook is retrospective; given a Facebook entity (e.g. a person, or page) access their posts, profile, likes etc. From the perspective of analyzing social data, this subtle difference significantly alters the effort and planning needed to curate a data set (González-Bailón, Wang, Rivero, & Borge-Holthoefer, 2014). The technical challenges also differ significantly from receiving a continuous stream of data (i.e., tweets) vs. Facebook's paginated results. The latter incites large numbers of API calls, which are not limitless. On a side note, the validity period of an access token is also not infinite and must be refreshed periodically.

(**Mixed Method**) **Analysis** as illustrated in Figure 1, is inherently iterative and interdisciplinary. Foreseeable is repeated interaction with the social media adapters and apps. Whilst approaches from computer science and computational social science are becoming more prevalent, the question of research methodology is often a poignant discussion point and challenge that cannot be overlooked. Computer scientists and social scientists speak very different languages. Therefore, the realization of a Social Observatory needs to accommodate a vast array of (interdisciplinary) methodological approaches.

Irrespective of methodology, an important feature of a Social Observatory is the ability to view a community at a variety of resolutions; starting from an individual micro layer, and progressively zooming out via ego-centric networks, social groups, communities, and demographic (sub) groups, up to the macro layer: community. This ability is of significant importance for understanding a community as a whole; different granularities present differentiated views of the setting. Interpretation is hence domain specific in nature, and should be decided according to the proposed research questions. The architecture supports both inductive and deductive research.

Necessary to address at this point are the ethical boundaries of an unobtrusive approach of Big Data analyses of social data. Both Twitter and Facebook have terms and conditions allowing for the anonymized assessment of data which the use has indicated to be public. Specifically Facebook has argued that this is tantamount to informed consent, and this is a common position across social media platforms. This study agrees that when information is placed in public fora and domains, it is subject to public review. This is in line with the ethical guidelines of (Markham & Buchanan, 2012). In the case of obtrusive design (i.e., greedy apps), informed consent must continue to be in place as the standards of human subject research demand. A further ethical (and security) concern is that the provide architecture can also be used irresponsibly. In the case of public-facing data, this is of a lesser concern. Obtrusively-designed architectures still require user consent (e.g., downloading an app), as such research works are neither the work of hacking nor 'Trojan horses,' thus guaranteeing a moderately informed subject base.

1.1 Implementation: a Facebook Social Observatory Adapter

The first step towards a Social Observatory focuses on a Facebook social adapter for several reasons. Firstly, Facebook lends itself to the case study, especially due to the large number of "open" Facebook entities; where Facebook pages are a prime example. Secondly, when extracting data from Facebook, the researcher receives

near complete datasets. Finally, there is lack of general-purpose Facebook data acquisition tools available. Those that are available tend to rely either on crawling techniques, which cannot fully acquire paginated Facebook data, or data extraction via the Graph API that typically focus on the logged-in user or do not return data in full. Whilst such approaches are useful, especially in classroom settings, they do not provide mechanisms to curate research worthy datasets. This chapter presents a general and extensible Facebook data acquisition and analysis tool: FBWatch.

The objective is simple: an interface-based tool allowing social as well as computational scientists to access complete Facebook profiles irrespective of programming ability or data size, as no such tool is available. In extracting data from Facebook, the researcher first needs to define what is accessed: an entity that has a unique Facebook identifier. FBWatch is implemented such that it can access any Facebook entity that is public, or for which it has received user permissions.

FBWatch is implemented using the Ruby on Rails framework, and consists of five top-level components and modules: 1) Sync is the module responsible for fetching data from Facebook. It executes Graph API calls, converts graph data to the internal data structures and stores it in the database.b2) Metrics are the analysis components of FBWatch and responsible for analyzing fetched data. They contain parameters used for case studies and data structures for storing results. A metric can therefore be any result of an analysis (see Section 4). 3) Tasks are an abstraction for running Sync and Metric jobs as background processes. 4) A relational database for storing Facebook resource data, and running more complex queries regarding connections between Facebook entities. Any SQL-Server can be used provided that it supports UTF-8 encoding, as this is needed for handling foreign languages. MySQL and PostgreSQL both proved adequate. 5) A web front-end as an access point and controller for FBWatch. Here the user can request the retrieval of new Facebook entities, refresh previously fetched entities, group entities together for comparative analysis, execute metric calculations, visualize metrics as well as the social network of individual or grouped entities, and download datasets for use in third party analysis tools (see Section 3).

