International Journal on Advances in Internet Technology

2018 vol. 11 nr. 1&2

The *International Journal on Advances in Internet Technology* is published by IARIA.
ISSN: 1942-2652
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Internet Technology
Volume 11, Number 1 & 2, 2018

Xi Chen, University of Washington, USA
IlKwon Cho, National Information Society Agency, South Korea
Andrzej Chydzinski, Silesian University of Technology, Poland
Noël Crespi, Telecom SudParis, France
Antonio Cuadra-Sanchez, Indra, Spain
Javier Cubo, University of Malaga, Spain
Sagarmay Deb, Central Queensland University, Australia
Javier Del Ser, Tecnalia Research & Innovation, Spain
Philipe Devienne, LIFL - Université Lille 1 - CNRS, France
Kamil Dimililer, Near East Universiy, Cyprus
Martin Dobler, Vorarlberg University of Applied Sciences, Austria
Jean-Michel Dricot, Université Libre de Bruxelles, Belgium
Matthias Ehmann, Universität Bayreuth, Germany
Tarek El-Bawab, Jackson State University, USA
Nashwa Mamdouh El-Bendary, Arab Academy for Science, Technology, and Maritime Transport, Egypt
Mohamed Dafir El Kettani, ENSIAS - Université Mohammed V-Souissi, Morocco
Armando Ferro, University of the Basque Country (UPV/EHU), Spain
Anders Fongen, Norwegian Defence Research Establishment, Norway
Giancarlo Fortino, University of Calabria, Italy
Kary Främling, Aalto University, Finland
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Ivan Ganchev, University of Limerick, Ireland / University of Plovdiv “Paisii Hilendarski”, Bulgaria
Shang Gao, Zhongnan University of Economics and Law, China
Kamini Garg, University of Applied Sciences Southern Switzerland, Lugano, Switzerland
Rosario Giuseppe Garroppo, Dipartimento Ingegneria dell'informazione - Università di Pisa, Italy
Thierry Gayraud, LAAS-CNRS / Université de Toulouse / Université Paul Sabatier, France
Christos K. Georgiadis, University of Macedonia, Greece
Katja Gilly, Universidad Miguel Hernandez, Spain
Mariusz Głąbowski, Poznan University of Technology, Poland
Feliz Gouveia, Universidade Fernando Pessoa - Porto, Portugal
Kannan Govindan, Crash Avoidance Metrics Partnership (CAMP), USA
Bill Grosky, University of Michigan-Dearborn, USA
Jason Gu, Singapore University of Technology and Design, Singapore
Christophe Guéret, Vrije Universiteit Amsterdam, Nederlands
Frederic Guidec, IRISA-UBS, Université de Bretagne-Sud, France
Bin Guo, Northwestern Polytechnical University, China
Gerhard Hancke, Royal Holloway / University of London, UK
Arthur Herzog, Technische Universität Darmstadt, Germany
Rattikorn Hewett, Whitacre College of Engineering, Texas Tech University, USA
Quang Hieu Vu, EBTIC, Khalifa University, Arab Emirates
Hiroaki Higaki, Tokyo Denki University, Japan
Dong Ho Cho, Korea Advanced Institute of Science and Technology (KAIST), Korea
Anna Hristoskova, Ghent University - IBBT, Belgium
Ching-Hsien (Robert) Hsu, Chung Hua University, Taiwan
Chi Hung, Tsinghua University, China
Edward Hung, Hong Kong Polytechnic University, Hong Kong
Raj Jain, Washington University in St. Louis , USA
Edward Jaser, Princess Sumaya University for Technology - Amman, Jordan
Terje Jensen, Telenor Group Industrial Development / Norwegian University of Science and Technology, Norway
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Georgios Kambourakis, University of the Aegean, Greece
Atsushi Kanai, Hosei University, Japan
Henrik Karstoft , Aarhus University, Denmark

Dimitrios Katsaros, University of Thessaly, Greece
Ayad ali Keshlaf, Newcastle University, UK
Reinhard Klemm, Avaya Labs Research, USA
Samad Kolahi, Unitec Institute Of Technology, New Zealand
Dmitry Korzun, Petrozavodsk State University, Russia / Aalto University, Finland
Slawomir Kuklinski, Warsaw University of Technology, Poland
Andrew Kusiak, The University of Iowa, USA
Mikel Larrea, University of the Basque Country UPV/EHU, Spain
Frédéric Le Mouël, University of Lyon, INSA Lyon / INRIA, France
Juong-Sik Lee, Nokia Research Center, USA
Wolfgang Leister, Norsk Regnesentral (Norwegian Computing Center), Norway
Clement Leung, Hong Kong Baptist University, Hong Kong
Longzhuang Li, Texas A&M University-Corpus Christi, USA
Yaohang Li, Old Dominion University, USA
Jong Chern Lim, University College Dublin, Ireland
Lu Liu, University of Derby, UK
Damon Shing-Min Liu, National Chung Cheng University, Taiwan
Michael D. Logothetis, University of Patras, Greece
Malamati Louta, University of Western Macedonia, Greece
Maode Ma, Nanyang Technological University, Singapore
Elsa María Macías López, University of Las Palmas de Gran Canaria, Spain
Olaf Maennel, Loughborough University, UK
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Yong Man, KAIST (Korea advanced Institute of Science and Technology), South Korea
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Chengying Mao, Jiangxi University of Finance and Economics, China
Brandeis H. Marshall, Purdue University, USA
Sergio Martín Gutiérrez, UNED-Spanish University for Distance Education, Spain
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Shawn McKee, University of Michigan, USA
Stephanie Meerkamm, Siemens AG in Erlangen, Germany
Kalogiannakis Michail, University of Crete, Greece
Peter Mikulecky, University of Hradec Kralove, Czech Republic
Moeiz Miraoui, Université du Québec/École de Technologie Supérieure - Montréal, Canada
Shahab Mokarizadeh, Royal Institute of Technology (KTH) - Stockholm, Sweden
Mario Montagud Climent, Polytechnic University of Valencia (UPV), Spain
Stefano Montanelli, Università degli Studi di Milano, Italy
Julius Müller, TU- Berlin, Germany
Juan Pedro Muñoz-Gea, Universidad Politécnica de Cartagena, Spain
Krishna Murthy, Global IT Solutions at Quintiles - Raleigh, USA
Alex Ng, University of Ballarat, Australia
Christopher Nguyen, Intel Corp, USA
Petros Nicopolitidis, Aristotle University of Thessaloniki, Greece
Carlo Nocentini, Università degli Studi di Firenze, Italy
Federica Paganelli, CNIT - Unit of Research at the University of Florence, Italy
Carlos E. Palau, Universidad Politecnica de Valencia, Spain
Matteo Palmonari, University of Milan-Bicocca, Italy
Ignazio Passero, University of Salerno, Italy
Serena Pastore, INAF - Astronomical Observatory of Padova, Italy
Fredrik Paulsson, Umeå University, Sweden
Rubem Pereira, Liverpool John Moores University, UK
Yulia Ponomarchuk, Far Eastern State Transport University, Russia
Jari Porras, Lappeenranta University of Technology, Finland

Neeli R. Prasad, Aalborg University, Denmark
Drogkaris Prokopios, University of the Aegean, Greece
Emanuel Puschita, Technical University of Cluj-Napoca, Romania
Lucia Rapanotti, The Open University, UK
Gianluca Reali, Università degli Studi di Perugia, Italy
Jelena Revzina, Transport and Telecommunication Institute, Latvia
Karim Mohammed Rezaul, Glyndwr University, UK
Leon Reznik, Rochester Institute of Technology, USA
Simon Pietro Romano, University of Napoli Federico II, Italy
Michele Ruta, Technical University of Bari, Italy
Jorge Sá Silva, University of Coimbra, Portugal
Sébastien Salva, University of Auvergne, France
Ahmad Tajuddin Samsudin, Telekom Malaysia Research & Development, Malaysia
Josemaria Malgosa Sanahuja, Polytechnic University of Cartagena, Spain
Luis Enrique Sánchez Crespo, Sicaman Nuevas Tecnologías / University of Castilla-La Mancha, Spain
Paul Sant, University of Bedfordshire, UK
Brahmananda Sapkota, University of Twente, The Netherlands
Alberto Schaeffer-Filho, Lancaster University, UK
Peter Schartner, Klagenfurt University, System Security Group, Austria
Rainer Schmidt, Aalen University, Germany
Thomas C. Schmidt, HAW Hamburg, Germany
Zary Segall, Chair Professor, Royal Institute of Technology, Sweden
Dimitrios Serpanos, University of Patras and ISI/RC ATHENA, Greece
Jawwad A. Shamsi, FAST-National University of Computer and Emerging Sciences, Karachi, Pakistan
Michael Sheng, The University of Adelaide, Australia
Kazuhiko Shibuya, The Institute of Statistical Mathematics, Japan
Roman Y. Shtykh, Rakuten, Inc., Japan
Patrick Siarry, Université Paris 12 (LiSSi), France
Jose-Luis Sierra-Rodriguez, Complutense University of Madrid, Spain
Simone Silvestri, Sapienza University of Rome, Italy
Vasco N. G. J. Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
Radosveta Sokullu, Ege University, Turkey
José Soler, Technical University of Denmark, Denmark
Victor J. Sosa-Sosa, CINVESTAV-Tamaulipas, Mexico
Dora Souliou, National Technical University of Athens, Greece
João Paulo Sousa, Instituto Politécnico de Bragança, Portugal
Kostas Stamos, Computer Technology Institute & Press "Diophantus" / Technological Educational Institute of Patras, Greece
Cristian Stanciu, University Politehnica of Bucharest, Romania
Vladimir Stantchev, SRH University Berlin, Germany
Tim Strayer, Raytheon BBN Technologies, USA
Masashi Sugano, School of Knowledge and Information Systems, Osaka Prefecture University, Japan
Tae-Eung Sung, Korea Institute of Science and Technology Information (KISTI), Korea
Sayed Gholam Hassan Tabatabaei, Isfahan University of Technology, Iran
Yutaka Takahashi, Kyoto University, Japan
Yoshiaki Taniguchi, Kindai University, Japan
Nazif Cihan Tas, Siemens Corporation, Corporate Research and Technology, USA
Alessandro Testa, University of Naples "Federico II" / Institute of High Performance Computing and Networking (ICAR) of National Research Council (CNR), Italy
Stephanie Teufel, University of Fribourg, Switzerland
Parimala Thulasiraman, University of Manitoba, Canada
Pierre Tiako, Langston University, USA

International Journal on Advances in Internet Technology
Volume 11, Numbers 1 & 2, 2018

CONTENTS

The BBC Broadcast of 'India's Daughter': An Examination of the Interactions between Mainstream Media and Social Media

Balbir S. Barn*, Ravinder Barn† and Franco Raimondi*
*Dept. of Computer Science
Middlesex University, London, UK
Email: b.barn, f.raimondi@mdx.ac.uk

†School of Law
Royal Holloway University of London
Egham, UK
Email: r.barn@rhul.ac.uk

***Abstract*—The pervasiveness of social media has resulted in increased public involvement in key discussions about social issues, as well as creating greater affordances for individual expression and collective mobilisation. In December 2012, the rape and murder of a 23-year-old Indian student in New Delhi, India, was followed by widespread condemnation and public action organised and coordinated through social media. In March 2015, a controversial BBC (British Broadcasting Corporation) documentary, "India's Daughter", about the incident was broadcast despite restrictions imposed by the Indian Government. This paper explores the interplay between Mainstream Media ("Fourth Estate") and Social Media ("Fifth Estate") through a case study analysis using computational techniques to analyse 250,000 tweets collated following the broadcast of the documentary. The primary contribution of the paper is a contextualisation of our findings and analysis within the theoretical frameworks of social movement theory and postcolonialism in order to understand the interactions between mainstream media and social media. Limitations and issues around implications for conducting inter-disciplinary social media research are also discussed.**

Keywords–Postcolonialism; India's Daughter; Social Network Analysis; Mainstream Media; Social Media; Protest Movement; Twitter

1. INTRODUCTION

In December 2012, the rape and murder of a 23-year-old Indian student in New Delhi, India, was followed by widespread condemnation and public action organised and coordinated through social media. In March 2015, a controversial BBC (British Broadcasting Corporation) documentary, "India's Daughter", about the incident was broadcast despite restrictions imposed by the Indian Government. Analysis of social media commenting on the broadcast was first reported at the Third International Conference on Human and Social Analytics [1]. As the most populous democracy in the modern world, India has witnessed an increasing growth in the use of the internet in general, and social media in particular. Although accurate statistics are difficult to obtain, estimates of the micro-blogging site of 140 characters, Twitter, range from 23 million to 35 million [2]. This figure has more than doubled in the last 3 years. According to a collective called 'India on the Internet 2014', Twitter users in India total 35 million while 125 million people are on the social networking site (SNS), Facebook. Further, it is estimated that almost 9 out of 10 web users in India visit a social networking site. In a climate of smart phones and their applications, and SNSs, such a figure is not so surprising. A core feature of such social media sites is their reliance on individual users for content creation and active user involvement. Consequently, such engagement on social media platforms results in the availability of what has been termed as big data. Academic researchers have been drawn to such sources of information to help promote understandings in a number of key areas including social movements and social protests, general and local political elections, trends in patterns of health and health behaviours, consumption of social media by individuals and groups, social influence in e-commerce, work stress, and national happiness [3].

Twitter is notable in that it has rapidly become important and popular as key tool for organising and generating communication for protestors around the world [4]. Examples of where Twitter has played a significant role include: the Iranian protests of 2009-2010 [5], the so-called Egyptian revolution of 2011 [6] and also the various Occupy protests that took place around the world [7]. It is also clear that that the messaging technology is viewed differently depending upon context. Hence it is seen as subversive by autocratic regimes as well as a suitable technology for surveillance [8]. In addition to the organising and communicating aspects of Twitter, researchers have also commented on how Twitter and social media in general is also being used to reconstruct and extend journalism and notions of what constitutes a Habermasian public sphere [9], [10]. That is, the realm of social life in which, something approaching public opinion can be formed and where access is guaranteed to all citizens. The perception is that networked individuals have the capacity to use social media to enhance their role in news production and dissemination to achieve a growing independence from the mainstream media ("Fourth Estate") [11].

In South Delhi, India, the 16 December 2012 rape and murder of a 23 year old physiotherapy student by six men, marked a Twitter watershed moment where some commentators asked whether this heralded an 'Indian Spring' [12]. The street protests across the nation in which social media was said to have played a part were described by some as the new 'unifying force' [13] through the formulation of a shared public opinion on social media.

Researchers have analysed social media outlets from that period [14] [15]. Much of the protest and mobilisation has

been framed using social movement theories such as that by McAdam et al. [16] and resource mobilization theory [17].

Several years after the Delhi rape incident, a BBC documentary titled *"India's Daughter"* was broadcast on 4 March 2015 in the UK and on 8th March 2015 in New York. The broadcast of the documentary directed by Leslee Udwin was controversial, in that, the Indian Government sought to have it banned and the BBC chose to bring forward the broadcast to an earlier programme slot. Early indicators of the controversial aspects of the documentary were immediately brought to the foreground. These included: the extent to which mainstream media occupied the so-called egalitarian and democratising space of social media; the postcolonial texture of the debate and the overall sense of how mainstream media handled the case under question. Twitter naturally formed a predominant backdrop to this broadcast given its initial role in mobilisation of public opinion in the original 2012 event (mentioned above) [12], [13], [18].

We contend that this 2015 debate on Twitter bears likeness to a protest movement, albeit tempered by the influence of mainstream media.

Within a context where the mainstream media and social activists now largely occupy this micro-blogging space [13], research that examines the interplay between mainstream media and social media through specific case study instances can have public policy implications. Hence, the research reported in this paper makes a key analytical contribution of public reaction, through tweets, to the broadcast of the documentary on the BBC in the UK and on Youtube, Vimeo and other sources in India. This interplay is explored through two perspectives. Two explorations are undertaken. Firstly, we reflect on the reaction to the documentary broadcast as another mode of social mobilisation. Secondly, we examine how postcolonialism and anti-postcolonialism rhetoric manifests itself.

This is an extended version of a paper published in the IARIA HUSO 2017 conference [1]. Here we present additional theory development of social mobilisation as well as further commentary and results.

The remainder of the paper is structured as follows: In Section 2, approaches to research methods for social media research as well as outlining key concepts and debates around social mobilisation and postcolonial theory. The latter is significant given that postcolonialism was an important emergent theme. Crucially, the historical relationship between the two countries where the incident occurred and the film was shown provided a particular salience and backdrop to the analysis of this paper. In Section 3, we present the aims and details of the research methodology we have used. Section 4 presents overview results arising from the blog analysis and computational analysis of the collected tweets. Section 5 entails a discussion of some tweets followed by commentary on the validity of the results. Finally, in Section 6, we present concluding remarks and outline some further research considerations.

2. Related work and underpinning theory

In this section we outline key concepts and debates around Twitter methodological concerns, social mobilisation and postcolonial theory given that postcolonialism was an important emergent theme.

A. Twitter based methods

Twitter evolved from a status, phatic oriented and "inconsequential chirping" tool, to an event-following reporting tool. More recently, it has evolved to its current form where it has settled into a data set from which researchers extract collections and one that is archived by the US Library of Congress [19]. Throughout this evolution, Twitter's data set has been source of study for social scientists where methodological approaches which have largely remained as black boxes whose outputs on analyses such as "reach" taken on trust. It is possible that Twitter, through its publicly accessible routes to tweet data, has moved computational social science from being the preserve of governments and private companies [20].

Methods for analysing tweet data sets rely on understanding the interplay between three layers of communication in Twitter. Bruns and Moe [21] identify these three layers as: Micro (inter-personal communications); Meso (follower-followee relationships) and and the macro level of hashtag-based exchanges. The Meso layer affords the notion of a personal public space but it is the Macro level, hashtag based exchanges that dramatically change the nature of discourse.

Regardless of the level, twitter data collection is central to any method. Most methods and examples of Twitter based social media research use the Streaming API either through existing tools or bespoke development to sample or filter (via keywords) tweets to form a collection [22]. Collections are analysed at the micro, meso and macro levels through statistical tools that offer descriptive measures of the networks implicit in the tweet collections [23].

For a large collection, another widely used technique is sentiment analysis and time series analysis [24]. Automatic sentiment analysis is increasingly popular and is used for predicting the sentiment content of texts based upon the features it identifies such as the words used. Sentiment analysis techniques can be applied over an individual tweet, (sub) collections, categorised collections for example or over a time period. There are several automated sentiment analysis tools available including Sentistrength [25] and VaderSentiment [26].

For some forms of social media analysis, qualitative textual analysis can also be appropriate. Thus a key challenge facing social scientists is determining what approach or technique to use for a given research question. Later in this paper, we outline the approach taken with respect to the various methods used in this paper.

B. Social movements and protest

As noted earlier, Twitter is both important and popular as a key tool for organising and generating communication for protestors around the world [4]. Examples of where Twitter has played a significant role include: the Iranian protests of 2009-2010 [5], the so-called Egyptian revolution of 2011 [6] and also the various Occupy protests that took place around the world [7].

In 2017, following rape and sexual assault allegations against one of the top Hollywood film makers, Harvey Weinstein, the recent #MeToo social media movement has been persistent in its focus on gender based violence, patriarchy, male power and domination [27]. The national and transnational

impact of this has been phenomenal with counter arguments protesting about a 'witch hunt' against men [28].

It is also clear that that the messaging technology is viewed differently depending upon temporality and context. Hence, it is seen as subversive by autocratic regimes, as well as a suitable technology for surveillance [8].

Scholarly research continues to point to the ways in which mainstream media play a key role in determining the agenda of social media [29]. In an analysis of 104,059 tweets related to the Delhi rape incident and social protests that took place across urban India, in line with previous scholarship, Ahmed and Jaidka (2013) conclude that traditional media still plays a pivotal role in disseminating information [13]. For instance, the authors report that less than 10% of the tweets were actually from ordinary citizens / individuals. Such a finding certainly lends credence to previous observations that have questioned the egalitarian, and democratising promises of such space [30]. Questions also arise as to whether a new public sphere is being reconstructed where ordinary citizens really do have an opportunity to form public opinion.

Nevertheless, digital activism was prevalent in the 2012 incident and again during the period around the broadcasting of the documentary in 2015. Dey defines digital activism as "political participation, activites and protests organised in digital networks beyond representational politics." [31]. Such activism as a social movement emerges as an outcome of the existence of number of interrelated factors defined by McAdam et al. [16] as:

Political-opportunities: the establishment of link between institutionalised politics and social movements.

Mobilising structures: the collective set of means which people generate solidarity and commitment to a movement or collective action. Such structures could include family and friend networks, protests and demonstration events.

Framing processes: the shared understanding, meanings and definitions that people bring to their situation.

Ray et al. [15] provide some preliminary insights into social movement mobilisation of the case under question. They analysed 1585 top tweets (those presented by Twitter in a search) around the December 2012 incident. The analysis was conducted by coding the tweets into ten content oriented categories. To understand the role of social movements, the tweets were also categorised according to the critical factors underpinning social movement formation: political opportunity, mobilising structures and framing processes. Their analysis suggests that Twitter offered mechanisms to expose political opportunity as an alternative vehicle to traditional media. For example, access to powerful political allies or celebrities. Hashtags served as mobilisation structures especially when linked to specific meetings and events. Similarly, hashtags such as "awareness", "delhirape", served as a language and symbol set to help frame the shared understanding. Eipe et al. echo this and also state that social media, more broadly, creates a less-confined political space by establishing connections with other social movements in a global community [14].