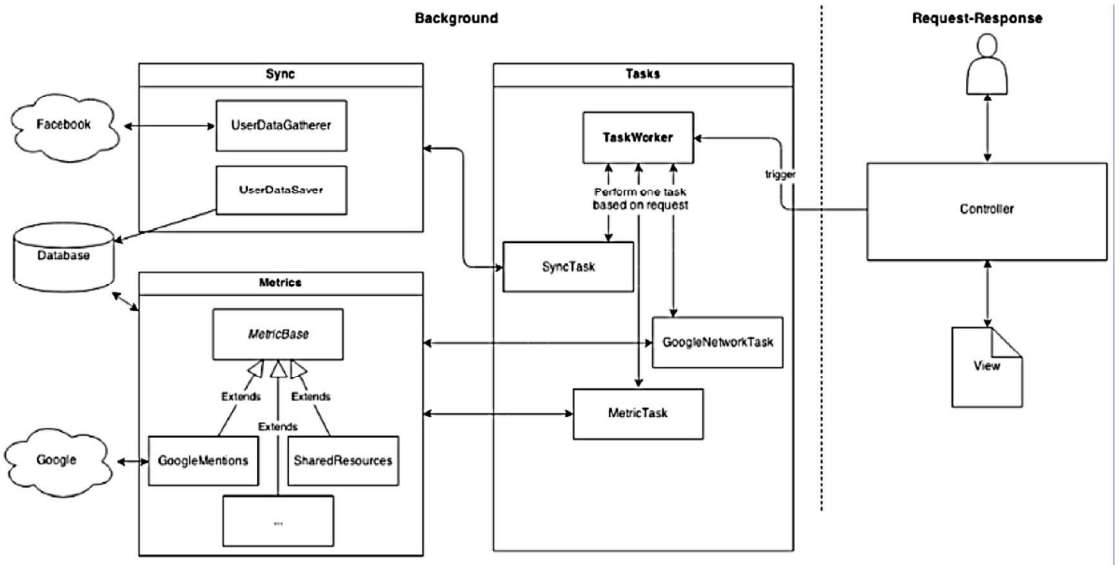


Figure 2. Workflow illustrating the steps to acquire, analyses, and interpret Facebook

Figure 2 shows the architecture of FBWatch, and highlights a typical request involving either the data fetching, or the metrics calculation. Upon a request, the controller triggers a background worker class and returns an appropriate view to the user who is notified that a task was started. The worker then performs one of two tasks, depending on whether Facebook data is to be retrieved, or retrieved data is to be analyzed.

The first step in the process flow the user providing the Facebook URL of one or more entities of interest, which are parsed for their username or Facebook ID. To synchronize the data of Facebook resources, a background sync task is started by FBWatch. The user can check the status and progress of the task, as required. Depending on the size and number of entities, synchronization can take several hours, and can also encounter several errors that need to be handled manually. Once synchronization has successfully completed, this will be visible and the user informed of how many feed entries have been retrieved. If errors were encountered that could not be handled this will also be displayed.

To access data, Koala, a lightweight and flexible Ruby library for Facebook, is used. It provides a simple user interface to the Graph API and the Facebook Query Language. As the Graph API returns the data in JSON format, Koala automatically parses the resulting string and converts it into the appropriate data structure using Arrays and Hashes and aligns the primitive data types into Ruby's data types. Furthermore, the library supports the use of the OAuth protocol to authenticate within Facebook through the use of the OmniAuth Ruby library. A valid, i.e. Facebook authenticated, instance of Koala is generated on a per-session basis and stored in the session context. At this time this is also the only real authentication the application performs directly. To mitigate exposing all data fetched by FBWatch, HTTP authentication is enforced on the server.

Synchronizing a Facebook resource is done in a two-step process. First, any basic information of that resource is pulled by calling the Graph API link facebook-id. Basic information contains the information visible at the top

of a Facebook page and in the about section, like first and last names, website, the number of likes etc. Second, the actual feed data is retrieved.

This is not trivial. First of all, not all data will and can be received at once, as Facebook limits the number of results per query; 25 per default. Increasing this limit drastically reduces the number of Graph API calls, and thus, speeds up the data gathering process. By default FBWatch uses a limit of 900, increasing speed and managing scalability. Facebook also only returns a subset of the comments and likes of a feed item; four by default. The resulting data contains a paging feature, similar to the one of the feed itself in a single feed item. Comments as well as like arrays have to be fetched using multiple API calls, dramatically increasing runtime. The UserDataGatherer module automatically navigates the paging system until it receives an empty data array. FBWatch also stores the link representing the first response from Facebook. This allows FBWatch to easily update a resource at some point in the future. If, however, a problem occurs, the last feed query is stored to enable the future continuation of a sync task.