More critical comment is reported by Losh [32]. Two examples of criticism are noted. Firstly, participants in activism need to have "skin in the game" as embodied actors. Secondly, Silicon Valley has a universalising missionary mentality that also stifles creativity, thereby masking new solutions, through personalisation technologies.

C. Media and postcolonialism

In addition to the protest movement discussion, it also makes sense to provide an understanding of Western mainstream media's handling of the case under question. This is particularly important to help ground the response of the postcolonial society and to also more properly explore the role of the impact of social media on public policy.

Over the last several decades, postcolonial theory has emerged as a major intellectual critical approach. The theory is generally regarded as having been founded on the contributions of key writers including Frantz Fanon, Edward Said, Gayatri Spivak, and Homi Bhabha. Primarily, postcolonial theory seeks to problematise key historical and contemporary notions, structures and processes including colonialism, race, ethnicity, culture, racism, gender, identity, inequality, and globalisation. In short, the theory seeks to 'critique and aims to transcend the structures supportive of Western colonialism and its legacie' [33]. The watershed moment for the polemics of postcolonialism was the publication of Edward Said's *Orientalism* in 1978 [34]. In *Orientalism*, Said meticulously brought out, through close textual studies, the prejudices about and biases against the non-West that informed the colonial discourse and its meaning productions. He showed how the non-West or the orient came to occupy the space of an exotic 'Other' in the canon of Western knowledge and how this 'orientalism' as a discourse was responsible in justifying the colonial and imperial projects of the Western powers. Postcolonialism, therefore, became the rallying point to challenge the presumptions of Western knowledge systems, to comprehend the epistemologies of the non-West, to create a space where, as Gayatri Chakravorty Spivak puts it "the marginal can speak and be spoken, even spoken for" [35]. Such an epistemological framework is indeed one of the key components of postcolonial theory. Others include its critique of power in the forms of economic/cultural/economic/political/ideological domination (both historically and in the present time), its stance on the processes of otherisation, and essentialism. In the context of postcolonial theory, otherisation refers to a process by which one group uses social and psychological means to exclude or marginalise another group by focussing on differences. Whereas essentialism is understood to be the essence or "whatness" of something. In postcolonialism, essentialism implies the action of how a colonising power decides what is and isn't a particular identity. More often than not, differences and/or commonalities between groups may be overlooked to maintain a power relation.

Given this backdrop, in an examination of verbal and visual texts in United States mainstream news media reporting of the Delhi rape case, Durham argues that India / Third World is 'represented as a primitive and undisciplined space populated by savage males and subordinate women' [36]. She further asserts that in the geopolitics of sexual assault, the USA news media reinscribed social geographies of power and sex in terms of gender. Such an ethnocentric framework portrays the Third World woman as oppressed and lacking in agency, and the nation-state as incompetent and complicit in her subordination. The mediated deployment of space and place and Delhi in particular as the 'rape capital' of India serve as a key signifier

of the political economy of gender and sexuality, and hence the process of ranking one society over the other.

In her analysis of over 1500 USA (United States of America) mainstream news articles published over a period of two months, following the December 2012 rape incident, Roychowdhury (2013) argues that through its coverage, the news reporting not only created a polarity between the new and old India within a neo-liberal consumer world; but also stressed the 'notions of Western gender progressivism' as evidenced through its language including words such as 'traditional societies', 'medieval', 'rape as a weapon of power against modernity'. Here, in spite of the evidence on crimes of rape against women in the west, western space with its so-called modern cosmopolitanism is presented as safer for women [37]. The December 2012 case is used as a platform to present a dichotomy of the modern Indian woman victim, and the backward / savage /misogynist brown man. Roychowdhury cites Spivak's 1988 writings, and argues for its ongoing appeal as witnessed in western media, that is, "white men saving brown women from brown men." [38].

We use this texturised context of postcolonialism to examine the extent to which Spivak's theoretical framework can be employed to explore the interaction between the mainstream media and Twitter in the context of the BBC documentary, India's Daughter. Indeed, the broadcast of this film, and its discussion on Twitter provides a useful anchor to extend the postcolonial lens referred to by scholars, such as Durham and Roychowdhury.

3. Aims and Methods

In this study, our primary aim was focused on the interaction between the mainstream media and Twitter. In doing this, we sought to apply the theory of postcolonialism to understand the dynamics of this interaction. Additional areas of interest included an identification and exploration of the debates and discussion generated as a consequence of this controversial BBC documentary. To this end, we employed a mixed-methods approach to help understand the situation namely: a series of blogs written at the time of the broadcasting of the film in the UK (United Kingdom), and USA in March 2015; and then the collection and analysis of tweets over a period of 4 weeks (3 March 2015 - 3 April 2015). Notably, the film was broadcast during this period to coincide with International Women's Week. The topicality and contemporaneous nature of the study required drawing upon those social media blogs, written within a week or so of the broadcasting of the film. The blogs were read and analysed manually by two of the authors and led to the identification of the dominant themes used in the subsequent analysis. These blogs constituted not only as part of our data collection, but they were also useful in contextualising and in making sense of our Twitter data. Sixteen blogs from prominent bloggers were analysed. Most of the bloggers were female (11), and only three were male. Two of the bloggers were unknown.

A key challenge in conducting social media based research is the lack of standard approaches in appropriate methods for data collection and analysis. This concern tends to be further compounded by a limited range of integrated tools to support research methods that can enable the full range of types of analyses required. The Collaborative Online Social Media Observatory (COSMOS) is an example of a distributed digital social research platform that addresses these requirements [39]. However, at the time of this research, the tool was not readily available and furthermore did not integrate with our efforts at developing a learning set through manual analysis. Other tools such as Prometheus is a peer-to-peer service that collects social data from a number of sources and applies social inferencing techniques, but it is mostly concerned with privacy-aware social data management [40].

Given these concerns, we chose to access the Twitter data stream using the published Twitter Application Programming Interface (API) via our own bespoke software. Twitter offers a streaming API that can be filtered on keywords; in our case we employed the following list: "IndiasDaughter", "Leslee Udwin", "Udwin" and "banbbc". The script was kept running to collect tweets that included the keywords from 3rd March 2015 to 3rd April 2015 following the broadcast of the documentary. Over 254,000 tweets were collected amounting to around 1GB of data. Such a volume of data requires computational approaches to analysis. Figure 1 provides an illustration of the general steps in our method.

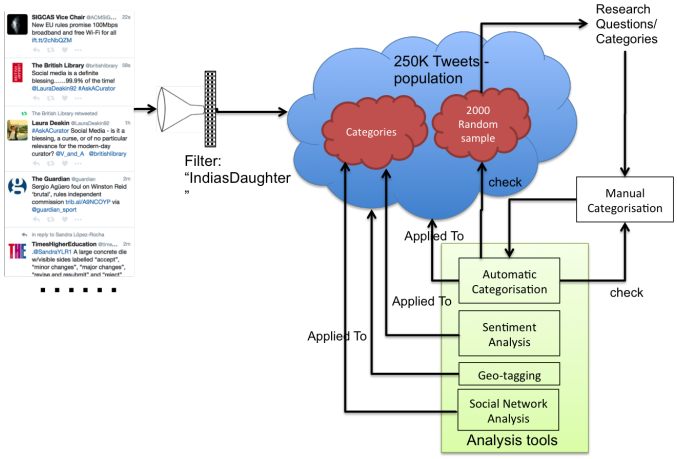


Figure 1. Method overview.

Small scale analysis of tweets of say less than 10000 is relatively straightforwardly done by human processing. Computational approaches provide additional insights that would not necessarily be possible by manual analysis. Given the volume of the tweet set, we were interested in several types of analysis. These analytical tools included: Automatic categorisation: the use of machine learning to categorise text or other data; Sentiment analysis: the use of natural language processing and text analysis to identify attitudes of a respondent with respect to a topic; Geo-tagging: plotting the location information of tweets on a map; and Social network analysis: the use of network and graph theory to investigate social structures. We discuss the use of the analytical tools below using the diagram in Figure 1 to provide an overall context.

4. Results

As indicated earlier, our research questions have centred around several exploratory areas. The blogs, existing literature and our own research questions prompted and influenced these exploratory areas which would become coding categories in our thematic analysis of the tweets.

A. Analysis of the Blogs: Emergent themes/categories

In the analysis of blogs from March 2015, several divergences were identified. Firstly, some bloggers set out to support the ban and to justify it [41]. Others proposed taking a legal stand on the matter and argued that as the case is still subjudice, the telecast should be postponed until such time as the judgement is pronounced by the court, but in no way supporting the ban [42]. Bloggers also chose to challenge the ban, ask for it to be lifted immediately, and the telecast to take place as per the schedule [43], [44]. Others took an informed and critical stand, and commentated from various perspectives such as feminism, postcolonialism or even with India's general use of "bans". Arguments for condemning and supporting the ban in the same breadth were presented [45], [43]. An analysis of the blogs helped generate a useful framework that could be applied to our Twitter dataset. A total of 7 prominent themes were identified. These included notions of legality of broadcasting the documentary (Legality/Ban), the postcolonial mindset of the film and the response from others (Postcolonialism), representations and discussions about the lawyers involved (Lawyers), contemporary feminist thinking in India (Feminism), the role of traditional media in discussions, in this case BBC and New Delhi TV (NDTV, mainstream media), representations and discussions about the role of Government of India (Government), and finally the role and value of punishment (Punishment).

B. Categorisation of tweets

Thematic analysis or categorisation is a powerful qualitative data analysis tool. The challenge is to deploy it for 1 gigabyte of data. We elected to use machine learning techniques and the use of training sets. A random sample of 2000 tweets were extracted from the tweet population and classified manually against the categories listed above by two of the researchers. Where there was discrepancy, discussion was used to agree a final classification. This tweet set of 2000 tweets was used as a "Training Set" to refine / parameterise machine learning algorithms which were then used on the entire tweet population to categorise the 254K tweets. We partitioned the 2000 training set using 3 folds. We used NTLK (Natural Language Toolkit), a Python based toolkit for natural language processing [46]. This software comes with an open source library and toolkit for natural language processing to do stemming and tokenisation, using all the words as features. We have employed a Naive Bayesian classifier to build our model. The table in Figure 2 overleaf summarises the categorisation results. Table I shows sample tweets for each classification.

This training set had around 72% accuracy (manual versus automatic categorisation). This is consistent with other research [47]. We finally applied the model to the whole dataset of 254K tweets. Tweets related to postcolonialism amounted to 26,816, representing 10.5% of the overall total and the second largest of the analytical categories. The largest category centred around tweets about legality and banning of the broadcast. More importantly, the postcolonialism tweets amounted to 23% of the tweets that were classified against the desired classifications by the machine learning algorithm. We used the 'Other' category to denote discussion that did not fall into the categories of interest.

TABLE I. Categories and sample tweets

Category	Sample Tweet
Legality / Ban	RT @GladImIndian: After government's request, YouTube pulls down Nirbhaya documentary in India: TV report #IndiasDaughter #NirbhayaDocumentary #IndiasDaughter is on BBC Four now - but was BANNED in India today because of its infamous interview with a rapist. Watch now and RT
Postcolonialism	RT @IndianWatching: @narendramodi All night tirelessly we have tried to counter SM efforts of Plan UK the foundation for which Udwin worked RT @ShekharGupta: Must make Brits pay for demeaning #IndiasDaughter also Bharat Mata. Stop playing cricket, London junkets;
Lawyers	@IndiasDaughter Sickening ideas of lawyer who defended rapist with equally disgusting ideas about role of #women in #society BBC4 RT @ImKazKohli: Showing the corrupt side of India, the lawyer was paid all he cares about his money #IndiasDaughter
Feminism	RT @gsurya: Just watched #IndiasDaughter on YouTube, brilliantly made, shows a mirror to these patriarchal medieval Sanghi types RT @neelvan_ruak: "We must know that Women are perhaps softer and weaker than Men" - Sheila Dixit on Protecting Women, and Rapes. #wtf
Media	One can only hope that the Indian government reads this. RT @soniafaleiro: I reviewed #IndiasDaughter for @guardian: http://t.co/Y6rO3H9dge RT @IndiaSpeaksPR: Sonia Singh assured that NDTV has NO LINK to the #IndiasDaughter film. Maybe thats why the makers thanked Prannoy Roy
Government	RT @xAnarchyPistolx: #IndiasDaughter Bcuz the government thinks it's better to ban the truth than actually bother to prevent rape #IndiasDaughter d doc is a grt exposŐ on the Indian mindset. Twitter ppl r elitist. hv seen louts in Haryana Delhi spk the exact same way
Punishment	Devastated. It may take 2 hands to clap but it will only take 1 hand to send you to your death. #IndiasDaughter #IAmIndiasDaughter #RIPJyoti #IndiasDaughter shows you tht even after committing rape with such BRUTALITY,; being sentenced to DEATH, the Rapist has ZERO REMORSE. WOW!

C. Sentiment Analysis

The use of natural language processing and text analysis to identify attitudes of a respondent with respect to a topic is a popular analytical tool used in tweet analysis. Both the original incident (through the brutality of the crime) and the subsequent controversial aspects of the televised documentary generated a wide range of emotions and efforts to assess the overall sentiment was deemed appropriate. We used the open source vaderSentiment0.5 tool [26] to conduct a sentiment analysis of the tweet data set. VaderSentiment represents sentiments on a scale from -1 to 1 representing negative sentiment at one end (-1) and positive sentiment at the other end (1). When the full set of tweets were subjected to a sentiment analysis using VaderSentiment, we found that tweets related to postcolonialism were ranked 3rd in association with negative sentiment. The overall compound sentiments for all categorisations is shown in Figure 2.

Analytical Category	(n)
Legality/Ban	42,870
Postcolonialism	26,816
Lawyers	17,150
Feminism	12,129
Media	8,757
Government	4671
Punishment	2186
Other	139, 841
Total	254,430

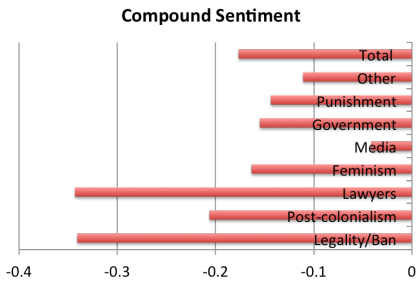


Figure 2. Automatic categorisation and overall sentiments.

D. Social Network Analysis

In this paper we are predominantly concerned with the postcolonial texture of the debate surrounding the broadcasting of the film. Hence, social network analysis on the tweet set associated with postcolonialism was conducted using two open source social network analysis tools: NodeXL [48] and Gephi [49]. The Postcolonialism tweet set (JSON file) was transformed using bespoke scripts into a form readable by NodeXL. In this data set, nodes are Twitter users and the edges represent tweets that can either be retweets or mentions. This data set was further cleaned by merging duplicate edges and the addition of weights to reduce the edge count. NodeXL was used primarily as a means for creating the GraphML format for use in the visualisation within the Gephi toolset. Gephi is better supported on MacOS and we were able to compute and visualise various social network analysis descriptive statistics. Within Gephi, the data set comprised 2000 nodes and 13243 undirected edges. Modularity computations were performed on the network. These measured the strength of division of a network into communities. We used the modularity algorithm included in Gephi [50] and produced seven communities of interest. Each of the top four communities (in size) were centred around key mainstream media actors such as @BBC, @NDTV @BDUTT, @BBCIndia and @TimesOf India. Also apparent was how these same actors were similarly ranked highly in a range of network centrality measures such as Betweenness Centrality and Eigenvalue centrality.

Degree Centrality is a measure of a node with respect to its in-bound connections and its outbound connections. If a node/actor receives many ties they are often regarded as prominent or important. Nodes that have a high out-degree are actors that are influential.

Betweenness Centrality is a measure of node that is based on the extent to which a node falls on the geodesic paths between other pairs of nodes in the network. In social network analysis, nodes with a high value for betweenness centrality are an indication of influence on information flow in a network. Hence a node with a high value is an important conduit for information flowing between nodes in the network.

Eigenvalue Centrality considers in-bound and out-bound connections and also the node's connection to other important nodes. Hence, the measure is seen as an indicator of the power of the node.

Tiryakian et al. note that "Individuals with high betweenness centrality tend to be influential because they are well informed and can affect the flow of information in a network. As a result, they are often information gatekeepers." [51]. For example, @BBC was top-ranked for both Eigenvector (0.00929400), and Betweenness centrality (13556362.612). We also observe that the use of Eigenvalue centrality to denote power is open to debate and recent results have indicated that in Twitter, users with high eigenvector centrality need not be influential users [52]

The film was shown in both USA and the UK of which the latter has a postcolonial relationship with India. Analysis of geo-tagged tweets was not considered meaningful given the low numbers of geo-tagging. None the less even from the limited, usual 1% of geo-tagged tweets (2637 tweets), we can observe how mobilisation and activity is centred around areas where there are known sub-continent diasporas. Thus figure 3 very clearly delineates tweets from diasporic communities from the UK, Scandinavia, United Arab Emirates, and USA. For the latter, the large city connurbations as well as the east and west coasts are clearly visible as sources of tweets.

To mitigate against the lack of location data, automated analysis of Twitter handle biographies could be used for analysing how sentiments and other issues vary between the countries where the films where shown. Our future work will incorporate such approaches.

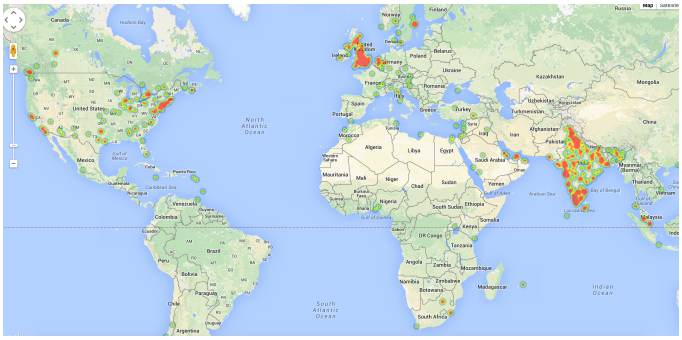


Figure 3. All geo-tagged tweets.

Our results for the centrality statistics are shown in table 1. These data are the non-normalised results. The top 25 results are shown. From the table it is clear that there is considerable overlap between the centrality measures of various nodes. @BBC for example is the most powerful node in the network and one with the highest betweenness measure. Several BBC based Twitter accounts feature as important conduits for information flow. Several of the nodes were Hindu nationalists, Others were bloggers such as @Sootradhar and @thekinshu. Presenters on Indian television programmes were represented and included @dibang, and @BDUTT for example. Most were individuals.

Figure 4 depicts the core of the 2000 node undirected network by limiting the map to vertices (nodes) that have degree range of 25-89 edges and that also have an Eigenvalue > 0.1111762270211876 (normalised). These parameters were used primarily for presentation purposes. Additionally, the vertices are scaled by Betweenness centrality to indicate the roles that vertices are playing in brokerage and diffusion of information. The various communities to which vertices belong are also indicated by the colour.

TABLE II. Centrality measures

Vertex	Eigenvector Centrality	Vertex	Betweenness Centrality
BBC	0.00929400	BBC	13556362.612
lesleeudwin	0.00635000	BBCIndia	9210704.414
YesIamSaffron	0.00344600	AlanaBowker	7989346.749
MahaveerM_	0.00344600	lesleeudwin	7436802.195
Sootradhar	0.00304900	LutyensInsider	4750346.108
ndtv	0.00301000	Sootradhar	3742481.265
TRobinsonNewEra	0.00290400	BBCWorld	2470768.572
dibang	0.00289000	gvicks	2415990.807
rishibagree	0.00261900	ndtv	2385230.780
LutyensInsider	0.00260200	vicky895	2364758.802
PMOIndia	0.00250600	dibang	2340111.494
DrJwalaG	0.00228700	shaliniscribe	2269173.050
BBCWorld	0.00228700	sunnysingh_nw3	2145012.461
thekinshu	0.00193500	M_Lekhi	1981712.913
Keisar_	0.00184800	Keisar_	1967994.765
mediacrooks	0.00169900	rishibagree	1806544.314
narendrapjoshi	0.00168700	MahaveerM_	1738846.955
mariawirth1	0.00155700	rvaidya2000	1634460.299
rvaidya2000	0.00149300	mariawirth1	1634460.299
sankrant	0.00147700	seemagoswami	1523056.837
AskAnshul	0.00143200	YesIamSaffron	1458497.885
BDUTT	0.00142300	SuchayVora	1275945.185
surnell	0.00138700	radhabharadwaj	1259731.335
htTweets	0.00134600	PMOIndia	1206559.332
rohitagarwal86	0.00129200	anniegowen	1176381.602

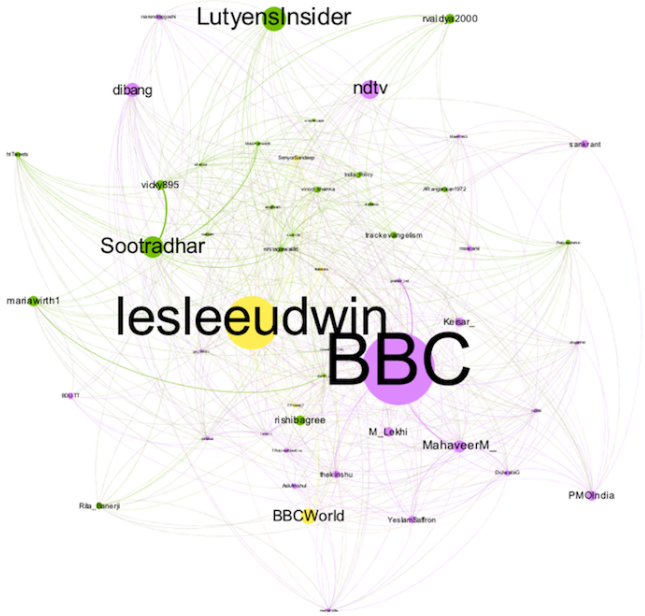


Figure 4. Power diffusion

5. Discussion

Our results indicate that mainstream media appear to be effective in utilising the communication spaces created by social media. Thus, for example, the BBC and NDTV Twitter accounts feature in both measures in Table II. However, we can observe some impact of social media, namely Twitter, through the centrality measures observed for some influential bloggers (@Sootradhar and @thekinshu) but noticeably not from those bloggers who were perceived to have set the agenda through their articles. The positioning of social media as a force for dissemination and a viable growing independent "Fifth Estate" as an alternative to mainstream media based journalism [53] should be questioned.