The second part of the Sync module stores fetched data via the UserDataSaver. Aside from transforming Facebook JSON into internal data models, data entry needs to be optimized such that it scales. In order to decrease runtime, multiple INSERT and UPDATE statements are grouped into transactions. However, not all statements can be executed in one transaction due to interdependencies between data models. Thus, saving the data in the correct order is important. In order to take into account all possible dependencies, four transactions are used: 1) resources and their basic data are updated as well as all new Facebook entities that posted or interacted on the feed at the root level. 2) Feed entries. 3) Resources which interacted at a lower level, i.e. with a comment, like or tag. 4) The comments, likes and tags.

Once an entity has been fetched, it can at any time be resynchronized to retrieve any new feed items and their properties or continue to fetch all historic data if the synchronization was not successfully completed before. If a resource is no longer available on Facebook or no longer relevant for the analysis it also can be disabled or removed. Apart from the ability to traverse Facebook data automatically using the provided paging mechanism, the other main feature of the UserDataGatherer is error handling. The Facebook API is not reliable all the time, and is badly documented. Therefore, flexible error handling is required. The most pertinent hurdle is a limit to the amount of calls a single application can execute for a given access token in a certain time frame from the same IP address. While it is not officially documented, as per Facebook, apps tend to be limited to 600 calls every 10 minutes. For large resources, this limit is hit multiple times. FBWatch handles this by pausing the sync task, and retrying periodically (every five minutes) to resume it. This can require up to 30 minutes. FBWatch also handles when a resource cannot be queried, be it that it was deleted or disabled, when a username has been changed, and other miscellaneous errors.

1.2 Data Model

The data models representing social network data is loosely based on the Facebook Graph API format. A resource model corresponds to one Facebook entity but also constitutes the most important object in FBWatch. All overlapping properties of the different types of Facebook resources are saved in this data model: the free text name, the unique Facebook ID, the unique username and the full link to the resource on the Facebook system. Additional data relevant for the application is saved in this data model as well: a flag indicating whether or not a resource is active, i.e. if it should be synchronized, and the date of the last synchronization.

Other information returned by Facebook differs greatly for different entity types and is thus stored as an array of key-value pairs. Here, information such as the number of likes for pages, a website URL or the first and last names of real users, their gender and email address is represented. Furthermore, configuration data of the application is stored: information of the last synchronization so that it can be resumed more easily and no duplicates are retrieved. The value of stores the URL of the first link of the paging feature of the first feed page, i.e. where at the moment of synchronization newer data would be available. A property is called ‘last link’ stores the link to the last feed page unsuccessfully queried if an error occurred.

The core data structure is the feed (or timeline); a set of feed items. A feed item is modeled such that any type of textual activity can be represented, i.e. posts, comments and stories. Obviously, stories play an important role in user feeds. Note, however, that stories often appear right next to the actual activity, especially for comments; therefore, the content will be duplicated without care. So as to not lose too much information when handling different types of feed entries, a few additional properties are needed to the standard Facebook set. In order to simplify the data model differences in the available post types are mostly ignored. Post types are links, photos, statuses, comments, videos, swfs (flash objects) and check-ins as well as the corresponding stories. After analyzing the properties of these entries, the following attributes were selected: the unique facebook ID, timestamps representing when the entry was created and when it was last updated, the originator of the entry, optionally also the receiver of the entry and the comment and like count if present.

The originator and receiver are represented as separate resources, hence, only their unique IDs are stored here. The count of comments and likes are taken from the comments and likes properties of the Facebook format if present. A normal post has an attribute message which holds the text the user posted. A story, however, does not have a message, but rather a story property. The different sub-types of a post additionally have attributes containing the link, photo URL, etc. Each of these properties are mapped onto a single property. In order to distinguish between different types of feed items this property can be any of message, story or comment. The attribute then holds either story or comment for these two data types and the concrete post type for messages. A foreign key to the resource which this feed item belongs to, i.e. on which timeline it is posted. Last, to link comments to their respective post, a parent property is included, which is null for top-level posts.