The BBC finds mention in around 50% of the tweets that were posted on the issue of the ban (categorised as Postcolonialism), but there were very few tweets from the various BBC official handles in that data set (<10). The BBC World News, with the handle @BBCWorld, on the other hand retweeted some of the tweets coming from @BBCIndia. These include the following: @BBCIndia: Letter from Director of BBC Television to Indian broadcast ministry on #IndiasDaughter. The letter that was attached to this tweet, was the one written by Danny Cohen, the Director of Television at the BBC, to the Joint secretary at the Ministry of Information and Broadcasting, Government of India. The following quote from the letter helps disambiguate the BBC's position:

...We appreciate your concern but we feel India's Daughter has a strong public interest in raising awareness of a global problem and the BBC is satisfied with the editorial standards of the film... The purpose of including the interview with the perpetrator was to gain an insight into the mind-set of a rapist with a view to understanding the wider problem of rape and not just in India. [54]

Further, the letter goes on to make clear whilst there is a notional respect for Indian legal jurisdiction, the BBC would bring forward the broadcast of the documentary to an earlier slot. Reasons are not given, and one possible inference is potential legal action preventing the broadcast.

...It should be noted, although the BBC is happy to take your views into consideration, we are not planning to transmit the film in any territory which lies under Indian legal jurisdiction...after lengthy and careful consideration we have decided to show India's Daughter on BBC Four in the UK at 22:00 tonight. [54]

What is attempted in this letter has a direct bearing on our postcolonial understanding, for it redraws the battle-lines for this debate and depicts 'rape', the central issue of the film, as a *global* and not an India specific malaise. And in so doing, it attempts to steer away from any alleged bias in the representational politics despite the obvious conflation in the title of the documentary which asserts a national identity to the malaise of rape. The letter's reference to *jurisdiction* appears to underplay the ubiquity and reach of social media, and seems to embolden the role of mainstream media. The Twitter discussions amplify some of the positions identified by the bloggers. Thus, Mehra notes:

There are many questions regarding why some sections of the widely-read Western media are so obsessed with reporting and analysing rapes in India while often conveniently ignoring similarly grievous nature of the problem in their own backyards. and asks: *Can you imagine if an Indian film-maker were to make a documentary on the thousands of cases of sexual abuse and 'grooming' of young girls in Rotterham (sic) between 1997-2013, and title it "Britain's Daughter"?* [55].

The blogs and tweets from other central nodes (including bloggers and Hindu nationalists) not only accused BBC of being prejudiced in their framing of the narrative, but also came down upon the broadcaster for doing so at the expense of not giving adequate attention to gruesome incidents and issues at 'home'. For example:

@*YesIamSaffron* How Many Documentaries ve u Seen @BBC Reporting US Rapes or Biggest Child Exploitation Scandals in UK? #BanBBC #UHF http://t.co/hglD7lfHOP (297 RTs)

@*thekinshu* Dear @lesleeudwin can you pl tell me why you afraid to make a Documentary on your own Rape Case in London.? #BanBBC

The above tweets demonstrate a strong dynamic of nationalism, and the 'other' that portrays the BBC and the documentary film maker as an oppressive force. There is a challenge to the claim made by the 'other' to be the holder of morality, and superiority. Moreover, the 'other's' stance on human rights and gender is also questioned in highlighting its own social problems. Crucially, the BBC and Leslee Udwin serve to represent the Western 'other'. In doing so, they become emblematic of the UK, the ex-colonial power and its representation of postcolonial India.

A remarkable anomaly was the hijacking of the debate by the UK anti-muslim, far right party leader, Tommy Robinson and how Hindu nationalist Twitter users used a tweet (retweeted over 2000 times) from Tommy Robinson for their own ideological stance (see @YesIamSaffron above) bringing to the foreground a curious postcolonial paradox.

@*TRobinsonNewEra* We are in the middle of the biggest rape and child exploitation scandal in our countries history, and the @BBC are focusing on india's rapes (2078 RTs)

It would seem that networks and how power is transferred through such networks appear to be structured by colonial and national tensions and complicated by religion, and other social divisions.

The broadcasting of the documentary received widespread analysis from a range of contributors from both academia and elsewhere. Titha Bhattacharya, writing in the International Socialist Review, argued strongly against any restrictions on its telecast, partial or total [56]. Her review recognised Udwin's historical credentials but contrasts the reasons given for the crime ("the logic of traditional masculinity") with what Bhattacharya considers as a far more important factor, namely the promise of neoliberalism being tantalisingly beyond reach of the working class. Separately, Banaji raises concerns around vicious sexual and gender violence that go beyond national boundaries and it is the lack of reference in the documentary that both weaken the documentary and allow the debate about postcolonialism and nationalism to dominate.

> *"Nevertheless, the film would have been far more resonant and powerful, had the Indian context been linked creatively, even briefly, to wider histories of rape around the globe..."* [57]

Banaji's other themes resonate with our own: She notes the postcolonial/colonial history of orientalism and its resultant outcomes of an impulse leading to disavowal of problems such as misogyny amongst western communities. At the same time, the documentary narrative ends up otherising certain types of Indian men while distancing Indian elites from this world view. The former point is repeated made in the postcolonial category tweet set and discussed above. Banaji also notes that the significance of judicial ethics and legal context do not appear to have had much impact on the BBC nor Udwin, most pertinently demonstrated by the BBC's letter offering an explanation for bringing forward the broadcast of the documentary.

A. Social Movements

The broadcast of the documentary and the resulting furore presents a peculiar type of social movement in that competing movements denoted by our categories such as Legality/Ban, Feminism, and Postcolonialism are observable. Each appears to represent a political opportunity, exhibits mobilising factors and documents a shared symbolic language. Considering the latter and consistent with Ray et al. [15], social media appears to have played a key role in exposing political opportunities to express nationalistic views as a response to perceived post-colonial sentiments. A groundswell of sentiment was achieved through mobilisation structures such as the Twitter handles of traditional media, key bloggers and perversely, the UK anti-muslim, far right party leader, Tommy Robinson. The network centrality measures presented in table II indicate the role that key nodes played in information diffusion (sharing of information). Mobilisation structures were further enhanced through hashtags related to the documentary. We suggest that the a shared understanding is implied through frequency word counts. Thus the top twenty words forming a symbolic language (cleaned) were: IndiasDaughter, India, Ban, Documentary, rape, BBC, watch, daughter, rapist, indian (sic), Udwin, shame, women, NDTV, film, Leslee, Like, People and Nirbhaya. The symbolic language for a framing process seems to be embedded in these frequently cited words.

B. Threats to Validity

Internal validity issues centre around whether the micro posts (tweets) are appropriately categorised by the machine learning algorithms. Here, care was taken to use a manual process of categorisation of 2000 randomly selected tweets to help develop the training set for the machine learning software. While there is confidence that the categorisation has operated at the reference 72% accuracy [47] it is noted that a significant percentage of tweets could not be automatically categorised. Note, however, no claims on external validity (wider generalisations to different domains) are being made.

Retweets form a significant portion of the tweet data set. It is possible that the outcomes may have been different if the tweets analysed had not been retweets. However, retweets are a core feature of Twitter. Moreover, as Halavais [58] points out:

> *"Retweeting a message represented both an affirmation of the contents of a particular tweet, and a way of spreading a conversation more widely."*

From this perspective, we propose that retweets do not affect the outcomes significantly as retweets invite a structure for conversation and comment as well as being a 'people's microphone'.

Methodological notions of validity, reliability and repeatability present a concern for much of social media research as it is challenging to be definitive that data collected from social media is actually representative of the phenomena of interest. Twitter users commenting on a particular phenomena are self-selecting. Such concerns can be partially mitigated by the use of large scale analysis (using large data sets) but nonetheless risks such as social media being generative of the behaviours

it aims to document are paramount [59]. A further concern is related to decisions around treating journalists as individual citizens or as part of the overall machinery of the Fourth Estate. We view the latter as a more representative definition.

Social media research requires inter-disciplinary thinking, but arriving at a common ground whereby a sociological theory can be adequately expressed for computational purposes is an ongoing research challenge [60].

Social network analysis, and in particular the centrality measures can offer some insight about power diffusion in networks, but as noted earlier, are open to debate. Eigenvalue centrality for example, unless correlated with inbound/outbound data may not be a good indicator of power.

Furthermore, there are limited open source seamless software tool chains that addressed the types of analyses that we utilised. There are risks in moving data between software tools. Importantly, we restrict our claims to the data that we have collected and make no generalisations for other contexts.

The ethics of using data published in social media is also of concern. The approach taken in this paper considers two key dimensions, risks of identification / disclosure to users and ethical risks around the content of the micro-blogs. The data collection (both the blogs and the tweets) are from users who would be classified as low risk users as either the user is not identifiable from a Twitter profile or is from a public, official or bot account. Ethical risks related to the content of the tweets are also limited. While the content is at times provocative and antagonistic, the classification of the users as low risk does not warrant opt in permissions before publication. Possibilities of masking identities can address confidentiality concerns, but it is important to note that Twitter Terms and Conditions state that tweets must be given in their original form and attributed to the individual who posted the tweet. Furthermore, informed consent becomes near impossible when dealing with data at large scale. The dominant use of hashtags by Twitter users also supports the notion that such users were broadcasting their thoughts specifically on a subject in a public discussion [61]. Copyright is of less concern. A user may be considered to be the author of a tweet, but the tweet may not be protected under copyright law as there is general consensus that protection generally requires either 'originality' or 'sweat of the brow', i.e., significant expenditure of labour [62].

6. Conclusion

Social media is becoming an important communication channel in the modern world, however its role as a democratic tool, its reach and therefore direct impact remain questionable in terms of policy formation. The recent presidential election in USA has certainly brought to the foreground, the role of social media and the need for policy discussions [63]. In particular, there is strong evidence of the critical role of social media in social mobilisation efforts for causes. Our study shows that the traditional media and their components such as individual journalists continue to play a central role in these new spaces. The influence and reach of the ordinary citizen is less well pronounced so it is uncertain that the so-called Fifth Estate is really coming to the fore. Methodological concerns remain challenging. For example, common grounds whereby a sociological theory, for example, postcoloniality (as in this paper) can be adequately expressed for computational purposes remains an open question. Software tools that can support both sociological reasoning and computational analysis in a linked and coherent way have the potential to make a significant impact on methodological concerns of validity and reliability. However, considering that there many computational analysis styles possible, this coherent linking may be some way off. Some of our future research effort is directed at developing a software tool chain that social scientists will be able to use independently of computer scientists.

References

[1] B. S. Barn, R. Barn, F. Raimondi, and U. Mukkerjee, "Exploring the interaction between the Fourth Estate and Twitter: A Case Study," in The Third International Conference on Human and Social Analytics (HUSO). IARIA, 2017.

[2] I. Facts, "Topic: Internet usage in India," 2016, Retrieved: 31-05-2018. [Online]. Available: https://www.statista.com/topics/2157/internet-usage-in-india/

[3] C. R. Sunstein, # Republic: Divided democracy in the age of social media. Princeton University Press, 2018.

[4] M. Tremayne, "Anatomy of protest in the digital era: A network analysis of Twitter and Occupy Wall Street," Social Movement Studies, vol. 13, no. 1, 2014, pp. 110–126.

[5] E. Morozov, "Iran: Downside to the "Twitter revolution"," Dissent, vol. 56, no. 4, 2009, pp. 10–14.

[6] A. M. Attia, N. Aziz, B. Friedman, and M. F. Elhusseiny, "Commentary: The impact of social networking tools on political change in Egypt's "Revolution 2.0"," Electronic Commerce Research and Applications, vol. 10, no. 4, 2011, pp. 369–374.

[7] M. C. Stoddart and D. Tindall, "We've also become quite good friends: Environmentalists, social networks and social comparison in British Columbia, Canada," Social Movement Studies, vol. 9, no. 3, 2010, pp. 253–271.

[8] E. Morozov, The net delusion: The dark side of Internet freedom. PublicAffairs, 2012.

[9] J. Habermas, The structural transformation of the public sphere: An inquiry into a category of bourgeois society. MIT press, 1991.

[10] V. Belair-Gagnon, S. Mishra, and C. Agur, "Reconstructing the Indian public sphere: Newswork and social media in the Delhi gang rape case," Journalism, vol. 15, no. 8, 2014, pp. 1059–1075.

[11] N. Newman, W. H. Dutton, and G. Blank, "Social media in the changing ecology of news: The fourth and fifth estates in Britain," International Journal of Internet Science, vol. 7, no. 1, 2012, pp. 6–22.

[12] R. Barn, "Social media and protest - the Indian Spring?" Huffington Post, 2013, Retrieved: 31-05-2018. [Online]. Available: http://www.huffingtonpost.co.uk/professor-ravinder-barn/india-social-media-and-protest_b_2430194.html

[13] S. Ahmed and K. Jaidka, "The common man: An examination of content creation and information dissemination on Twitter during the 2012 New Delhi Gang-Rape Protest," in International Conference on Asian Digital Libraries. Springer, 2013, pp. 117–126.

[14] J. J. Eipe, T. Varghese, and S. M. Veranani, "'india against corruption' movement: An Online Version of a Non-violent Mass Movement," Quarterly Journal of the Gandhi Peace Foundation, vol. 34, no. 3, 2012, pp. 343–353.

[15] D. Ray and M. Tarafdar, "How does twitter influence a social movement?" 2017.

[16] D. McAdam, J. D. McCarthy, and M. N. Zald, Comparative perspectives on social movements: Political opportunities, mobilizing structures, and cultural framings. Cambridge University Press, 1996.

[17] J. C. Jenkins, "Resource mobilization theory and the study of social movements," Annual review of sociology, vol. 9, no. 1, 1983, pp. 527–553.

[18] F.-M. Titzmann, "The voice of the youth, locating a new public sphere between street protest and digital discussion," Studying Youth, Media and Gender in Post-Liberalisation India: Focus on and beyond the Delhi Gang Rape, vol. 6, 2014, p. 79.

[19] K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann, Twitter and society. P. Lang, 2014, vol. 89.

[20] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann et al., "Life in the network: the coming age of computational social science," Science (New York, NY), vol. 323, no. 5915, 2009, p. 721.

[21] A. Bruns and H. Moe, "Structural layers of communication on Twitter," in Twitter and society. Peter Lang, 2014, vol. 89, pp. 15–28.

[22] D. Gaffney and C. Puschmann, "Data collection on Twitter," Twitter and society, 2014, pp. 55–67.

[23] A. Bruns and S. Stieglitz, "Metrics for understanding communication on Twitter," in Twitter and society. Peter Lang, 2014, vol. 89, pp. 69–82.

[24] M. Thelwall, "Sentiment analysis and time series with Twitter," Twitter and society, 2014, pp. 83–96.

[25] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," Journal of the Association for Information Science and Technology, vol. 63, no. 1, 2012, pp. 163–173.

[26] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[27] G. McSherry, "#metoo: Bringing victims of sexual assault into the spotlight won't stop rape. Bringing justice to perpetrators will," 2018, Retrieved: 31-05-2018. [Online]. Available: https://isreview.org/issue/97/indias-daughter

[28] J. Kinos-Goodin, "Catherine deneuve says the #metoo movement has turned into a 'witch hunt'," 2018, Retrieved: 31-05-2018. [Online]. Available: https://goo.gl/XbebTW

[29] G. Lazaroiu, "The social construction of participatory media technologies," Contemporary Readings in Law and Social Justice, vol. 6, no. 1, 2014, p. 104.

[30] F. Rebillard and A. Touboul, "Promises unfulfilled? Journalism 2.0, user participation and editorial policy on newspaper websites," Media, Culture & Society, vol. 32, no. 2, 2010, pp. 323–334.

[31] A. Dey, "A brief exploration of the effects of ICTs and social media on the gender activism in india post december 16th 2012," CECS-Publicações/eBooks, 2016, pp. 187–204.

[32] E. Losh, "Hashtag feminism and Twitter activism in India," Social Epistemology Review and Reply Collective, vol. 3, no. 3, 2014, pp. 11–22.

[33] J. Go, "For a postcolonial sociology," Theory and Society, vol. 42, no. 1, 2013, pp. 25–55.

[34] E. W. Said, Orientalism. Vintage, 1979.

[35] G. C. Spivak, Outside in the teaching machine. Routledge, 2012.

[36] M. G. Durham, "Scene of the crime: News discourse of rape in India and the geopolitics of sexual assault," Feminist Media Studies, vol. 15, no. 2, 2015, pp. 175–191.

[37] R. Barn and R. A. Powers, "Rape Myth Acceptance in Contemporary Times: A Comparative Study of University Students in India and the United Kingdom," Journal of Interpersonal Violence, May 2018. [Online]. Available: https://doi.org/10.1177/0886260518775750

[38] P. Roychowdhury, ""The Delhi gang rape": The making of international causes," Feminist studies, vol. 39, no. 1, 2013, pp. 282–292.

[39] P. Burnap, O. Rana, M. Williams, W. Housley, A. Edwards, J. Morgan, L. Sloan, and J. Conejero, "COSMOS: Towards an integrated and scalable service for analysing social media on demand," International Journal of Parallel, Emergent and Distributed Systems, vol. 30, no. 2, 2015, pp. 80–100.

[40] N. Kourtellis, J. Finnis, P. Anderson, J. Blackburn, C. Borcea, and A. Iamnitchi, "Prometheus: User-controlled p2p social data management for socially-aware applications," in Proceedings of the ACM/IFIP/USENIX 11th International Conference on Middleware. Springer-Verlag, 2010, pp. 212–231.

[41] N. Menon, "India's Daughter and sexual violence: The issues at stake - Europe Solidaire Sans Frontieres," 2015, Retrieved: 31-05-2018. [Online]. Available: http://www.europe-solidaire.org/spip.php?article34577

[42] K. Krishnan, "Rape, rape culture and the debate over India's Daughter," 2016, Retrieved: 31-05-2018. [Online]. Available: http://scroll.in/article/711369/rape-rape-culture-and-the-debate-over-indias-daughter

[43] V. Narayan, "Why the documentary India's Daughter should be aired," 2015, Retrieved: 31-05-2018. [Online]. Available: https://vinodnarayan.com/2015/03/04/should-the-documentary-indias-daughter-be-aired/

[44] S. Dev, "What does banning the documentary India's Daughter accomplish exactly?" 2015, Retrieved: 31-05-2018. [Online]. Available: https://goo.gl/mHYYUN

[45] S. Krishnan, "India's Daughter: When men come together, but women stand divided," 2015, Retrieved: 31-05-2018. [Online]. Available: http://www.tarshi.net/inplainspeak/indias-daughter-when-men-come-together-but-women-stand-apart/

[46] I. Facts, "NLTK 3.2.4 documentation," 2017, Retrieved: 31-05-2018. [Online]. Available: http://www.nltk.org

[47] A. Schulz, E. L. Menca, and B. Schmidt, "A rapid-prototyping framework for extracting small-scale incident-related information in microblogs: Application of multi-label classification on tweets," Information Systems, vol. 57, 2016, pp. 88–110.

[48] M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave, "Analyzing (social media) networks with NodeXL," in Proceedings of the fourth international conference on Communities and technologies. ACM, 2009, pp. 255–264.

[49] M. Bastian, S. Heymann, M. Jacomy et al., "Gephi: an open source software for exploring and manipulating networks." ICWSM, vol. 8, 2009, pp. 361–362.

[50] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, 2008, p. P10008.

[51] E. A. Tiryakian, A. Gruzd, B. Wellman, and Y. Takhteyev, "Imagining Twitter as an imagined community," American Behavioral Scientist, vol. 55, no. 10, 2011, pp. 1294–1318.

[52] P. Howlader and K. Sudeep, "Degree centrality, eigenvector centrality and the relation between them in Twitter," in Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE International Conference on. IEEE, 2016, pp. 678–682.

[53] W. H. Dutton, "The fifth estate emerging through the network of networks," Prometheus, vol. 27, no. 1, 2009, pp. 1–15.