1.3 Summary

The developed artifact demonstrated a first prototype of forming a general service that is capable of facilitating Big Data analyses based on Facebook data. The resulting software was designed to be modular enough to be extended in many different possible ways in order to support a multitude of research questions. As an endeavor like this is a large project only a first foundation was implemented. Nevertheless, as a first exploratory work in that direction the feasibility of a larger service was demonstrated. The aim of targeting software towards non-computer scientists is met for the main workflow. For this main workflow the other usability requirement of response times of less than ten seconds is met. Clearing data or loading the deep details of a resource can take more than ten seconds. For future applications to be performed on a different set of resources, the application provides a simple workflow without the need to adjust any source code. Modifying the scoring or adding new metrics requires programming knowledge, but is feasible.

In order to facilitate different analyses, the metric system was modularly defined. By providing a general base class where all specific metric classes can register themselves, it can be easily extended. Should external systems be required to perform additional analyses, the fetched data can be exported into a JSON format and put to other software. The structure of the JSON format was designed to be close to the one Facebook provides itself. Since not all returned data is saved and some parts are stored differently, the JSON feed of Facebook and FBWatch are not a one-to-one match. Only small differences exist, though, and any Facebook format parser should be adapted easily to the artifact's format. In general, the data input is extensible.

In summary, it can be said that the contribution of this research is twofold. First, it provides an exploratory social network observatory. Essential information and challenges were discovered and a robust error handling introduced. Second, a comprehensive solution for retrieving a new market perspective from the customer point of view was presented focusing on Facebook data. Additionally, the information contained within should provide guidelines and a solid base for conducting further social network research and for creating further social observatories. With internet services and online social network services developing at a rapid pace and more and more services being created the possibilities of facilitating the data which they collect stays an interesting topic of research. It remains to be seen whether or not more services will open up their platforms and provide access to at least some part of their data warehouses giving academic researchers and in particular social scientists new ways of studying people's behavior and get a new perspective on markets.

2 References

Burnap, P. et al., 2014. COSMOS : Towards an integrated and scalable service for analysing social media on demand. *International Journal of Parallel, emergent and Distributed Computing*, pp.37–41.

Cioffi-Revilla, C., 2014. *Introduction to Computational Social Science*, Berlin: Springer Texts in Computer Science.

Hackenberg, R., 1970. The Social Observatory : Time series data for health and behavioral research. *Social Science and Medicine*, 4, pp.343–357.

Lasswell, H.D., 1967. Do We Need Social Observatories? *The Saturday Review*, pp.49–52.

Stylometry-based Fraud and Plagiarism Detection for Learning at Scale

Markus Krause, markus@hci.uni-hannover.de, Leibniz University

Fraud detection in free and natural text submissions is a major challenge for educators in general. It is even more challenging to detect plagiarism at scale and in online classes such as Massive Open Online Courses. In this paper, we introduce a novel method that analyses the writing style of an author (stylometry) to identify plagiarism. We will show that our system scales to thousands of submissions and students. For a given test set of ~4000 users our algorithm shows F-scores of over 90%.

1 Introduction

Fraud, cheat, and, plagiarism detection are major challenges for learning at scale. Verifying that a student solved an assignment alone is extremely hard to verify in an online setting. Even in offline courses with a couple of hundred students, detecting plagiarism or cheating is nearly impossible for a teacher. Various attempts have been proposed to solve this issue (Meuschke, 2013). Intrinsic plagiarism detection is a promising method for large-scale online courses. Intrinsic plagiarism detection uses stylometry (analysis of literary style) to identify stylistically different segments in a text. Features used for stylometry are sentence length, vocabulary rich-ness, frequencies of words, or word lengths. With carefully chosen features, stylometry is robust even against automated imitation attempts as automatically altering the grammatical structure of a sentence without changing the meaning of the text is challenging (Brennan, Afroz, & Greenstadt, 2012; Krause, 2014). Narayanan and Paskov also demonstrated that stylometry is scalable. They reliably identify authors in a large corpus with texts from 100,000 individuals (Narayanan & Paskov, 2012). Monaco et al. (Monaco, Stewart, Cha, & Tappert, 2013) gives an overview of different identification methods using stylometry.

In this paper, we propose a new method using stylometry for plagiarism detection. We illustrate the feasibility of our method and demonstrate that our approach scales to thousands of authors. Our approach uses a relatively small feature set (164 features) compared to methods such as *Writeprints* (>30,000 features) (Abbasi & Chen, 2008). In contrast to other approaches, we use an input resampling and smoothing method on the training, testing, and evaluation data and train a classifier for each author. We also use a standard SVM in contrast to most other approaches that use a single class SVM (Abbasi & Chen, 2008). With this method, the SVM can not only learn positive instances but also negative. We hypothesis that:

H1: Negative instances in the training set increase the quality of fraud prediction.

To illustrate the performance of our method we predict if a text was written by a given author or not comparing the writing style of the questionable text to a known sample. We report two main units F-score and Cohen's Kappa. The F-score is defined as the harmonic mean of precision and recall. Cohen's Kappa is a measure for inter rater agreement (Cohen, 1960) and also useful in estimating classifier performance. Measuring the disagreement between the algorithms predictions and the expected classes.

We test our method on a corpus build from 17,000 blogs. The corpus was initially composed and published by Schler et al. (Schler, Koppel, Argamon, & Pennebaker, 2005). We selected all blogs with at least 35,000 characters (approx. 3100 words per author).

2 Feature Extraction

For our experiments, we extracted a set of features from each blog in our corpus. Many approaches use features that an algorithm can easily alter, for instance digits. An algorithm can easily detect fractional numbers and add additional numbers to better resemble another author e.g. altering 0.98 to 0.982. This does not change the meaning of the text and a reviewer would not be able to recognize such changes. Similar approaches also work for whitespaces such as line breaks, tabs, and space. Besides omitting certain features, we also expanded others. An often-used feature is average word length. Instead of the average length, we use word length frequencies. Furthermore, we added new features not yet explored. We describe individual feature sets below. They sum up to 164 individual features.

Character Frequency (48 features)
The relative frequency of individual characters. This feature set contains the relative frequencies of a-z and A-Z.

Word Length Frequency (20 features)
The relative frequency of word length. In some rare cases the part of speech tagger was not able to filter certain artifacts e.g. long numbers, some e-mail addresses (without the @ sign). This results in particular long words. To filter such elements we only use words of up to 20 characters.

Sentence Length Frequency (35 features)
The relative frequency of sentence length. Similar to the word length feature we filter out overly long sentences longer than 35 words.

Part of Speech Tag Frequency (35 features)
For this feature set we use the Penn Treebank part of speech tag set. We use the Natural Language Toolkit (*NLTK* [2]) python library to extract these tags from a corpus. We calculate the relative frequency of each tag.

Word Specificity Frequency (20 features)

The specificity of words used by an author is a discriminating feature and a relevant predictor in other Natural Language Tasks (Kilian, Krause, Runge, & Smeddinck, 2012; Krause, 2013). However, to our knowledge this feature have not been used for stylometry yet. To estimate the specificity of a word we use *wordnet* (Miller, 1995). For each word, we predict the lemma of the word and its part of speech. With the lemma and the part of speech, we retrieve all relevant *synsets*. The algorithm calculates the distance between each *synset* and the root node of *wordnet*. We define specificity as the average depth of these *synsets* rounded to the nearest integer. The algorithm calculates the relative frequency of each depth. The depth is limited to 20 as higher values tend to be extremely rare.

3 Method

To represent the diversity of an authors' writing we resample the input data. We merge all documents of an author and split the resulting text into sentences. Each word in a sentence is annotated with its part of speech (POS). To generate POS we use the Penn-Treebank tagger from the *NLTK* library. From these annotated sentences, we calculate a feature vector for each sentence. So that each author has a corresponding set of vectors. We split each set into three equally sized subsets a *training set*, a *test set*, and an *evaluation set*.

Each set is further processed. From each set, we randomly select n sentence vectors to have equal numbers of vectors in each set of these sets for each author. We select b bootstrap samples from the *training set*. A bootstrap sample is drawn by randomly selecting a vector from a set and repeating this as many times as vectors in the set without removing the picked vector from the set. As a final step, we average over all picked vectors to create a single bootstrap sample.

To train the support vector machine for an author we take the b bootstrap samples of an author as positive examples. We then select i other authors randomly from the initial set of ~4000 authors. These authors are called impostors. We again randomly select $\frac{b}{\mathrm{i}}$ bootstrap samples from the training sets of these i authors for b negative examples. As we use the described bootstrap method to generate samples, we can use any number for b reasonably smaller than the total number of possible permutations of sentence vectors per author.