[54] D. Cohen, "BBC - letter from Danny Cohen to Sh. Rakesh Singh regarding concerns about Storyville - India's Daughter - Media Centre," 2015, Retrieved: 31-05-2018. [Online]. Available: http://www.bbc.co.uk/corporate2/mediacentre/statements/dc-letter-re-indias-daughter

[55] B. Mehra, "India's Daughter: An Indian Daughter's Opinion," 2014, Retrieved: 31-05-2018. [Online]. Available: http://swarajyamag.com/culture/indias-daughter-an-indian-daughters-opinion/

[56] T. Bhattacharya, "India Daughter Neoliberalism's dreams and the nightmares of violence," 2014, Retrieved: 31-05-2018. [Online]. Available: https://goo.gl/dM4x8a

[57] S. Banaji, "Five issues raised by BBC india's daughter documentary," 2015, Retrieved: 31-05-2018. [Online]. Available: http://blogs.lse.ac.uk/southasia/2015/03/07/five-issues-raised-by-the-bbc-indias-daughter-documentary/

[58] A. Halavais, "Structure of Twitter: Social and technical," Twitter and society, 2014, pp. 29–42.

[59] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," PloS one, vol. 11, no. 3, 2016.

[60] D. Murthy, "Digital ethnography an examination of the use of new technologies for social research," Sociology, vol. 42, no. 5, 2008, pp. 837–855.

[61] L. Townsend and C. Wallace, "Social media research: A guide to ethics," 2015, Retrieved: 31-05-2018. [Online]. Available: http://www.gla.ac.uk/media/media_487729_en.pdf

[62] M. Beurskens, "Legal questions of Twitter research," Twitter and society, 2014, p. 123.

[63] J. Boorstin, "Facebook, Snapchat and Twitter played a bigger role than ever in the election," 2017, Retrieved: 31-05-2018. [Online]. Available: https://goo.gl/YJ62dm

Modeling and Simulation of Complex Agents for Analyzing Communication Behavior in Social Media

Fabian Lorig, Stephanie C. Rodermund, Jan Ole Berndt, Ingo J. Timm

TriLabS@CIRT
Business Informatics 1
Trier University
54296 Trier, Germany
Email: [lorigf,rodermund,berndt,itimm]@uni-trier.de

***Abstract*—Personal and professional communication increasingly take place on online platforms like Facebook, Twitter, and Google+. This trend is accompanied by new forms of communication and novel communication dynamics which can no longer be analyzed using conventional methods only. By means of agent-based social simulation, actors in social media can be modeled and their behavior can be investigated. In this work, we use complex agents to model and simulate Twitter users with respect to analyzing communication behavior accompanying a television program. To this end, we present a case study and show how different actor constellations within a population of agents influence the dynamics of this communication.**

Keywords–Agent-Based Modeling; Intelligent Software Agents; Agent-Based Social Simulation; Social Media Analysis; Social Actor Types.

I. INTRODUCTION

It is no longer possible to imagine a world without information and communications technology. *Social media* like Facebook, Twitter, or Google+ have become predominant means of communication for both private and professional users and social activities increasingly take place on online platforms. This trend is accompanied by new forms of communication which also result in novel communication dynamics. For analyzing and understanding such emerging phenomena, the suitability of conventional approaches is limited. To overcome shortcomings of existing approaches, innovative methods such as agent-based modeling and computer simulation can be used. They allow for a systematic and adaptive investigation of communication dynamics in social media [1].

Understanding communication processes and dynamics is important in both commerce and politics, e.g., to derive communication strategies for social media. By this means, marketing campaigns can reach a vast target audience through viral communication dynamics [2]. However, if the same dynamics distributes negative opinions, emerging mass criticism can endanger a company's commercial success. Therefore, it is crucial to anticipate likely reactions of social media users to such campaigns to avoid unintended effects or to develop appropriate counter strategies to those effects [3].

Nonetheless, the inherent distribution of social media and the dynamics of user interactions therein make it difficult to analyze and understand that kind of communication. Thus, manual analysis has been complemented using methods from computational linguistics, data mining, and simulation [4]. These methods help recognize conversation topics, discern user communities, and model information diffusion in social networks.

Especially agent-based social simulations [5] are a promising technique for understanding complex dynamics of interrelated communication activities. They model the behavior of humans by means of artificial agents to explore the effects of different social actor constellations and various situations in an experimental environment. For instance, viral dynamics of mass phenomena in social media like the *harlem shake* [6] can be reproduced by using artificial agents for representing media users [7]. Each agent can react to other agents' communication activities in a simulated media environment. This interaction leads to complex dynamics. Exploring various user constellations and agent decisions in a controlled experiment helps understand these dynamics in real world social media.

However, agent-based simulation for social media analysis requires a model of user motivations and resulting behaviors to yield realistic results. Agents must be complex enough to explain *why* particular communication processes emerge and which effects potential reactions to them will provoke. Thus, in this paper, we present an agent-based model of user behavior for analyzing communication dynamics in social media. This is a first step toward a simulation-based decision-support method for developing and testing social media communication strategies as proposed by Berndt et al. [3].

The paper is structured as follows. Section II provides an overview of the foundations of social media analysis, social actor theory, and agent-based modeling as a technique for dynamic analysis. Subsequently, Section III describes our concept of complex agents for modeling user behavior. This concept covers individual social actors, their respective decision-making, as well as populations of media users. Section IV applies that concept to communication processes on Twitter, which accompany a German television program and specifies the agents' decision functions accordingly. In Section V, we evaluate our model by simulating user behaviors in the presented scenario. Furthermore, a sensitivity analysis is conducted to identify how compositions of different actor types influence the communication behavior of individual agents. Finally, Section VI concludes on our findings and gives an outlook on future work.

The presented model was first introduced in earlier work of the authors [1]. As extension of this preceding contribution

and to illustrate the agent decision-making, this work provides and discusses formal specifications of the decision behavior of each of the four applied actor types. Moreover, the results of a sensitivity analysis are presented to shed light on the influence that different actor types have on the communication behavior of the agents.

II. Foundations

To analyze, model, and simulate user behavior in social media, it is necessary to understand communication processes within those media. Such processes depend on the underlying platforms that structure possible communication, the observable communicative activities, as well as the social actors performing these activities. Thus, the following sections discuss approaches and theories for analyzing and modeling these aspects. In addition, we give an overview of the state of research in agent-based modeling of human behavior to provide a foundation for our approach to user behavior analysis.

A. Social Media

Social media structure communication processes by providing options to their users to connect with each other. In terms of graph theory, such a structure can be described by a set of users (nodes) and relationships between the users (edges) [8]. Graphs can be unidirectional, defining the direction of the relationship, or bidirectional, connecting two nodes without providing information regarding that direction.

For instance, the online social network Twitter can be modeled as a directed graph. In contrast to most other platforms, which consist of bidirectional relationships between users, a distinction between *followers* and *followees* is made on Twitter. That is, a user actively and voluntarily decides, which other users to *follow* for receiving their status updates. Following another Twitter participant makes the following user become a *follower*. However, a *followee*, i.e., the user being followed, does not need to follow his or her followers.

When analyzing the structure of social media, a typical task is to identify and assess the importance of the most influential users by means of centrality measures [9]. The *degree* of centrality corresponds to the total number of edges a node has. Hence, it is a measure of a node's interconnectedness in a graph. Nodes having a high *degree* (compared to other nodes) act as hubs for information diffusion within a social network.

By contrast, a graph's *density* denotes the interconnectedness of an entire network. It is used to compare different network structures and their impact on information propagation. The *density* is defined by the ratio of the number of existing edges and the maximum number of edges in case every pair of nodes would be connected by an edge (complete graph).

B. Communication

Human communication can be considered as a sequence of actions by individuals, where the behavior of a sender influences the behavior of a receiver [10]. The sender uses a set of characters to encode a message, which is transmitted using an information medium. The receiver uses an own set of characters to decode and interpret the message and returns a feedback using the same mechanism [11]. The formulation and transmission of messages by the sender as well as the corresponding reaction by the receiver form the communicative activities available to users of social media.

However, the shift of communication into technical media is accompanied by a loss of information. The transmission of messages is ensured, yet, the receiver does not know whether a message was interpreted correctly. On Twitter, communication results can only be returned by replying to a Tweet using another Tweet. Consequently, conversations are formed as sequences of messages, which refer to or forward previous ones [12]. To that end, Twitter provides mechanisms for replying to other tweets and for addressing a tweet to a certain person. Using the @-symbol followed by the name of a user or by putting the prefix *"RT"* (retweet) at the beginning of a tweet, dialogs or conversations can be defined and identified.

In addition to the structuring of dialogs, Twitter users can use another operator for classifying the content of a message. The content provides information about the intention as well as the context of communication. On Twitter, the #-symbol (hashtag) is used for categorizing messages and for marking keywords. This simplifies filtering Tweets according to certain topics, which makes this kind of communication easily accessible to media studies and communication research. In fact, Twitter has been widely used for conducting studies of certain subjects or events, e.g., spread of news and criticism [4], [13], the activity of diseases [14], or political communication [15].

C. Social Actors

Communication is inherently social. In fact, sociality can be considered to consist entirely of communication [16]. Social systems emerge from interconnected communicative activities being selected by social actors. Those actors are influenced by an observed social situation. They decide about their reactions to that situation. This results in observable behaviors that lead to a new situation in effect (Figure 1). For example, a user can observe an ongoing conversation about a specific topic (1). She may decide to utter a controversial opinion about that topic (2). Her utterance becomes observable to other users in the form of her respective Tweet (3). This changes the conversation and provokes further reactions. Thus, the conversation on the macro-social level (4) both influences individual behaviors and emerges from them on the micro-social level.

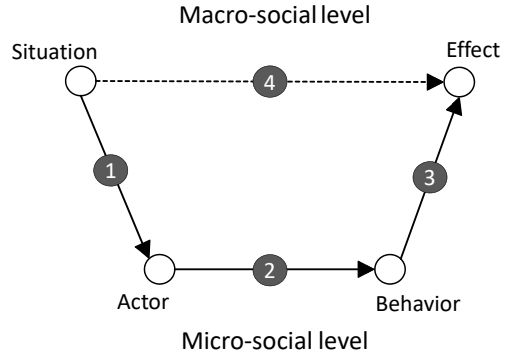


Figure 1. Emergence of macro-social effects from micro-social behavior [17].

There are several analyses of user behavior in social media available. For instance, activity frequencies on Twitter (i.e., Tweets, Responses, Retweets) have been related to user attributes and traits such as gender, age, region and political opinion [18]. While such an analysis reveals *how* social media users interact with each other, it cannot explain *why* they do it. To answer that question, other studies cover motivations

for communication. These motivations can be categorized into groups such as *smalltalk*, *entertainment*, or *information and news sharing* [19]. Additionally, they can be derived from psychological personality traits [7] [20]. Such approaches provide valuable insights into the decision-making process of social actors in diverse situations ranging from casual comments on a television series [21] to crisis communication [22].

In addition to social media specific and psychologically founded motivational categories, there are also theories of actor behavior in sociology. Sociologists distinguish between four basic social actor types, which differ in their behavior [23]. Firstly, the *homo economicus* is a rational decision-maker who strives to maximize her personal utility. Such an actor attempts to reach personal goals as efficiently as possible. Secondly, the *homo sociologicus* obeys social norms and obligations. This actor type tries to conform with expectations to avoid negative sanctions. Thirdly, the *emotional man* is driven by uncontrollable emotions such as love, anger, respect, or disgust. This leads to affective behavior in response to, e.g., unfulfilled expectations [24]. Finally, the *identity keeper* has the goal to establish and maintain a desired social role. Such an actor seeks social acknowledgment by provoking positive reactions toward stereotypical behaviors. In the remainder of this paper, we will show how these actor types can be applied to agent-based modeling of user behavior in social media.

D. Related Work: Agent-Based Modeling of Human Behavior

As discussed in the preceding section, communication processes in social media emerge from individual activities of the participating users. For investigating emergent phenomena, agent-based modeling has been established as a standard means. Artificial agents are capable of decision-making, communication, and goal-directed behavior [25]. By modeling real world actors as software agents, individual behavior and anticipation of behavior on the micro level can be simulated resulting in emergent effects on a macro level [26] [27]. In terms of social sciences, using such actor models for simulation studies is referred to as agent-based social simulation [5].

The majority of agent-based models in social media analysis is concerned with *information propagation*. They aim at identifying a group of users, which can propagate information, i.e., a message, to as many users as possible [28]. The users are frequently modeled as agents being connected by neighborhood relations in cellular automata [29] or general network graphs [30]. These agents often have particular behavioral rules that fire if a certain activation threshold is reached. Such a threshold denotes the required strength of influence (e.g., a number of received messages) on an agent until it becomes active itself. This method is particularly relevant for planning advertising strategies since viral marketing campaigns make use of information propagation effects [2], [31].

While threshold models are usually investigated by means of simulation studies, there are also *analytical approaches* to agent-based modeling of opinion formation. These focus on the interactions among agents, which lead to the diffusion and adoption of opinions in a process of compromising [32]. They model these interactions by means of thermodynamics [33] or the kinetic theory of gases [34]. These methods describe the emergence of macro-social phenomena from micro-social interactions using differential equations. This allows for analyzing the resulting opinion dynamics mathematically.

However, there is a discrepancy between these threshold and analytical models on the one hand, and the mentioned sociological perspectives on decision-making on the other. While these methods describe *how* opinion and communication dynamics occur in agent-based social simulations, they lack the descriptive power to analyze *why* this happens. That is, they focus on the dynamics between interacting agents and treat the agent population as a homogeneous mass. For instance, in kinetic theory, gas molecules behave solely according to their current states and their mutual influences without having individual habits. The same holds for cellular automata in which all cells, i.e., agents, are usually homogeneous and strongly restricted in their neighborhood relations. As a result, the discussed approaches largely disregard modeling individual motivations for decision-making such as described by social actor types.

For utilizing agent-based social simulation to understand human behavior and to develop communication strategies, it is necessary to apply more elaborate agent decision approaches. Agents must have individual motivations to allow for analyzing who participates in communication processes for which reason [3]. Since, in social media, different users react differently to the same message, this should also be the case for artificial agents in a simulation model. In fact, a wide range of agent decision-making architectures based on philosophy, psychology and cognitive science is readily available [35]. In addition, sociological theory and agent-based modeling have been combined in the interdisciplinary field of *socionics* [36]. In that context, the described social actor types can be utilized to explain social behavior in an agent-based simulation.

Dittrich and Kron model social characters by means of actor types and combinations between these types [23]. They simulate the so-called "bystander dilemma" in which persons must decide whether or not to help a victim of physical violence. In their model, agents implementing the *homo sociologicus* and *identity keeper* roles feel obliged to help while *homo economicus* and *emotional man* flee the situation. Combining these dispositions on both an individual and on a population level leads to complex macro-social behaviors. This makes that approach a promising candidate for a transfer to modeling user behavior in social media as described in the following section. In this regard, Berndt et al. [1] present a first approach towards the use of sociological agent types in agent-based modeling for the simulation of user behavior.

III. Concept: Modeling User Behavior

In this section, we adapt the agent-based decision-making approach by Dittrich and Kron [23] to modeling communicative user behavior in social media. That is, we model the selection of messages about a specific topic to be published on a social media platform within a limited time frame [37].

Our modeling and simulation concept is structured as shown in Figure 2. Each decision-making situation receives an input of one or more keywords, which describe the situation (e.g., a list of hashtags or abstract topic description). The respective output consists of messages being published at the social network platform by the population of agents. In order to produce that output, each agent observes the situation and calculates expected values for its potential reactions according to its respective social actor type and depending on the activities of other agents. It then selects its next message (or chooses not to publish any message) with respect to these expected values. The following sections describe the actor

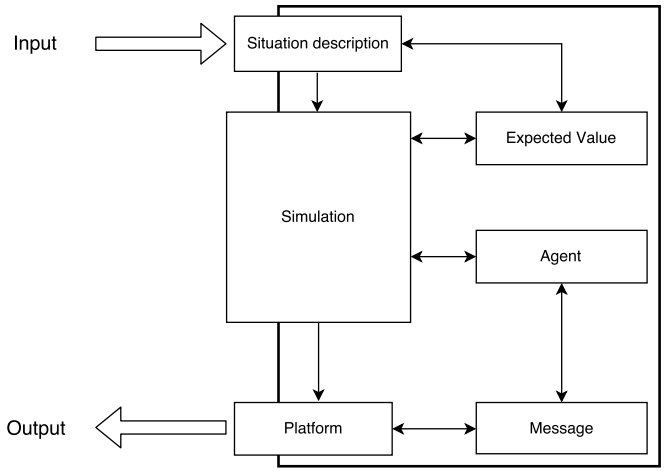


Figure 2. Structure of the user modeling and simulation concept.

types, their combinations, and the resulting agent populations.

A. Social Actor Types and Decision-Making

Besides the current situation, its social actor type determines an agent's decision-making. To that end, we model each type by means of a function EV that returns an expected value for each available activity option. For a *homo economicus*, this amounts to a standard utility function. Contrastingly, a *homo sociologicus* prefers socially adequate behaviors over controversial actions. Such an agent makes its behavior dependent on contributions to a conversation by other agents. In addition, while the *identity keeper* has a genuine desire to further any kind of discussion, the *emotional man* only becomes active when being emotionally affected by the situation.

All of the expected value functions should cover the same range of values to make them comparable with each other. That range depends on the number of available activity options and their effects in a particular application scenario. Each option can either have a positive, neutral, or negative effect on an agent's goals. For instance, a scenario with five possible messages can be encoded through the following set of values: $\{-1, 0, 1, 2, 3\}$. In this case, a message is either detrimental to an agent's goals (-1), it can be neutral towards them (0), or it furthers its motivations to different extents (1–3). Then, the agent can select its actions as follows.

$$\arg\max_{a} EV_i(s, a)$$

Each actor type i ($i \in \{$*homo economicus*, *homo sociologicus*, *emotional man*, *identity keeper*$\}$) maximizes its expected value for all available actions a in the current situation s. If there are several options with the same value, an agent decides randomly among them. This results in a specific message (i.e., Tweet) being selected and published at the simulated social network platform for all other agents to observe. A formal specification of each actor types' expected value functions is presented in Section IV

Using the described value maximization approach to select a message to be communicated leads to a restriction in the amount of behavioral randomness. This is especially useful for evaluating the sensitivity of the resulting emergent effects on the population level to the agent population. Different compositions of agents within a population will lead to different interactions with low variance.

In order to increase the variance of agent behaviors, fluctuating populations can be introduced. Alternatively, a random selection of messages, weighted by their respective expected values, can be introduced. This will then increase the randomness on an individual instead of the population level. However, adding this stochasticity decreases the explanatory impact of modeling social actor types because their respective motivations become less pronounced in the selected communication activities.

B. Actor Type Combinations and Populations

According to the preceding decision-making model, each agent can implement one of the four available actor types. However, these are only prototypical examples for categorizing motivations. In fact, an actor's social disposition will often be more adequately described by a mixture of several basic motivations [23]. Consequently, we allow for combinations of actor types within individual agents to represent that phenomenon.

For mixing several actor types, each agent is defined by four weights w_i, one for each actor type i, with $\sum_i w_i = 1$. The weights denote the ratio with which those types contribute to its decision-making. Then, an agent with mixed types selects its activities by maximizing the weighted sum of the respective expected values (with a randomized selection in case of several maxima).

$$\arg\max_{a} \sum_{i} EV_i(s, a)\ w_i$$

In addition to combining actor types within an individual agent, it is also possible to mix different agents within the overall agent population. That is, a population can either consist of homogeneous agents that all implement the same actor type combination, or it can comprise different agents. Homogeneous populations are particularly useful for model validation and calibration. They make the effects of different value functions easily observable and adjustable. Contrastingly, heterogeneous populations are more realistic. They lead to complex interaction dynamics, which are necessary for replicating and explaining user behaviors in social media as described in the following sections.

IV. Application: Agent-Based Analysis of Social Media Communication

In this section, we apply our agent-based modeling concept to an analysis of user behavior in communication processes on Twitter. In particular, we model live-tweeting behavior during an episode of the German television series "Tatort" (meaning *crime scene*). Running since 1970, "Tatort" is the most popular German TV series, which attracts a broad audience across all social groups, genders, and ages. We use a dataset of Tweets about the episode "Alle meine Jungs" (*all my boys*), of 18 May 2014. The dataset contains eight distinct phases of very high or very low Twitter activity, which correspond to specific scenes of the episode. These scenes provide the situation for the agents in our model to react to. Each of them is described by one or more out of five attributes as shown in Table I.

In our model, the agents can act repeatedly during each scene. At the beginning of a scene, they base their actions only on the respective description; subsequently, they can also react to other agents' Tweets. Thus, a dynamic communication system emerges from these interrelated activities. In the

TABLE I. SITUATION DESCRIPTIONS.

Scene	Description	Scene	Description
0	thrilling	4	funny
1	funny, music-related	5	thrilling, emotional
2	funny, music-related	6	thrilling
3	funny, music-related	7	judgmental

following, we describe the available actions and the decision-making of the four actor types in these situations.

A. Agent Activity Options

The Tweets in our dataset can be classified by their sentiment and tonality along two different dimensions. They are either positive or negative and they are either joking or not joking (i.e., serious). The possible combinations of these categories result in four different message types available to the agents. However, since not all users reply to every message, an agent also has the option not to tweet. Nevertheless, it can still decide to participate in the conversation about the current scene at a later time after observing Tweets by other agents. This results in the following five activity options for the agents.

1) No Tweet
2) Tweet – positive – joking
3) Tweet – positive – not joking
4) Tweet – negative – joking
5) Tweet – negative – not joking

Which of these options an agent selects at which time depends on its underlying combination of actor types, as well as on the activities of other agents as described in the following.