For the experiment we used $b=400$, $i=100$, and $n=100$. We generated 400 bootstrap samples for each author and 4 bootstrap samples for each of the impostors, therefore, training the SVM with 800 bootstrap samples. We determined this number with a series of tests. Larger values for b and i only marginally increased classification accuracy. Higher values for n increased classifier performance but seem to be unrealistic for course submissions. We repeated this process 1000 times for each author so we had 1000 classifiers for each author. For each trained classifier we randomly selected 100 authors from the initial ~4000. We again refer to these authors as impostors. These impostors are distinct from those we trained the classifier with! We generated 100 bootstrap samples for each impostor and 100 bootstrap samples for the author. All bootstrap samples are drawn from the sentence vectors selected for

evaluation earlier. We used each classifier to predict the class of each of the evaluation bootstrap samples and calculated F-score and Cohen's Kappa for each classifier.

We also use a traditional method on the same data set as a baseline as used for the *Writeprint* system (Abbasi & Chen, 2008). For this method, we use tenfold cross-validation by splitting author texts into 15 parts (5 for training, 5 for testing, 5 for evaluation). For example, in fold 1, parts 1–5 are used for training, while parts 6–10 are for testing; in fold 2, in parts 2–6 are used for training while parts 1 and 7–10, are for testing. From all possible permutations we randomly selected 1000 for each author. We trained a single class SVM with the 5 training samples. To calculate F-score and Kappa we used the evaluation samples of the author and the evaluation samples of a random set of 100 impostors. For both methods, we used the testing samples to estimate optimal parameters for the SVMs. This is a crucial step, as SVMs require parameter tuning to work effectively. To estimate optimal parameters we used grid search on a subsample of authors not used for testing or evaluating.

Results

We found our proposed method to be effective in detecting the authorship with a mean F-score of 0.91 and a 95% Confidence Interval of the mean of [0.89, 0.92]. The average Cohen's Kappa for our method is 0.8 with a 95% CI [0.77, 0.84]. These results are higher than the results from the traditional method with an F-score of 0.65 and a 95% CI [0.56, 0.74] and a Kappa score of 0.46 and a 95% CI of [0.36, 0.56]. Figure 1 illustrates the results of our experiment.

Figure 1: The binary SVM outperforms a traditional single class SVM. Error bars show the 95% Confidence Interval. We calculated the interval from 10,000 bootstrap samples.

Conclusion

In this paper, we presented a new method using stylometry to validate the authorship of student submissions. We illustrated that our method of input smoothing and resampling allows us to train standard binary SVMs. We also showed that the prediction quality of these binary SVMs trained with our method is higher than the quality of a comparable single class SVM trained with a traditional method. We illustrated that our method performs with high accuracy even for a large data set of more than 4000 authors and reliable predict if a given author wrote a text.

References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, *26*(2), 29.

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry. *ACM Transactions on Information and System Security*, *15*(3), 12:1–22.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37.

Kilian, N., Krause, M., Runge, N., & Smeddinck, J. (2012). Predicting Crowd-based Translation Quality with Language-independent Feature Vectors. In *HComp'12 Proceedings of the AAAI Workshop on Human Computation* (pp. 114–115). Toronto, ON, Canada: AAAI Press.

Krause, M. (2013). GameLab: A Tool Suit to Support Designers of Systems with Homo Ludens in the Loop. In *HComp'13 Proceedings of the AAAI Conference on Human Computation: Works in Progress and Demonstration Abstracts* (Vol. 1, pp. 38–39). Palm Springs, CA, USA: AAAI Press.

Krause, M. (2014). A behavioral biometrics based authentication method for MOOC's that is robust against imitation attempts. In *L@S'14 Proceedings of the first ACM conference on Learning@ scale conference (Work in Progress)* (pp. 201–202). Atlanta, GA, USA: ACM Press.

Meuschke, N. (2013). State - of - the - art in detecting academic plagiarism, *9*(1), 50–71.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Monaco, J., Stewart, J., Cha, S., & Tappert, C. (2013). Behavioral Biometric Verification of Student Identity in Online Course Assessment and Authentication of Authors in Literary Works. In *International Conference on Biometrics, ICB'2013* (p. 8). Madrid, Spain.

Narayanan, A., & Paskov, H. (2012). On the feasibility of internet-scale author identification. In *IEEE Symposium on Security and Privacy*. San Francisco, CA, USA: IEEE Press.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2005). Effects of Age and Gender on Blogging. In *Proceedings of the 19th AAAI Conference on Artificial Intelligence, AAAI-5*. AAAI Press.

ISSN 1869-9669
ISBN 978-3-7315-0344-6

ISBN 978-3-7315-0344-6