B. Agent Decision-Making

In our application example, the actor types defining the agents' decision-making represent typical behavioral roles and motivations in social media communication. These include the maximization of publicity, a desire for serious discussion, the expression of anger, as well as genuine content production. Motivations are represented by the *homo economicus*, *homo sociologicus*, *emotional man*, and *identity keeper*, respectively. For all actor types, we evaluate the available activity options with respect to those motivations in each situation in order to identify expected values for the agents' decisions. Table II summarizes the criteria and values for that evaluation. Furthermore, for each actor type, a specification of function EV is presented, which is used for the calculation of the expected values [38].

TABLE II. DECISION-MAKING BY SOCIAL ACTOR TYPES.

Homo Economicus	Homo Sociologicus	Emotional Man	Identity Keeper
No Tweet (0)	Must (3)	Unchanged (0)	Strengthened (3)
Utility function	Should (2)	Increased (-1)	Weakened (-1)
(0 to 3)	Can (1)	Decreased (2)	
Conversation size	Should not (-1)	Strongly	
threshold (-1)		decreased (3)	

In social media communication, a *homo economicus* agent attempts to maximize the impact of its contributions on the conversation. Such an agent gains the highest utility by provoking agreement with as many other agents as possible. Thus, its underlying utility function anticipates probable majority opinions. Actions supporting these are rated higher than less popular or even controversial contributions according to the distribution of actions in the original dataset. This agent type will maintain its ratings during an actual conversation regardless of other agents' behaviors.

In addition, we use a threshold of a minimal number of Tweets by other agents for this type of agent to become active itself. This threshold equals to the mean number of Tweets across all scenes (24 in the dataset). Until the threshold is reached, an agent will not participate in the conversation, leaving its utility unchanged. Thus, the *homo economicus* represents a casual media user who only joins ongoing conversations to represent common sense opinions shared by the expected majority of recipients.

The corresponding expected value function depends on the Tweets published in the current scene s so far as given by $tweets_s$. If the overall number of Tweets in $\sum_{a' \in A} tweets_s(a')$ does not exceed the threshold, the *homo economicus* has a value of -1 for all other actions than the no Tweet option. The threshold $\frac{1}{|A|} \sum_{a' \in A} \varphi_s(a')$ is the arithmetic mean of all Tweets throughout the scenes in the entire original data set. Otherwise, the agent selects its actions according to their share in the real world data set given by $\varphi_s(a)$. The prevalent action is yielded by the term $\max_{a' \in A}(\varphi_s(a'))$, which iterates over all possible actions in the respective scene. Moreover, the utility values for a *homo economicus* are normalized and rounded to the nearest natural number between 0 and 3.

$$EV_{HE}(s,a) = \begin{cases} -1 \text{ , if } \sum\limits_{a' \in A} tweets_s(a') < \frac{1}{|A|} \sum\limits_{a' \in A} \varphi_s(a') \\ \left\lfloor 3 \frac{\varphi_s(a)}{\max_{a' \in A}(\varphi_s(a'))} \right\rceil \quad \text{, otherwise} \end{cases}$$

Contrastingly, a *homo sociologicus* agent rates the available actions according to general social norms as well as other agents' behaviors. With respect to the scene description, its expected value function evaluates these options by their perceived strength of obligation. For instance, an agent *should not* joke about an emotional scene. However, if the majority of other agents has deviated from such norms before, the *homo sociologicus* will mimic these previously observed activities in order to gain acceptance by other agents. Hence, that type of agent represents a both morally concerned and opportunistic user who joins the dominant group as soon as one emerges. This behavior is typical, e.g., in massive online protests [4].

The expected value of a *homo sociologicus* agent depends on the norm for the current situation and the predominant action so far. The function $norm(c,a)$ returns a value of -1 for an action it *should not* select, 1 if the agent *can* execute an activity, 2 if it *should* do it, and 3 if it *must* choose the respective action. Table III shows the norms that affect an agent for each attributional category in the current scene description.

$$EV_{HS}(s,a) = \begin{cases} 3 \quad \text{, if a = } \arg max_{a' \epsilon A}(tweets_s(a')) \\ \sum\limits_{c \in s} norm(c,a) \quad \text{, otherwise} \end{cases}$$

$$\text{with } norm: \; C \times A \rightarrow \{-1, 0, 1, 2, 3\}$$

The *emotional man*, on the other hand, represents an outright dissatisfied and angry user. Such an agent strives to

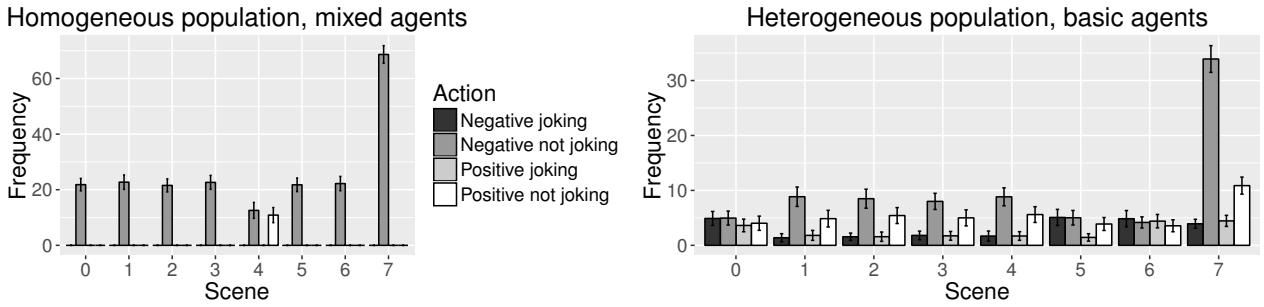


Figure 3. Activity frequencies of a homogeneous population of mixed actors (left) and a heterogeneous population of basic actor types (right).

TABLE III. Values of $anger(c, a)$ and $norm(c, a)$ for categories and actions.

Category $c \in C$	Action $a \in A$	$norm(c, a)$	$anger(c, a)$
thrilling	No Tweet	1	0
	Tweet - positive - joking	2	-1
	Tweet - positive - not joking	2	-1
	Tweet - negative - joking	-1	3
	Tweet - negative - not joking	1	2
funny	No Tweet	1	0
	Tweet - positive - joking	2	-1
	Tweet - positive - not joking	3	-1
	Tweet - negative - joking	-1	2
	Tweet - negative - not joking	-1	3
music-related	No Tweet	1	0
	Tweet - positive - joking	2	-1
	Tweet - positive - not joking	3	-1
	Tweet - negative - joking	-1	2
	Tweet - negative - not joking	1	2
emotional	No Tweet	2	0
	Tweet - positive - joking	-1	-1
	Tweet - positive - not joking	1	-1
	Tweet - negative - joking	-1	0
	Tweet - negative - not joking	1	0
judgmental	No Tweet	-1	0
	Tweet - positive - joking	1	-1
	Tweet - positive - not joking	3	-1
	Tweet - negative - joking	-1	3
	Tweet - negative - not joking	2	2

express that anger. This leads to predominantly negative and sometimes sarcastic (i.e., joking) contributions. By publishing negative Tweets, the agent decreases its anger until it no longer feels the need to communicate. Consequently, that behavior produces isolated criticism without any intention of engaging in an actual discussion.

The expected value for the *emotional man* depends on the output of an $anger$-function. That function evaluates the current attributional categories of the situation description according to their emotional implications for the agent. If an action *decreases* the agent's anger, its expected value is 2. If the agent can even *strongly decrease* it, the value is 3. In case an action would *increase* its anger instead, the $anger$-function returns -1 and if an action does not affect the anger at all, the yielded value is 0. Table III shows the results of the $anger$-function.

$$EV_{EM}(s, a) = \sum_{c \in s} anger(c, a),$$

$$\text{with} \quad anger: \ C \times A \rightarrow \{-1, 0, 1, 2, 3\}$$

Finally, the *identity keeper* is a genuine content producer. This type of agent has the goal of bringing forward any kind of discussion in order to maintain its participation in it. That is, the agent can strengthen its identity by providing arguments for other agents to react to. For that purpose, any kind of Tweet can be appropriate, especially controversial ones if they provoke reactions. Only remaining inactive weakens that identity. As a result, the *identity keeper* represents a user who enjoys a conversation for the sake of the conversation and who ensures a certain diversity of perspectives on the discussed topic. Thus, the expected value for the *identity keeper* is defined as follows:

$$EV_{IK}(s, a) = \begin{cases} -1 \text{ , if a } = \text{no tweet} \\ \\ 3 \text{ , otherwise} \end{cases}$$

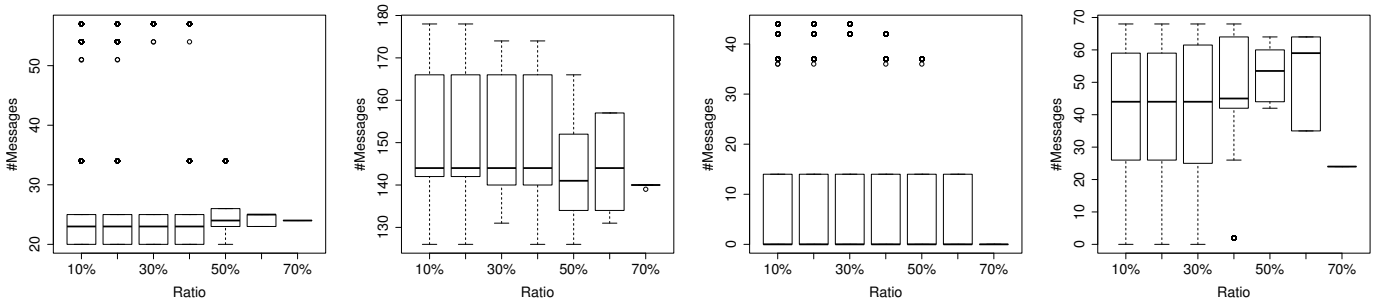
By combining the described actor type models within individual agents, it is possible to represent mixed motivations and to implement a wide variety of decision behaviors. Moreover, heterogeneous populations of different agent types will lead to complex interactions of these behaviors. The following section evaluates these effects.

V. Evaluation: Simulation of User Behavior

As a proof of concept for our agent-based modeling approach, we have implemented the aforementioned agent types and decision-making algorithms in a *JAVA* program. In the following, we use that program to simulate user behaviors emerging from different populations of various agents. Such a simulation gives a first impression of the range of effects that the model can (re)-produce. In particular, it allows for analyzing the interplay between several actor types on both the individual and the population level. In addition, the results of a sensitivity analysis are presented, which provides detailed insights in how variations of the input parameters affect the behavior of the agents in the model.

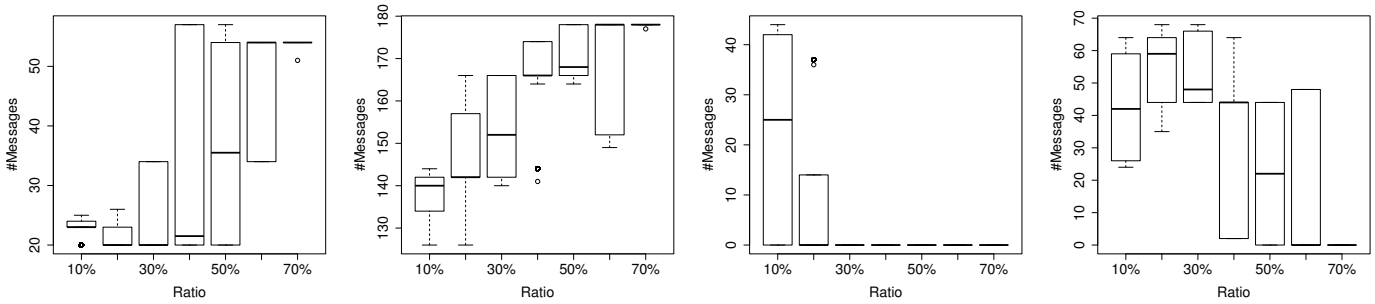
In our simulation, we compare two different settings. The first one consists of a homogeneous agent population with mixed actor types. That is, each agent combines all four types with equal weights. By contrast, the second setting comprises a heterogeneous agent population in which every agent implements one of the four basic actor types. Throughout the population, these agents are uniformly distributed. They communicate about all eight scenes. Their respective activity choices depend on the situation description for those scenes as well as on the previous actions of other agents.

For both settings, the population size is set to 164 agents that can join the conversation in each scene (as in the real world dataset). This number is relevant as long as the *homo economicus* uses a fixed conversation size threshold. The more agents there are, the sooner will a *homo economicus* impact the communication dynamics. While the threshold can be scaled



(a) positive joking (b) positive not joking (c) negative joking (d) negative not joking

Figure 4. Homo Economicus



(a) positive joking (b) positive not joking (c) negative joking (d) negative not joking

Figure 5. Homo Sociologicus

up or down according to the population size, we use the realistic one to enable comparisons of our simulation results with that dataset in future studies.

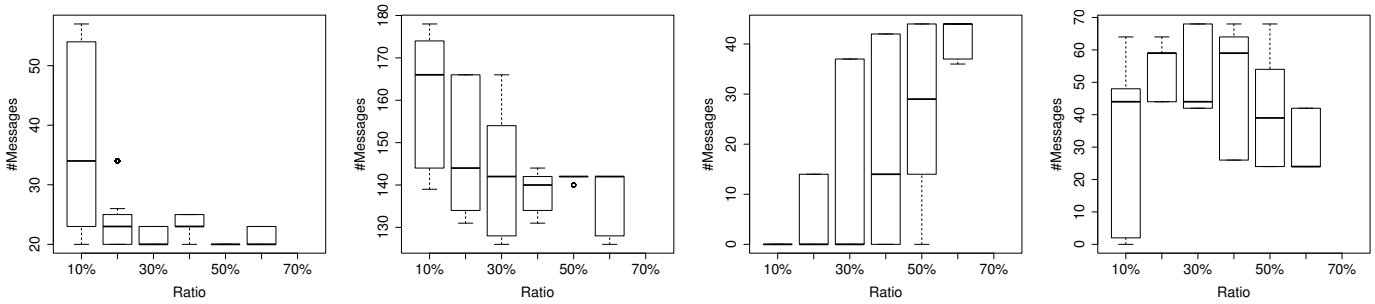
Figure 3 shows the arithmetic mean of our evaluation results together with the respective standard deviations out of 100 simulation repetitions (except for the "No Tweet" option). For the homogeneous population, the results show a majority of negative not joking Tweets. This is due to the fact that both *identity keeper* and *homo economicus* consider this activity as adequate. Moreover, the *emotional man* favors it over all others. Combining these within the agents leads to the observed uniformity, which even becomes amplified as the *homo sociologicus* imitates dominant behaviors. Only scene 4 leads to negative as well as positive Tweets. That scene is described as being funny. Hence, the positive actions favored by all other actor types override the negative option selected by the *emotional man*.

Contrastingly, the heterogeneous population leads to more diverse behavior. In that case, negative Tweets are still prevalent for most scenes. This is caused by the same effects as described: The *homo sociologicus* amplifies the behavior being initially driven by the other actor types, particularly the *emotional man*. However, since these agent types act simultaneously in a mixed population, all other actions are also observable. This leads to realistic effects, such as decisions not to tweet at all in scenes being described as thrilling.

After analyzing the interplay between different actor types, the behavior of each individual actor type is evaluated with respect to the identification of potential factors that influence its behavior. To investigate how different inputs of the simulation model influence the observed communication behavior of the actors, sensitivity analysis can be applied [39]. For this purpose, different parameter configurations must be simulated such that the influence of variations can be observed. In the presented simulation model, three different types of actors can be specified as parametrization of the model, which are then used to generate the population. The character of each actor is defined as a combination of the four presented motivations, i.e., *homo economicus*, *homo sociologicus*, *emotional man*, and *identity keeper*. As the sum of all motivations must be equal to 100%, each agent must embody all four motivations, and as motivations can be defined in 10% steps, a total number of 79 different agent specifications is possible. To investigate how different motivations influence the communication dynamics of the actors, only scenarios that consist of a single agent type are considered. By this means, interactions between different agent types can be avoided and unbiased results are generated.

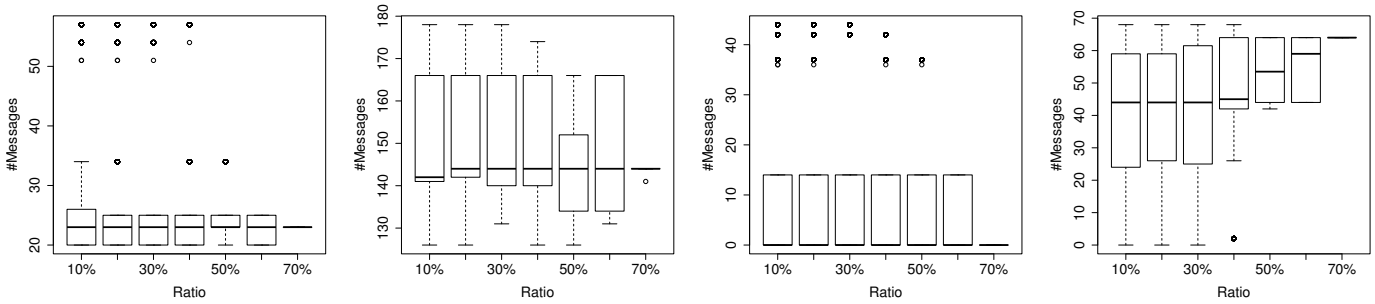
The results of the conducted sensitivity analysis are presented in Figure 4 - 7. Separated by the four presented motivations and for each possible message type, the number of tweeted messages is presented. To illustrate the distribution of the results, boxplots are used, which visualize the median value (bold line) as well as the range in which the middle 50% of the observed data fall (box). On each side of the box, antennas visualize 1,5 times the interquartile range and outliers that fall outside this range are marked as circles.

The behavior of *homo economicus*, a casual media user, can be described as balanced compared to the behavior observed from agents with other motivations. The number of joking tweets, regardless whether they are positive or negative, is



(a) positive joking (b) positive not joking (c) negative joking (d) negative not joking

Figure 6. Emotional Man



(a) positive joking (b) positive not joking (c) negative joking (d) negative not joking

Figure 7. Identity Keeper

considerably lower than the number of not joking tweets (median of number of joking tweets <25 compared to median of number of not joking tweets >40). Compared to the three other motivations, the number of tweets sent by *homo economicus* remains steady when the ratio of this motivation increases from 10% to 70%. Only the number of negative not joking tweets increases (approx. 25%) as the ratio of *homo economicus* exceeds 50% and the number of positive not joking tweets decreases slightly (approx. 5%).

In contrast to *homo economicus*, the impact of the ratio of *homo sociologicus* on the number of tweeted messages is much higher. For positive not joking messages, a total increase of more than 25% can be observed while the ratio increases from 10% to 70%. Also for negative not joking tweets, an impact of the ratio of *homo sociologicus* can be observed. Yet, after an initial increase of the number of tweets, a decrease occurs as the ratio exceeds 30%. Considering the number of joking messages, an even stronger influence of the ratio of *homo sociologicus* can be identified. For positive joking messages, a total increase of almost 300% and for negative joking messages a total decrease from 25 to 0 messages occurs. Summarizing it can be said that an increasing ratio of *homo sociologicus* results in an increase of the number of positive tweets, both joking and not joking, in favor of a decrease of the number of negative tweets.

The communication dynamics observed as the ratio of *emotional man* increases are diverse. An increasing ratio results in a decreasing number of positive joking tweets (-40%). The amount of positive not joking tweets decreases as the ratio increases from 10% to 20%. From this ratio onward to a ratio of 70%, no considerable decrease can be observed. The number of negative tweets, in contrast, increases as the ratio of *emotional man* increases. While an increasing number of negative joking tweets can be observed for a ratio between 40% and 60%, no steady behavior can be observed for the number of negative not joking tweets. Instead, the number of messages increases first and then decreases again.

Finally, the influence of the *identity keeper* is similar to the one of *homo economicus*. Considering the amount of positive joking, positive not joking, and negative joking messages, neither an increase not a decrease can be observed as the ratio increases. Only the number of negative not joking tweets increases slightly (<25%) as the ratio increases from 50% to 70%.

Summarizing, both *homo economicus* and *identity keeper* positively affect the number of negative not joking tweets. Further impacts that can be identified are an increasing number of positive joking tweets for a high ratio of *homo sociologicus* as well as a deceasing number of positive joking tweets for a low ratio of *emotional man*. The amount of positive not joking tweets as well as of negative not joking tweets can mostly be attributed to *homo sociologicus* as well. This observation is supplemented by the positive impact high ratios of *homo economicus* and *identity keeper* have on the number of negative not joking messages. Overall, the observed behavior corresponds to the sociological theories presented in Section II-C. *Homo economicus* behaved efficiently and did not write more messages as its ratio increased. The norms of *homo sociologicus* mainly lead to positive messages, however, also some negative not joking messages are sent which can be

attributed to the judgmental parts of the television program. The behavior of the *emotional man* is a result of emotional implications, which explains the increasing amount of negative messages as the ratio increases. Finally, the communication behavior of the *identity keeper* is not related to the content and rather to social aspects. Hence, messages sent by this actor type are mainly not joking.

Overall, these results presented in this evaluation show that the combination of actor types both within individual agents and their mixing in heterogeneous populations drastically impacts the emergent dynamics of simulated social media communication. This demonstrates that modeling motivations of individual agents can produce behavioral heterogeneity, which other models have to introduce artificially, e.g., by means of random noise [34]. In contrast to those approaches, we can directly control, which type of agent reacts to which particular communicative situation in what manner. The composition of an agent population then models the affinity or aversion of a user group in social media to certain topics, opinions, and communication styles. Hence, we conclude that our model adds this composition of populations as an important variable to existing methods for studying information diffusion in social simulations.

However, it is important to select and calibrate the agent types carefully for such a simulation to yield meaningful results for understanding user behavior. To that end, it is necessary to analyze available real world data and identify typical activity patterns [4]. Then, potential underlying motivations can be derived from those observations in order to define the required actor types and their combinations [7] [20]. With this work, we have shown how such actor types can be modeled for exploring user behavior in agent-based social simulations.

VI. Conclusions and Future Work

In this paper, we have developed an agent-based model of user behavior in social media. This model facilitates dynamic analyses of complex communication processes, which are difficult to assess by means of conventional approaches. In such a context, agent-based social simulations allow for experimentally exploring emergent behaviors [4].

Our model focuses primarily on the decision-making of social actors communicating about a specific topic. This is in contrast to existing work on information diffusion, which analyzes the impacts of social network structures on the spread of messages. Instead, we have modeled motivational causes for user behaviors by utilizing complex agents based on sociological theory. To that end, we have presented a general concept for representing and combining four different actor types in agent-based social simulations.

In addition, we have applied this concept to model and analyze Twitter communication about a German television program. Our evaluation shows that particular combinations of different motivations either within individual agents or across an entire population drastically impact communication dynamics. Therefore, we conclude that it is crucial to consider these motivations carefully in order to realistically model and explain user behavior in social media. Furthermore, the impact different actor types have on the communication behavior were shown by means of a sensitivity analysis. The observed dynamics correspond to the sociological descriptions of the respective actor types.

While our model provides a promising first step to agent-based simulations of social media usage, we consider both extensions and applications of the model for future work. Firstly, we are working on calibrating the model to accurately imitate the user interactions observed in our real world example. This will provide insight into the achievable realism when combining the four basic actor types into complex agents and populations. As a first result, we have already demonstrated that our model is indeed capable of reproducing the communication activities of real world users [38].

Secondly, it would be interesting to integrate the agent decision method with existing information diffusion approaches [28]. This will complement those methods with motivational aspects of *why* information is spread within a social network. In that context, the population composition will provide an additional variable, which impacts communication dynamics. The various actor types can then produce behavioral heterogeneity on a more detailed and explanatory deeper level than the addition of abstract random noise to an equational modeling approach [34].

Finally, it will also be necessary to model the activity options for the agents in more detail. This covers particularly the message contents. In order to simulate, e.g., the shaping of opinions in political discourses, a classification of communication contents and their impact on the interaction is required. To achieve this, we plan to utilize content modeling and annotation techniques from media and communication studies [40] for encoding discourses in agent-based social simulations.

In addition to extending the existing model, we plan to utilize the model as part of comprehensive simulation studies. By this means, research hypotheses regarding the behavior of the model can be answered systematically and in a methodologically sound way [41]. Potential hypotheses on the model's behavior that can be investigated in such a way include the influence actor types have on the communication behavior of other actor types or the identification of actions that influence communication dynamics in a specific way.

Acknowledgments

We would like to thank Carla Schmidt, Christof Barth, and Hans-Jürgen Bucher for providing us with real world data and a media studies perspective on our application example.

References

[1] J. O. Berndt, S. C. Rodermund, F. Lorig, and I. J. Timm, "Modeling user behavior in social media with complex agents," in Third International Conference on Human and Social Analytics (HUSO 2017), 2017, pp. 18–24.

[2] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network." Theory of Computing, vol. 11, no. 4, 2015, pp. 105–147.

[3] J. O. Berndt, F. Lorig, I. J. Timm, C. Barth, and H.-J. Bucher, "A systematic approach to agent-based dynamic analysis of social media communication," International Journal On Advances in Internet Technology, vol. 10, no. 1&2, to appear 2017.

[4] I. J. Timm, J. O. Berndt, F. Lorig, C. Barth, and H.-J. Bucher, "Dynamic analysis of communication processes using twitter data," in 2nd International Conference on Human and Social Analytics (HUSO 2016). IARIA, 2016, pp. 14–22.

[5] P. Davidsson, "Agent based social simulation: A computer science view," Journal of Artificial Societies and Social Simulation, vol. 5, no. 1, 2002.

[6] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena." in 5th International AAAI Conference on Weblogs and Social Media, 2011, pp. 450–453.

[7] F. Lorig and I. J. Timm, "How to model the human factor for agent-based simulation in social media analysis?" in 2014 ADS Symposium (part of SpringSim Multiconference). SCS, 2014, p. 12.

[8] F. Vega-Redondo, Complex Social Networks. Cambridge University Press Cambridge, MA, 2007.

[9] L. C. Freeman, "Centrality in social networks conceptual clarification," Social networks, vol. 1, no. 3, 1978, pp. 215–239.

[10] C. R. Berger, "Interpersonal communication," The International Encyclopedia of Communication, 2008.

[11] C. E. Shannon, "A mathematical theory of communication," ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, 2001, pp. 3–55.

[12] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in 43rd Hawaii International Conference on System Sciences (HICSS). IEEE, 2010, pp. 1–10.

[13] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks." ICWSM, vol. 10, 2010, pp. 90–97.

[14] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," PloS one, vol. 6, no. 5, 2011, p. e19467.

[15] A. Maireder and S. Schlögl, "24 hours of an #outcry: The networked publics of a socio-political debate," European Journal of Communication, 2014, pp. 1–16.

[16] N. Luhmann, Social Systems. Stanford, USA: Stanford University Press, 1995.

[17] P. Hedström and P. Ylikoski, "Causal mechanisms in the social sciences," Annual Review of Sociology, vol. 36, 2010, pp. 49–67.

[18] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in 2nd International Workshop on Search and Mining User-generated Contents. ACM, 2010, pp. 37–44.

[19] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM, 2007, pp. 56–65.

[20] S. Vandenhoven and O. De Clercq, "What does the bird say? exploring the link between personality and language use in dutch tweets," in 2nd International Conference on Human and Social Analytics (HUSO 2016). IARIA, 2016, pp. 38–42.

[21] S. Schirra, H. Sun, and F. Bentley, "Together alone: motivations for live-tweeting a television series," in 32nd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2014, pp. 2441–2450.

[22] V. Hoste, C. Van Hee, and K. Poels, "Towards a framework for the automatic detection of crisis emotions on social media : a corpus analysis of the tweets posted after the crash of germanwings flight 9525," in 2nd International Conference on Human and Social Analytics (HUSO 2016), 2016, pp. 29–32.

[23] P. Dittrich and T. Kron, "Complex reflexive agents as models of social actors," in Proceedings of the SICE Workshop on Artificial Society/Organization/Economy, ser. Meeting of Systems Engineering, vol. 25, Tokyo, 2002, pp. 79–88.

[24] H. Flam, "Emotionalman': I. the emotionalman'and the problem of collective action," International Sociology, vol. 5, no. 1, 1990, pp. 39–56.

[25] M. Wooldridge and N. R. Jennings, "Agent theories, architectures, and languages: a survey," in International Workshop on Agent Theories, Architectures, and Languages. Springer, 1994, pp. 1–39.

[26] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," Proceedings of the National Academy of Sciences, vol. 99, no. suppl 3, 2002, pp. 7280–7287.

[27] J. O. Berndt and O. Herzog, "Anticipatory behavior of software agents in self-organizing negotiations," in Anticipation Across Disciplines. Springer, 2016, pp. 231–253.

[28] C. Zhang, J. Sun, and K. Wang, "Information propagation in microblog networks," in 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, 2013, pp. 190–196.

[29] S. Monica and F. Bergenti, "A stochastic model of self-stabilizing cellular automata for consensus formation," in 15th Workshop Dagli Oggetti agli Agenti (WOA 2014), Catania, 2014.

[30] A. Tsang and K. Larson, "Opinion dynamics of skeptical agents," in International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2014), 2014, pp. 277–284.

[31] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009, pp. 199–208.

[32] S. Monica and F. Bergenti, "Opinion dynamics in multi-agent systems: selected analytic models and verifying simulations," Computational and Mathematical Organization Theory, 2016, pp. 1–28.

[33] F. Schweitzer and J. A. Hołyst, "Modelling collective opinion formation by means of active brownian particles," The European Physical Journal B-Condensed Matter and Complex Systems, vol. 15, no. 4, 2000, pp. 723–732.

[34] S. Monica and F. Bergenti, "An analytic study of opinion dynamics in multi-agent systems with additive random noise," in AI* IA 2016 Advances in Artificial Intelligence. Springer, 2016, pp. 105–117.

[35] T. Balke and N. Gilbert, "How do agents make decisions? a survey," Journal of Artificial Societies and Social Simulation, vol. 17, no. 4, 2014.

[36] K. Fischer, M. Florian, and T. Malsch, Socionics: scalability of complex social systems. Berlin: Springer, 2005.

[37] R. Belkaroui, R. Faiz, and A. Elkhlifi, "Conversation analysis on social networking sites," in 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems (SITIS). IEEE, 2014, pp. 172–178.

[38] S. C. Rodermund, F. Lorig, J. O. Berndt, and I. J. Timm, "An agent architecture for simulating communication dynamics in social media," in Proceedings of the 15th German Conference on Multiagent System Technologies, MATES, Leipzig, Germany, 2017, pp. 19–37.

[39] J. P. Kleijnen, "An overview of the design and analysis of simulation experiments for sensitivity analysis," European Journal of Operational Research, vol. 164, no. 2, 2005, pp. 287–300.

[40] K. A. Neuendorf, The Content Analysis Guidebook. Sage, 2016.

[41] F. Lorig, D. S. Lebherz, J. O. Berndt, and I. J. Timm, "Hypothesis-driven experiment design in computer simulation studies," in Proceedings of the 2017 Winter Simulation Conference. Las Vegas, NV, USA: IEEE, Dec. 2017, pp. 1360–1371.

An Investigation of Users' Actions Expressed in Tweets Submitted by Using Music Player Applications

Yasuhiko Watanabe, Kenji Yasuda, Ryo Nishimura, and Yoshihiro Okada
Ryukoku University
Seta, Otsu, Shiga, Japan
Email: watanabe@rins.ryukoku.ac.jp, t130522@mail.ryukoku.ac.jp,
r_nishimura@afc.ryukoku.ac.jp, okada@rins.ryukoku.ac.jp

***Abstract*—What users are doing at a certain point in time is important for designing various services and applications in social media, such as targeted advertisement, news recommendation, and real-world analysis. As a result, in this study, we investigated tweets which users submitted when they were listening to music by using music player applications. We collected 2,000 tweets including hashtags generated by music player applications and investigated what users described in these tweets. We found 10 % of them were tweets where actions while listening to music were described. We applied machine learning techniques to detect tweets where two kinds of actions while listening to music, moving to somewhere or going to bed, were described. Furthermore, we examined whether we can detect tweets where two kinds of action phases, start and middle, were described. In both cases, we obtained the high accuracy and precision. The experimental result shows that our method is useful for providing behavior based services and applications in social media.**

Keywords–music player application; music content; behavior based service; Twitter; social media.

I. INTRODUCTION

Social media, such as Twitter and Facebook, generate large quantities of data about where users are and what they are thinking or doing at a certain point in time. Take tweets on Twitter, (exp 1) and (exp 2), for example. We can understand the submitters of these two tweets were listening to music. This is because #nowplaying in (exp 1) and (exp 2) show that these tweets were submitted by using music player applications. Users who are using music player applications are thought to be listening to music.

(exp 1) *#nowplaying: "soundscape" from "soundscape - Single" by TRUE (saisei kaisuu: 35) #songsinfo*
(#nowplaying: "soundscape" from "soundscape - Single" by TRUE (plays: 35) #songsinfo)

(exp 2) *#nowplaying kagerou by ONE OK ROCK on #onkyo #hfplayer*

#nowplaying is a hashtag generated by various music player applications. Furthermore, #songsinfo in (exp 1) is a hashtag generated by a music player application, SongsInfo. Also, #onkyo and #hfplayer in (exp 2) are hashtags generated by a music player application, HF Player. These hashtags and the other words in (exp 1) and (exp 2) were all generated and embedded into these tweets automatically by music player applications when users submitted these tweets by using them. As a result, these hashtags enable us to understand that these users were listening to music when they submitted these tweets by using music player applications. As mentioned, (exp 1) and (exp 2) consist of words and hashtags all of which were generated by music player applications. On the other hand, (exp 3), (exp 4), and (exp 5) include words generated not only by music player applications but by users.

(exp 3) *#nowplaying: "Grow Slowly" from "Hafa Adai" by iguchi yuka (saisei kaisuu: 3) #songsinfo suki desu motto kiiteiru*
(#nowplaying: "Grow Slowly" from "Hafa Adai" by Iguchi Yuka (plays: 3) #songsinfo I like and listen to it so many times)

(exp 4) *basu wo nogashita node aruki masu !!#nowplaying: "walk on Believer ♪" from "walk on Believer ♪" by toyosaki aki (saisei kaisuu: 96) #songsinfo*
(I will walk because I missed the bus !! #nowplaying: "walk on Believer ♪" from "walk on Believer ♪" by toyosaki aki (plays: 96) #songsinfo)

(exp 5) *tenshon age te yakin ikuzo #nowplaying NIGHT FLIGHT by Perfume on #onkyo #hfplayer*
(I cheer myself up and go to night shift #nowplaying NIGHT FLIGHT by Perfume on #onkyo #hfplayer)

Specifically, the following words in (exp 3), (exp 4), and (exp 5) were generated not by music player applications but by users.

- *suki desu motto kiiteiru* (I like and listen to it so many times) in (exp 3),
- *basu wo nogashita node aruki masu !!* (I will walk because I missed the bus !!) in (exp 4), and
- *tenshon age te yakin ikuzo* (I cheer myself up and go to night shift) in (exp 5)

In this study, we describe user generated words in tweets submitted by using music player applications as *comments*. We will explain comments in tweets submitted by using music player applications in Section III. The comments in (exp 3), (exp 4), and (exp 5) express user's impression, action, and reason, respectively.

We can know that the submitters of (exp 3), (exp 4), and (exp 5) were listening to music when they submitted these tweets into Twitter. Furthermore, comments in these tweets enable us to understand what they were thinking and doing while listening to music. What users are thinking and doing at a certain point in time is important for designing various services and applications on social media, such as targeted advertisement, news recommendation, and real-world analysis. As a result, we investigated tweets submitted by using music

player applications and show what Twitter users are thinking and doing while listening to music [1]. In this paper, we conduct a detailed investigation on tweets submitted by using music player applications and discuss whether they can be classified by using machine learning techniques.

The rest of this paper is organized as follows: In Section II, we survey the related works. In Section III, we investigate tweets submitted by music player applications and show what users are thinking and doing while listening to music. In Section IV, we apply machine learning techniques to classify tweets submitted by music player applications and discuss whether we can detect what users are doing and what action phases they are in while listening to music. Finally, in Section V, we present our conclusions.

II. Related works

Twitter enables us to easily submit short messages in real time from anywhere with internet access. As a result, Twitter data is a valuable resource for predicting various trends and events. Taking this in consideration, there are many studies that have treated Twitter as a social sensor [2]. Aramaki et al. reported that Twitter messages reflect the real world and influenza related tweets can be extracted by using Twitter API and NLP techniques [3]. Also, Culotta showed that influenza-related Twitter messages can be identified by using a document classification method and a small number of flu-related keywords can forecast future influenza rates [4]. Sakaki et al. investigated the real-time nature of Twitter and proposed an event notification system that monitors tweets and delivers notification promptly [5]. Jansen et al. reported that microblogging is an online tool for customer word of mouth communications and potentially rich for companies to explore as part of their overall branding strategy [6]. Furthermore, Twitter data was used for inferring on-line Internet service availability [7], measuring public interest and concern about health-related events [8], observing information diffusion in social media [9], and examining situational features during emergency events [10].

Timestamps and geotags embedded into tweets are useful for treating Twitter as a social sensor. Some researchers conducted studies for event detection using geotags embedded into tweets. Lee and Sumiya proposed a method for detecting local events by applying a k-means clustering method to geotagged Twitter documents [11]. Kamath et al. studied the spatio-temporal dynamics of Twitter hashtags by using a sample of 2 billion geo-tagged tweets [12]. However, Watanabe et al. reported that less than one percent of Twitter posts are associated with a geolocation [13]. This is because Twitter users have been slow to adopt geospatial features and only a small amount of tweets comes with location information [14]. As a result, recent work has focused on geoinference for predicting the locations of posts. Yamaguchi et al. pointed out that most existing methods can be categorized into two kinds of approaches: a content-based approach or a graph-based approach [15].

First, we discuss studies based on the content-based approach. The content-based approach leverages user-generated contents in the form of texts. Cheng at al. proposed a method for estimating a Twitter user's city-level location based purely on the content of the user's tweets [14]. Eisenstein et al. proposed a method of multi-level generative model that enables prediction of an author's geographic location from tweets [16]. Hecht et al. reported that user's home country and state can be reasonably inferred by using simple machine learning techniques [17]. Han et al. proposed a method of finding location indicative words via feature selection and examined whether the reduced feature set boosts geolocation accuracy [18]. Schulz et al. proposed a multi-indicator approach for determining the location where a tweet was created and the location of the user's residence [19]. Yamaguchi et al. proposed an online location inference method that can update inference results using only newly arriving contents without using previous contents [15].

Next, we discuss studies based on the graph-based approach. The graph-based approach is based on the structure of social graphs where friends are connected. This approach is based on an idea: users' social networks are useful for revealing their locations. For example, Twitter users are more likely to follow others that are geographically closer to them. As a result, Rout et al. described this approach as network-based approach [20]. Wang et al. used communication records of 6 million mobile phone subscribers and found that the similarity between individuals' movements, their social connectedness and the strength of interactions between them are strongly correlated with each other [21]. Backstrom et al. pointed out that, by using user-supplied address data and the network of associations between members of the Facebook social network, we can directly observe and measure the relationship between geography and friendship [22]. Rout et al. proposed an approach to geolocating users of online social networks, based solely on their friendship connections [20]. Sadilek et al. reported that we can infer people's fine-grained location, even when they keep their data private and we can only access the location of their friends [23].

Kinsella et al. pointed out that understanding where users are can enable a variety of services that allow us to present information, recommend businesses and services, and place advertisements that are relevant to where they are [24]. We also may say that understanding what users are thinking and doing can enable a variety of services that are relevant to what they are thinking and doing. However, few studies have been made on predicting what users are thinking and doing while many studies have been made on predicting where users are. As a result, in this paper, we investigate tweets submitted by using music player applications and show what Twitter users are thinking and doing while listening to music. Furthermore, we discuss whether tweets submitted by using music player applications can be classified by using machine learning techniques.

III. Investigation of tweets submitted by using music player applications

In this section, we investigate tweets submitted by music player applications and show what the users are thinking and doing while listening to music.

A. The investigation object

Tweets can be classified into three types [25]:

- reply
  A reply is submitted to a particular person. It contains "@username" in the body of the tweet. For example, (exp 6) is a reply to @eitaso.

(exp 6) *@eitaso ore to nagoya de seigi no uta wo utawanaika ? (^L^) #nowplaying futten toppa ☆ LOVE IS POWER ☆ / chikyu bouei bu*
(@eitaso Let's sing a song of justice in Nagoya? (^L^) #nowplaying futten toppa ☆ LOVE IS POWER ☆ / chikyu bouei bu)

- retweet
  A retweet is a reply to a tweet that includes the original tweet.
- normal tweet
  A normal tweet is neither reply nor retweet. For example, (exp 3), (exp 4), and (exp 5) are normal tweets. Normal tweets are generally submitted to general public.

In order to investigate tweets submitted by music player applications and what the users are thinking and doing while listening to music, we collected the following 2000 tweets:

- 1,000 Japanese normal tweets including hashtag [26]
  - #nowplaying
  - #songsinfo

  obtained from 13 October 2016 to 11 December 2016. These 1,000 tweets were submitted by 244 users.
- 1,000 Japanese normal tweets including hashtag
  - #nowplaying
  - #onkyo
  - #hfplayer

  obtained from 13 October 2016 to 1 December 2016. These 1,000 tweets were submitted by 345 users.

We did not collect the following tweets even if they include the hashtags above: replies, retweets, and tweets that include no comments generated by users. As a result, (exp 1), (exp 2), and (exp 6) were not included in the collected 2000 tweets. Then, we extracted user generated comments from them by eliminating the following words.

- Uniform Resource Locators (URL),
- hashtags, and
- words generated automatically by music player applications.

As a result, we extracted *suki desu motto kiiteiru* (I like and listen to it so many times) from (exp 3) as a user generated comment. Also, we extracted *basu wo nogashita node aruki masu !!* (I will walk because I missed the bus !!) and *tenshon age te yakin ikuzo* (I cheer myself up and go to night shift) from (exp 4) and (exp 5), respectively.

B. Tweets which users submit when they use music player applications

We classified comments in tweets submitted by using music player applications into the following four types:

impressions comments expressing users' impressions and evaluations of contents which they played by using music player applications,
reasons comments expressing reasons why users played contents by using music player applications,

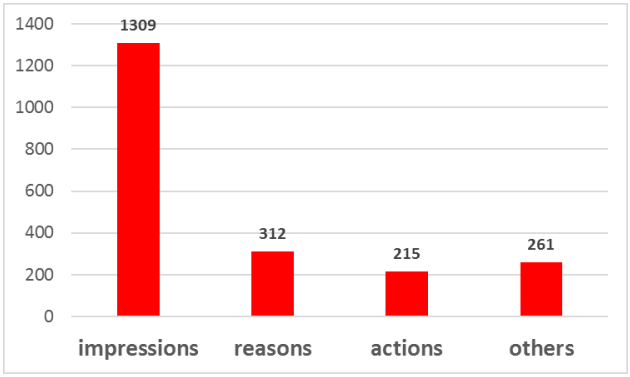


Figure 1. The classification result of the 2,000 tweets which users submit when they use music player applications (by human experts).

actions comments expressing actions which users carried out when they used music player applications, and
others comments that cannot be classified into the three types above.

Figure 1 shows the classification result of the obtained 2,000 Japanese tweets. We should notice that some comments can be classified into two types. For example, *yoi kyoku da!* (Good music!) in (exp 7) is classified into impressions. On the other hand, *ekurea katte kaero!* (Let's buy an eclair and go home!) is classified into actions.

(exp 7) *yoi kyoku da! ekurea katte kaero!*
(Good music! Let's buy an eclair and go home!)

We shall discuss the following kinds of comments in detail.

- comments expressing impressions,
- comments expressing reasons, and
- comments expressing actions.

1) Comments expressing impressions: We found many comments expressing users' impressions and evaluations of contents which they played by using music player applications. Figure 1 shows that more than half of the obtained 2000 tweets were classified into ones expressing users' impressions, such as (exp 8) and (exp 9).

(exp 8) *yoi. suki.*
(Good. I like it.)
(exp 9) *natsukashi sugi te naki sou*
(I was close to tears)

In addition, we found that many comments expressing users' impressions were related to time, such as (exp 10) and (exp 11).

(exp 10) *kono jikantai ni kiku jazz ha, honto ni kimochi ga ii.*
(It's fun listening to jazz in this time period.)
(exp 11) *shinya no Neptunus ha kakubetsu.*
(It is wonderful to listen to Neptunus very late at night.)

Especially, most of them were related to time periods when users played music by using music player applications.

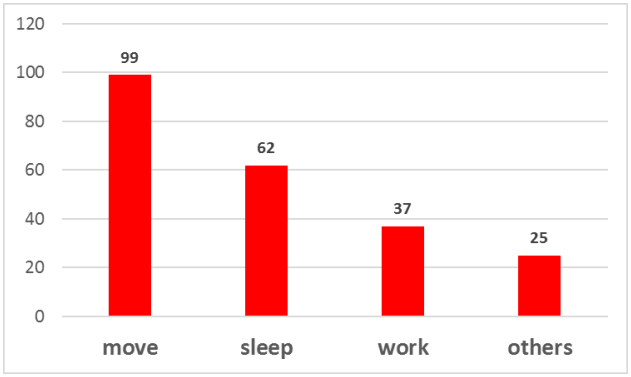


Figure 2. The classification result of the 215 tweets expressing users' actions (by human experts).

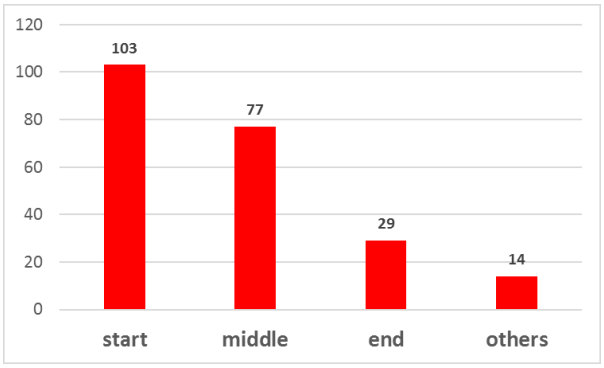


Figure 3. The classification result of the stages of users' actions in the 215 tweets expressing users' actions (by human experts).

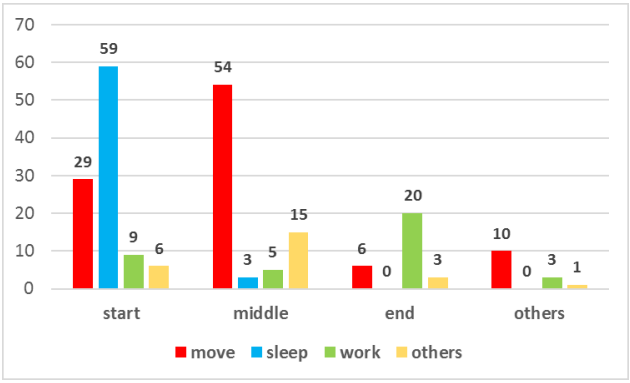


Figure 4. The classification result of the stages of users' actions: move, sleep, work, and others (by human experts).

2) Comments expressing actions: Psychology research has shown that people can attend only one task at a time [27]. Hyman et al. reported that people talking on their cell phones while walking they ran into people more often, and did not notice what was around them [28]. However, listening to music is an exception. We often do something while listening to music. Actually, we found many tweets where users described their actions while using music player applications. (exp 12), (exp 13), (exp 14), and (exp 15) are examples of comments expressing users' actions.

(exp 12) *tsuukin chu.. sawayakana hare.*
(On my way to work.. It's a crisp day.)
(exp 13) *oyasumi nasai*
(Good night)
(exp 14) *desaki deno gyomu shuryo. kiro he. yokohama live no set list.*
(I have finished my business out of the office. On my way home. The set list of the Yokohama live.)
(exp 15) *italo pop kiki nagara kare- shikomu yo*
(I will make curry with listening to Italo pop)

In our investigation, three kinds of most commonly actions described in tweets submitted by using music player applications are move, sleep, and work. For example, (exp 12) shows that the submitter was going to work with listening to music. (exp 13) shows that the submitter was going to sleep, and (exp 14) shows that the submitter had finished the job. As shown in Figure 1, we found 215 tweets expressing users' actions in the obtained 2,000 tweets which users submit when they use music player applications. We classified these 215 tweets expressing actions into four types: move, sleep, work, and others. Figure 2 shows the classification result of the tweets expressing users' actions. We found some tweets expressing users' actions can be classified into two types. For example, (exp 14) was classified into work and move. In particular, user's action expressed in *desaki deno gyomu shuryo* (I have finished my business out of the office) of (exp 14) was classified into work. On the other hand, user's action expressed in *kiro he* (On my way home) of (exp 14) was classified into move. Furthermore, some tweets expressing users' actions were classified into others. This is because they were classified into neither move, sleep, nor work. For example, (exp 15) was classified into others. As shown in Figure 2, many tweets expressing users' actions were classified into move and sleep. Hamamura and Iwamiya conducted the survey on the use of portable music player [29]. The survey was conducted on 72 college students. The result of their survey had partially in common with ours. In their investigation result, 65 students and 39 students of them used portable music players while moving and working, respectively. This investigation result is in good agreement with ours. On the other hand, in their investigation result, there were no students who used portable music players while sleeping. The result is not in good agreement with ours. Furthermore, Hamamura and Iwamiya reported that 19 students used portable music players while shopping. On the other hand, we found only one comment, (exp 16), submitted by a user who were shopping while listening to music.

(exp 16) *osanpo & okaimono !*
(walk & shopping !)

Many tweets expressing users' actions showed the phases of their actions. For example, (exp 13) showed that the tweet was submitted just before the user started his/her action. On the other hand, (exp 12) showed that the tweet was submitted when user's action was ongoing. As a result, we classified the 215 tweets expressing users' actions in Figure 1 into four types: start, middle, end, and others. Figure 3 shows the classification result of the tweets expressing users' actions. As shown in Figure 3, many tweets were classified into user's phase (start) and phase (middle). Some tweets expressing users' actions can be classified into two types. For example, (exp 14) was classified into user's phase (end) and phase (start).

In particular, user's action expressed in *desaki deno gyomu shuryo* (I have finished my business out of the office) of (exp 14) was classified into user's phase (end). On the other hand, user's action expressed in *kiro he* (On my way home) of (exp 14) was classified into user's phase (start). Furthermore, Figure 4 shows the classification result of the phases of user's actions expressed in 215 tweets (Figure 2), move, sleep, work, and others. Figure 4 shows that

- there were many tweets expressing users' action (sleep) in the tweets classified into user's phase (start),
- there were many tweets expressing users' action (move) in the tweets classified into user's phase (middle), and
- there were many tweets expressing users' action (work) in the tweets classified into user's phase (end).

Both (exp 17) and (exp 18) were classified into user's action (move) in Figure 2. On the other hand, (exp 17) and (exp 18) were classified into user's phase (middle) and phase (start) in Figure 4, respectively. The number of tweets expressing the phase (middle) of user's action (move), such as (exp 17), was more than twice the number of those expressing the phase (start) of user's action (move), such as (exp 18).

(exp 17) *kiki nagara doraibu now –*
(I am driving a car now while listening to music –)
(exp 18) *yakin! chikusho- itte kuru!*
(Night shift! Damn it. Let's go!)

As shown in Figure 4, most of tweets expressing users' action (sleep) were classified into user's phase (start). However, there was a small number of tweets classified into user's phase (middle), such as (exp 19).

(exp 19) *nere masen*
(I can't sleep.)

Many of tweets expressing user's action (work) were classified into user's phase (end). However, we found some tweets classified into user's phase (middle), such as (exp 20) and (exp 21).

(exp 20) *shigoto tiu nano yo ne*
(Working now.)
(exp 21) *kore wo kiki tutu tabunya no eigo no kyokasho wo hitasura yakushite iru*
(I have been listening to this song and translated English textbooks in other areas entirely.)

We found some tweets which expressed users' actions, however, did not show the phases of them. For example, (exp 22) did not show the phase of user's action.

(exp 22) *asa undou*
(morning exercise)

3) Comments expressing reasons: We found many comments expressing users' reasons why they were listening to music by using music player applications.

(exp 23) *kibun teki ni kikitaku natta*
(I have a craving for music)
(exp 24) *katte shimatta*
(I finally bought it!)

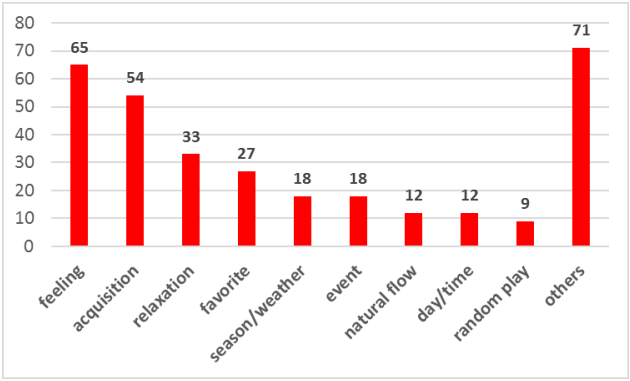


Figure 5. The classification result of users' results in the 312 tweets expressing users' reasons (by human experts).

(exp 23) and (exp 24) shows the reasons why the submitters of them were listening to music by using music player applications, feeling and acquisition, respectively. The submitter of (exp 23) felt an impulse and listened to music. On the other hand, the submitter of (exp 24) bought music contents and listened to it.

As shown in Figure 1, we found 312 tweets expressing users' reasons in the obtained 2,000 tweets which users submit when they use music player applications. We classified these 312 tweets expressing users' reasons why they were listening to music by using music player applications into ten types: (1) feeling, (2) acquisition, (3) relaxation, (4) favorite, (5) season/weather, (6) event, (7) natural flow, (8) day/time, (9) random play, and (10) others.

Figure 5 shows the classification result of the tweets expressing users' reasons why they were listening to music by using music player applications.

We classified (exp 23) into user's reason (feeling). This is because we thought the submitter of (exp 23) felt like listening to music. Also, we classified (exp 25) into user's reason (feeling). This is because we thought the submitter of (exp 25) did not listen to the song for a long time, and so, he/she felt like doing it.

(exp 25) *sugoku hisashiburi ni kiku.*
(I listen to this song after a long interval.)

We classified (exp 24) into user's reason (acquisition) because the submitter of (exp 24) bought and obtained the music content. Also, we classified (exp 26) into user's reason (acquisition) because the submitter of (exp 24) could not find the CD for a long time and found it.

(exp 26) *I found CD*

We classified (exp 27) and (exp 28) into user's reason (relaxation). In these tweets, the submitters played musics for relaxation.

(exp 27) *kibun tenkan* ♪
(relaxation ♪)
(exp 28) *toriaezu tenshon age*
(Let's get going)

(exp 29) and (exp 30) are examples of tweets classified into user's reason (favorite). This is because both the submitters of (exp 29) and (exp 30) addicted to the songs that they played.

(exp 29) *kinou kara do hamari shite iru*
(I have been addicted to the song since yesterday)
(exp 30) *1 nichi 1 kai ha kikanaito ikite ikenai karada ni*
(I will die if I do not play this song at least once a day)

(exp 31) and (exp 32) are examples of tweets classified into user's reason (season/weather). We thought both the submitter of (exp 31) and (exp 32) felt like listening to the song and submitting the tweets because they felt the sense of the season. As shown in Figure 5, 18 tweets were classified into this type. Seven of these tweets, including (exp 32), were submitted on November/24/2016, the first snow day of the winter. All these seven tweets touched the snow.

(exp 31) *kono kisetsu ha kore yana*
(This music suits the mood of the season)
(exp 32) *yuki to ieba kojinteki niha kore*
(I listen to the song when it snows)

We classified (exp 33) and (exp 34) into user's reason (event). The reasons why the submitter of (exp 33) and (exp 34) play the songs were the birthday of his/her friend and Halloween, respectively.

(exp 33) *Mikakoshi Happy Birthday !!!*
(My Mikako, Happy Birthday !!!)
(exp 34) *happi- harouin*
(Happy Halloween)

In both of (exp 35) and (exp 36), the submitters described that they selected and played the songs naturally. Take (exp 35) for example. *kotti mo* (this song) in (exp 35) meant implicitly that the submitter just before listened to the other song that had some kind of connection to this song. The connection let him/her select and play it. As a result, we classified (exp 35) and (exp 36) into user's reason (natural flow).

(exp 35) *kotti mo kika naku cha*
(I have to listen to this song)
(exp 36) *touzen no nagare*
(natural course)

We classified (exp 37) and (exp 38) into user's reason (day/time). The submitters of (exp 37) and (exp 38) listened to the songs because it was Sunday morning and night, respectively.

(exp 37) *nichiyoubi no asa ha, sawayaka ni heavy metal!! (((o(* °▽ °*)o)))*
(let's play heavy metal music refreshingly in Sunday morning!! (((o(* °▽ °*)o))))
(exp 38) *ichiou mada yonaka nano de kiku*
(I listen to this song because it is still night time)

We classified (exp 39) and (exp 40) into user's reason (random play). Both (exp 39) and (exp 40) were touched the songs that were selected randomly by music player applications.

(exp 39) *kyou no 1 kyoku me (random kettei)*
(Today's first song (random selection))
(exp 40) *soshite randam saisei de nagarete kita noga kore to iu*
(Then, random play and this song comes)

As shown in Figure 5, we found many tweets the comments of which were classified into user's reason (feeling) and (acquisition). This investigation result is not in good agreement with the survey on the use of portable music player conducted by Hamamura and Iwamiya [29]. They conducted the survey on 72 college students and reported that the reasons why the students used portable music players were relaxation (56 students), to kill time (51 students), to intercept environmental sound (27 students), to sharpen concentration (18 students), to improve operational efficiency (14 students), to avoid being talked to (13 students). The common reason of this survey and our investigation is only relaxation. This is because we investigated each tweets and the reason why the submitter listened to the song. On the other hand, Hamamura and Iwamiya did not survey every single use of portable music player. They surveyed the reasons why the college students used portable music players in their daily lives.

TABLE I. THE FEATURES USED IN MACHINE LEARNING METHODS FOR DATA TRAINING AND CLASSIFYING TWEETS EXPRESSING USERS' ACTIONS WHILE LISTENING TO MUSIC

$s1$	word unigrams of the comment
$s2$	word bigrams of the comment
$s3$	the number of words in the comment
$s4$	word unigrams of the first sentence of the comment
$s5$	word bigrams of the first sentence of the comment
$s6$	the number of words in the first sentence of the comment
$s7$	the last word of the first sentence of the comment
$s8$	character unigrams of the comment
$s9$	character bigrams of the comment
$s10$	character 3-grams of the comment
$s11$	the length of the comment
$s12$	character unigrams of the first sentence of the comment
$s13$	character bigrams of the first sentence of the comment
$s14$	character 3-grams of the first sentence of the comment
$s15$	the length of the first sentence of the comment

IV. DETECTION OF TWEETS EXPRESSING USERS' ACTIONS

What users are doing at a certain point in time is important to design various services and applications in social media that are relevant to what they are doing. If we detect users' actions while listening to music automatically, we can design behavior based services and applications in social media more precisely. For example, users may have free time to use services and applications when they are listening to music and going to somewhere. On the other hand, users may not want to be disturbed when they are lying down on their beds and listening to music. As a result, in this section, we discuss whether we can detect tweets including comments expressing users' actions, especially, move and sleep, from those including hashtags generated by music player applications by using machine learning techniques. Furthermore, we discuss whether we can detect tweets including comments expressing the phases of users' actions.

In this study, we used the 2,000 tweets investigated in Section III for the experimental data. The experimental data include 216 comments expressing users' actions. In this experiment, we used the support vector machine (SVM) and maximum entropy method (ME) for data training and classifying. Table I shows feature $s1$ to $s15$ used in machine learning on experimental data. $s1$ to $s7$ were obtained by using the results of morphological analysis on experimental

TABLE II. THE CLASSIFICATION RESULT OF USERS' ACTIONS IN THE 2,000 TWEETS INCLUDING HASHTAGS GENERATED BY MUSIC PLAYER APPLICATIONS.

(a) SVM classification result

	SVM results		
users' comments	action	others	recall
action	127	88	0.59
others	9	1776	0.99
precision	0.93	0.95	

(b) ME classification result

	ME results		
users' comments	action	others	recall
action	123	92	0.57
others	7	1778	1.00
precision	0.95	0.95	

data. In the experiments, we used a Japanese morphological analyzer, JUMAN, for word segmentation of tweets [30]. $s8$ to $s10$ and $s12$ to $s14$ were obtained by extracting character N-gram from experimental data. Odaka et al. reported that character 3-gram is good for Japanese processing [31]. $s4$ to $s7$ and $s12$ to $s15$ were obtained from first sentences of tweets. This is because, we thought, clue expressions of users' actions are often found at first sentences of tweets. We conducted this experiment using TinySVM [32] and maxent [33]. Table II shows the SVM and ME classification results of users' actions in the 2,000 tweets. The experimental result was obtained with 10-fold cross-validation. As shown in Table II, we obtained 95% accuracy each when we applied SVM and ME machine learning techniques to detect tweets including comments expressing user's actions. The SVM and ME precision of tweets including comments expressing user's actions were 93% and 95%, respectively. On the other hand, the SVM and ME recall of tweets including comments expressing user's actions were 59% and 57%, respectively. The experimental results show that our method failed to detect many tweets expressing users' actions. However, the precisions of our method show that our method is useful to collect tweets expressing users' actions precisely. In order to discuss the experimental result, we examined whether we can detect tweets including comments expressing users' actions, move and sleep, from those including hashtags generated by music player applications by using machine learning techniques.

The experimental data include

- 99 comments expressing users' action (move) and
- 62 comments expressing users' action (sleep).

Table III and Table IV show the classification result of users' action (move) and users' action (sleep) in the 2,000 tweets, respectively. As shown in Table III, we obtained 97% accuracy each when we applied SVM and ME machine learning techniques to detect tweets including comments expressing user's action (move). Also, as shown in Table IV, we obtained 99% accuracy each when we applied SVM and ME machine learning techniques to detect tweets including comments expressing user's action (sleep). Furthermore, the SVM and ME precision

TABLE III. THE CLASSIFICATION RESULT OF USERS' ACTION (MOVE) IN THE 2,000 TWEETS INCLUDING HASHTAGS GENERATED BY MUSIC PLAYER APPLICATIONS.

(a) SVM classification result

	SVM results		
users' actions	move	others	recall
move	48	51	0.48
others	5	1896	1.00
precision	0.91	0.97	

(b) ME classification result

	ME results		
users' actions	move	others	recall
move	41	58	0.41
others	2	1899	1.00
precision	0.95	0.97	

TABLE IV. THE CLASSIFICATION RESULT OF USERS' ACTION (SLEEP) IN THE 2,000 TWEETS INCLUDING HASHTAGS GENERATED BY MUSIC PLAYER APPLICATIONS.

(a) SVM classification result

	SVM results		
users' actions	sleep	others	recall
sleep	49	13	0.79
others	0	1938	1.00
precision	1.00	0.99	

(b) ME classification result

	ME results		
users' actions	sleep	others	recall
sleep	45	17	0.73
others	0	1938	1.00
precision	1.00	0.99	

of tweets including comments expressing user's action (move) were 91% and 95%, respectively. Also, the SVM and ME precision of tweets including comments expressing user's action (sleep) were 100% each. On the other hand, the SVM and ME recall of tweets including comments expressing user's action (move) were 48% and 41%, respectively. However, the SVM and ME recall of tweets including comments expressing user's action (sleep) were 79% and 73%, respectively. The reason why the recall of tweets including comments expressing user's action (sleep) was better than user's action (move) was that typical expressions, such as "*oyasuminasai* (good night)", were often used in comments expressing user's action (sleep). The experimental result shows that our method is useful to detect and collect tweets including comments expressing user's action (sleep). On the other hand, the recall of tweets including comments expressing user's action (move) shows that our method failed to detect many of them. However, the precision of them shows that our method is useful to collect them precisely.

Next, we discuss whether we can detect tweets including comments expressing the phases of users' actions, start, mid-

TABLE V. THE CLASSIFICATION RESULT OF THE PHASE (START) OF USERS' ACTIONS IN THE 2,000 TWEETS INCLUDING HASHTAGS GENERATED BY MUSIC PLAYER APPLICATIONS.

(a) SVM classification result

	SVM results		
action stages	start	others	recall
start	62	41	0.60
others	2	1895	1.00
precision	0.97	0.98	

(b) ME classification result

	ME results		
action stages	start	others	recall
start	59	44	0.57
others	1	1896	1.00
precision	0.98	0.98	

TABLE VI. THE CLASSIFICATION RESULT OF THE PHASE (MIDDLE) OF USERS' ACTIONS IN THE 2,000 TWEETS INCLUDING HASHTAGS GENERATED BY MUSIC PLAYER APPLICATIONS.

(a) SVM classification result

	SVM results		
action stages	middle	others	recall
middle	33	44	0.43
others	3	1920	1.00
precision	0.92	0.98	

(b) ME classification result

	ME results		
action stages	middle	others	recall
middle	29	48	0.38
others	1	1923	1.00
precision	1.00	0.98	

dle, and end. This is because the phases of users' actions enable us to provide more precise services and applications relevant to users' actions. The experimental data include

- 103 comments expressing the phase (start) of users' actions and
- 77 comments expressing the phase (middle) of users' actions.

Table V shows the classification results of the phase (start) of users' actions in the 2,000 tweets. Table VI shows the classification results of the phase (middle) of users' actions in the 2,000 tweets. As shown in Table V and Table VI, both of the precision of tweets including comments expressing the phase (start) and phase (middle) of users' actions were good. On the other hand, the recall of tweets including comments expressing the phase (start) was better than that of tweets including comments expressing the phase (middle). In order to discuss the experimental result, we examined whether we can detect tweets including comments expressing the phases of specific actions, move and sleep. The experimental data include

TABLE VII. THE CLASSIFICATION RESULT OF THE PHASE (START) OF USERS' ACTION (SLEEP) IN THE 2,000 TWEETS INCLUDING HASHTAGS GENERATED BY MUSIC PLAYER APPLICATIONS.

(a) SVM classification result

	SVM results		
action (sleep)	start	others	recall
start	50	9	0.85
others	0	1941	1.00
precision	1.00	1.00	

(b) ME classification result

	ME results		
action (sleep)	start	others	recall
start	45	14	0.76
others	0	1941	1.00
precision	1.00	0.99	

TABLE VIII. THE CLASSIFICATION RESULT OF THE PHASE (MIDDLE) OF USERS' ACTION (MOVE) IN THE 2,000 TWEETS INCLUDING HASHTAGS GENERATED BY MUSIC PLAYER APPLICATIONS.

(a) SVM classification result

	SVM results		
action (move)	middle	others	recall
middle	27	27	0.50
others	4	1942	1.00
precision	0.87	0.99	

(b) ME classification result

	ME results		
action (move)	middle	others	recall
middle	25	29	0.46
others	1	1945	1.00
precision	0.96	0.99	

- 59 comments expressing the start of users' sleep
- 54 comments expressing the middle of users' move

Especially, users who are in the middle of move are good targets for social media services, such as targeted advertisement and news recommendation. Table VII shows the classification results of the phase (start) of user's action (sleep) in the 2,000 tweets. Also, Table VIII shows the classification results of the phase (middle) of user's action (move) in the 2,000 tweets. In both cases, we obtained the high accuracy and precision. However, the recall of tweets including comments expressing the phase (start) of user's action (sleep) was good while that of the phase (middle) of user's action (move) was not good. This is because Twitter users often submit short messages including typical expressions, such as "*oyasuminasai* (good night)", in order to inform they cannot read any messages while they are sleeping. Nakao reported that there are many Japanese young SNS users who feel regret when they cannot reply to SNS messages rapidly [34]. As a result, many Twitter users submit messages including these typical expressions, such as

"*oyasuminasai* (good night)", before they are sleeping.

The experimental results show that our method could not detect many tweets expressing users' actions. However, the precisions of our method show that our method is useful to collect tweets expressing users' actions precisely. In other words, tweets detected by our method are useful to understand what users were doing and what phases users were in. As a result, our method is useful to provide social media services, such as targeted advertisement, news recommendation, and real-world analysis.

V. Conclusion

Social media such as Twitter generate large quantities of data about what users are thinking and doing at a certain point in time. In this respect it is important to design various services and applications in social media, such as targeted advertisement, news recommendation, and real-world analysis. As a result, in this study, we investigate tweets submitted by music player applications and show what the users are thinking and doing while listening to music. Furthermore, we apply machine learning techniques to detect tweets submitted by music player applications and discuss whether we can detect tweets expressing what the users are doing and what action phases they are in while listening to music. In both cases, we obtained the high accuracy and precision, however, the low recall. The low recall shows that our method often failed to detect tweets expressing users' actions. However, the high accuracy and precision show that most of detected tweets were classified correctly. In other words, tweets detected by our method are useful to understand what users were doing and what action phases they are in. As a result, our method is useful to provide social media services, such as targeted advertisement, news recommendation, and real-world analysis. We intend to use the results of this study for further investigation of tweets expressing users' emotions and sentiments. This is because more than half of the investigated tweets were classified into ones expressing users' impressions.

References

[1] Y. Watanabe, K. Yasuda, R. Nishimura, and Y. Okada, "An investigation of tweets submited by using music player applications," in Proceedings of the Ninth International Conference on Evolving Internet (INTERNET 2017), Jul 2017, pp. 24–29. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=internet_2017_2_30_40029 [accessed: 2018-5-25]

[2] T. Sakaki and Y. Matsuo, "Twitter as a social sensor : Can social sensors exceed physical sensors?" Journal of Japanese Society for Artificial Intelligence, vol. 27, no. 1, jan 2012, pp. 67–74.

[3] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: Detecting influenza epidemics using twitter," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1568–1576. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145600 [accessed: 2018-5-25]

[4] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in Proceedings of the First Workshop on Social Media Analytics, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 115–122. [Online]. Available: http://doi.acm.org/10.1145/1964858.1964874 [accessed: 2018-5-25]

[5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in Proceedings of the 19th International Conference on World Wide Web, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777 [accessed: 2018-5-25]

[6] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 11, Nov. 2009, pp. 2169–2188. [Online]. Available: http://dx.doi.org/10.1002/asi.v60:11 [accessed: 2018-5-25]

[7] M. Motoyama, B. Meeder, K. Levchenko, G. M. Voelker, and S. Savage, "Measuring online service availability using twitter," in Proceedings of the 3rd Workshop on Online Social Networks, ser. WOSN'10. Berkeley, CA, USA: USENIX Association, 2010, pp. 13–13. [Online]. Available: https://cseweb.ucsd.edu/~savage/papers/WOSN10.pdf [accessed: 2018-5-25]

[8] A. Signorini, A. M. Segre, and P. M. Polgreen, "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic," PLoS One, vol. 6, no. 5, May 2011. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019467 [accessed: 2018-5-25]

[9] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 497–506. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557077 [accessed: 2018-5-25]

[10] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1079–1088. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753486 [accessed: 2018-5-25]

[11] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," in Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, ser. LBSN '10. New York, NY, USA: ACM, 2010, pp. 1–10. [Online]. Available: http://doi.acm.org/10.1145/1867699.1867701 [accessed: 2018-5-25]

[12] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng, "Spatio-temporal dynamics of online memes: A study of geo-tagged tweets," in Proceedings of the 22Nd International Conference on World Wide Web, ser. WWW '13. New York, NY, USA: ACM, 2013, pp. 667–678. [Online]. Available: http://doi.acm.org/10.1145/2488388.2488447 [accessed: 2018-5-25]

[13] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs," in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ser. CIKM '11. New York, NY, USA: ACM, 2011, pp. 2541–2544. [Online]. Available: http://doi.acm.org/10.1145/2063576.2064014 [accessed: 2018-5-25]

[14] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating twitter users," in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 759–768. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871535 [accessed: 2018-5-25]

[15] Y. Yamaguchi, T. Amagasa, H. Kitagawa, and Y. Ikawa, "Online user location inference exploiting spatiotemporal correlations in social streams," in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ser. CIKM '14. New York, NY, USA: ACM, 2014, pp. 1139–1148. [Online]. Available: http://doi.acm.org/10.1145/2661829.2662039 [accessed: 2018-5-25]

[16] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, ser. EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1277–1287. [Online]. Available: http://dl.acm.org/citation.cfm?id=1870658.1870782 [accessed: 2018-5-25]

[17] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: The dynamics of the location field in user profiles," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '11. New

York, NY, USA: ACM, 2011, pp. 237–246. [Online]. Available: http://doi.acm.org/10.1145/1978942.1978976 [accessed: 2018-5-25]

[18] B. Han, P. Cook, and T. Baldwin, "Geolocation prediction in social media data by finding location indicative words," in COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India, M. Kay and C. Boitet, Eds. Indian Institute of Technology Bombay, 2012, pp. 1045–1062. [Online]. Available: http://aclweb.org/anthology/C/C12/C12-1064.pdf [accessed: 2018-5-25]

[19] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser, "A multi-indicator approach for geolocalization of tweets." in ICWSM, E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, and I. Soboroff, Eds. The AAAI Press, 2013.

[20] D. Rout, K. Bontcheva, D. Preoţiuc-Pietro, and T. Cohn, "Where's @wally?: A classification approach to geolocating users based on their social ties," in Proceedings of the 24th ACM Conference on Hypertext and Social Media, ser. HT '13. New York, NY, USA: ACM, 2013, pp. 11–20. [Online]. Available: http://doi.acm.org/10.1145/2481492.2481494 [accessed: 2018-5-25]

[21] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1100–1108. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020581 [accessed: 2018-5-25]

[22] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in Proceedings of the 19th International Conference on World Wide Web, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 61–70. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772698 [accessed: 2018-5-25]

[23] A. Sadilek, H. Kautz, and J. P. Bigham, "Finding your friends and following them to where you are," in Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, ser. WSDM '12. New York, NY, USA: ACM, 2012, pp. 723–732. [Online]. Available: http://doi.acm.org/10.1145/2124295.2124380 [accessed: 2018-5-25]

[24] S. Kinsella, V. Murdock, and N. O'Hare, ""i'm eating a sandwich in glasgow": Modeling locations with tweets," in Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents, ser. SMUC '11. New York, NY, USA: ACM, 2011, pp. 61–68. [Online]. Available: http://doi.acm.org/10.1145/2065023.2065039 [accessed: 2018-5-25]

[25] Y. Watanabe, K. Nakajima, H. Morimoto, R. Nishimura, and Y. Okada, "An investigation of a factor that affects the usage of unsounded code strings at the end of japanese and english tweets," in Proceedings of the Seventh International Conference on Evolving Internet (INTERNET 2015), Oct 2015, pp. 50–55. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=internet_2015_2_40_40038 [accessed: 2018-5-25]

[26] Twitter, "How to use hashtags," https://help.twitter.com/en/using-twitter/how-to-use-hashtags [accessed: 2018-5-25].

[27] S. Weinschenk, 100 Things Every Designer Needs to Know About People, 1st ed. Thousand Oaks, CA, USA: New Riders Publishing, 2011.

[28] I. E. Hyman, S. M. Boss, B. M. Wise, K. E. McKenzie, and J. M. Caggiano, "Did you see the unicycling clown? inattentional blindness while walking and talking on a cell phone," Applied Cognitive Psychology, vol. 24, no. 5, 2010, pp. 597–607. [Online]. Available: http://dx.doi.org/10.1002/acp.1638 [accessed: 2018-5-25]

[29] M. Hamamura and S. Iwamiya, "Survey on the use of portable audio devices by university students," The Journal of the Acoustical Society of Japan, vol. 69, no. 7, jul 2013, pp. 331–339.

[30] S. Kurohashi and D. Kawahara, JUMAN Manual version 5.1 (in Japanese). Kyoto University, 2005.

[31] T. Odaka et al., "A proposal on student report scoring system using n-gram text analysis method," The transactions of the Institute of Electronics, Information and Communication Engineers. D-I, vol. 86, no. 9, sep 2003, pp. 702–705. [Online]. Available: http://ci.nii.ac.jp/naid/110003171273/en/ [accessed: 2018-5-25]

[32] Taku Kudoh. TinySVM: Support Vector Machines. [Online]. Available: http://chasen.org/˜taku/software/TinySVM/index.html [accessed: 2018-5-25]

[33] M. Utiyama, "Maximum entropy modeling packages," http://mastarpj.nict.go.jp/˜mutiyama/software/maxent [accessed: 2010-7-27], 2008.

[34] M. Nakao, "Social media and children : Report from the world summit on media for children (wsmc)," The NHK monthly report on broadcast research, vol. 65, no. 3, mar 2015, pp. 76–80. [Online]. Available: https://ci.nii.ac.jp/naid/110009890194/en/ [accessed: 2018-5-25]

Functional Layered Architectures and Control Solutions in Internet of Vehicles – Comparison

Eugen Borcoci
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: eugen.borcoci@elcom.pub.ro

Marius Constantin Vochin
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: marius.vochin@upb.ro

Serban Georgica Obreja
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: serban@radio.pub.ro

***Abstract* — The Internet of Vehicles is a novel development trend in vehicular networking. Its driving factor is, on one part, the high growth of the vehicles number, including the intelligent ones and the need to solve numerous problems encountered in transportation systems related to safety, traffic management, information and entertainment services, autonomic vehicles challenges and so on. Internet of Vehicles extends the capabilities of the traditional Intelligent Transport System technologies but also takes benefit from new technologies used in Future Internet. It is considered by many authors as a sub-domain of Future Internet and specifically of Internet of Things. Internet of Vehicles will integrate the previous Vehicular Networks and also functionalities already developed in ITS. However, there is no unique definition of what Internet of Vehicles exactly is; some concepts and architectural aspects are still open research issues. This paper is not an exhaustive survey; it attempts a comparative critical analysis of several functional architectures and systems proposed for Internet of Vehicles. Recent approaches Fog/Edge computing – based systems and Software Defined Networking are also considered. An enriched SDN/Fog based architecture is proposed.**

Keywords — Internet of Vehicles; V(A)NET; Fog computing; Edge computing; Software Defined Networking; Network Function Virtualization.

I. INTRODUCTION

This paper is an extended version of the work [1], dedicated to a comparative analysis of some relevant functional architectures recently proposed for the *Internet of vehicles* (IoV). The aim here is not to detail certain functions or services, but to evaluate several variants of structured layering of functions and possible separation of functions among several architectural planes.

Vehicular communications, networks and many associated services have been intensively studied, designed, standardized and also implemented in the last two decades. The driving force has been and still is, the significant growth of the vehicles number all over the world, together with many problems related to transportation, but also due to the market needs of new services available in vehicular environment. The umbrella and framework for such developments is the *Intelligent Transport System* (ITS) [2].

Complementary support networking technologies have been developed in this area including the lower layers (physical and data link layers for wireless access) *Dedicated Short-Range Communications* (DSRC) and also higher functional layers *Wireless Access in Vehicular Environments* (WAVE) [3]. The IEEE 802.11a/p and respectively IEEE 1609 represent a mature set of standards for DSRC/WAVE networks. For wide area and high-speed mobility, another solution for wireless access of vehicles is based on 4G, *Long Term Evolution* (LTE) technology and recently on LTE-A (LTE-Advanced). Experiments have shown that the vehicles can operate with the speed of ~150 km/h. An alternative to LTE is WiMAX (World-wide interoperability for Microwave Access).

Vehicular Ad Hoc Networks (VANET) [4] have been defined to support basic vehicular communications types: *vehicle to vehicle (V2V)*, *vehicle to road (V2R)*, or *vehicle to Infrastructure (V2I)* in uni-directional or bi-directional communications (note that, some authors include V2R into V2I type). The basic VANET functional components are the *On-Board-Unit* (OBU), installed into the vehicles and *Road-Side–Unit* (RSU) placed on the roads. The RSUs communicates with vehicles, can inter-communicate and also could be linked to external networks like Internet. The main applications and services of VANET have been oriented to safety and traffic management use-cases.

The VANETs have several limitations related to their pure ad hoc network architecture (in V2V case), unreliable Internet service, incompatibility with personal devices, non-cooperation with cloud computing, low accuracy of the services, operational network dependency and restricted areas of applications and services. Therefore, extending the VANET architecture is considered today as a strong need and an opportunity.

Recently, *Internet of Vehicles* (IoV) concepts and architectures have been proposed as a significant enhancement in vehicular communication area. IoV could be seen as a global span of a vehicle network [5-9]. On the other part, IoV is considered as a special case of *Internet of Things* [10] [11], where the "things" are either vehicles or their subsystems. The IoV will connect the vehicles and RSUs through different *Wireless/Radio Access Technologies* (WAT/RAT), while traditional Internet and other heterogeneous networks will be used for wide area. In terms of services, IoV has as objectives to include the traditional VANET services but also will be open for development of novel ones, e.g., vehicle traffic management in urban or country areas, automobile production, repair and vehicle insurance, road infrastructure construction and repair, logistics and transportation, etc.

The IoV can be strongly supported by recent technologies like centralized *Cloud Computing* (CC) combined with *Fog* or *Edge Computing* [11] [12]; in comparison with CC, the Fog/Edge can offer for IoV a better time response, more flexibility and higher degree of functional distribution, context awareness, reduction in the amount of data exchanged between a cloud data center and a vehicle. All these features are more appropriate for vehicular world in comparison to centralized cloud computing approach.

In terms of management and control, *Software-defined networking* (SDN) technology [13] can offer to IoV its centralized up-to-date logical view upon the network, programmability, facilitating a flexible network management and on-the-fly modification of the network elements behavior.

Network Function Virtualization (NFV) [14] can add flexibility by virtualizing many network functions and deploying them into software packages. Dedicated *Virtualized Network Functions* (VNF) can be defined, then dynamically created/used/destroyed, assembled and chained to implement legacy or novel services. Challenges and open research issues exist, related to NFV and SDN cooperation and their adaptation to the vehicular networks requirements concerning high mobility, distributed character, aiming finally to realize new flexible and powerful IoV architectures and systems.

The large communities of users/terminal devices in IoV need powerful and scalable *Radio Access Technologies* (RAT). The 4G and the emergent 5G technologies, based on cloud computing architectures (*Cloud Radio Access Network*- CRAN) are significant candidates for constructing the IoV access infrastructure [15].

Despite IoV promises high capabilities, there still exist many challenges, both in conceptual and architectural aspects and also from implementation and deployment points of view. Many IoV advanced features and integration with the above technologies (CC, Fog/Edge, SDN, NFV) are still open research issues.

This paper attempts a comparative critical study of several functional layered architectures proposed for IoV, including recent ones based on Fog/Edge computing and Software defined networking (SDN) - control. An enriched functional architecture with Fog computing and SDN control is proposed in the paper. Other candidate support technologies for IoV, like Mobile Edge Computing are shortly discussed. The Sections III and V contains the main additional contributions w.r.t the original work [1].

The paper is organized as follows. Section II is a short overview of related work on IoT layered architecture. Section III exposes a comparative presentation of some IoV generic layered functional architectures. Section IV revisits the SDN-based architectures of IoV. Section V proposes a Fog-SDN oriented, enriched integrated architecture. Section VI presents a mapping example of the generic IoV architecture on Mobile Edge Computing (MEC) technology. Section VII draws some conclusions and exposes future work.

II. INTERNET OF THINGS LAYERED ARCHITECTURES

IoV is frequently seen as a part of the more general Internet of Things (IoT), so it is of interest to compare how the IoV architectures are generally consistent with previously proposed IoT architectures.

Among several architectural overviews and stacks suggested for IoT, Al-Fuqaha et al. [10] present an interesting IoT overview. They identify several IoT elements, i.e., *identification, sensing, communication, computation, services and semantics.* Several variants of IoT layered architectures are presented, where the most comprehensive has 5-layers:

- *Business (BL)*- highest layer
- *Application (AL)*
- *Service Management (SML)*
- *Object Abstraction (OAL)*
- *Objects (perception) (OL)*- lowest layer

If compared with the classical TCP/IP architecture, the above layers are defined in a more general way, but the layering principles are still preserved, in the sense that a given layer offers a set of services to the upper layer.

The *Object (perception)* layer (lowest) represents the IoT physical sensors and actuators, performing functionalities such as querying location, temperature, weight, motion, vibration, acceleration, humidity, etc. The digitized data are transferred to the OAL through secure channels.

The *Object Abstraction* layer transfers abstracted data to the *Service Management layer* through secure channels. Traditional Layer 2 networking transfer functions are included here, based on technologies like RFID, GSM, 3G, 4G, UMTS, WiFi, Bluetooth Low Energy, infrared, ZigBee, etc. Additionally, cloud computing capabilities are offered, and data management processes are handled at this layer.

The *Service Management* layer plays a middleware role, by pairing a service with its requester based on addresses and names. The SML supports IoT application programmers to work with abstracted heterogeneous objects. It also

processes received data, takes decisions, and delivers the required services over the network wire protocols.

The *Application* layer provides to the customers the requested services (with appropriate quality). The AL covers different vertical markets (e.g., smart home, transportation, industrial automation, health care, etc.).

The *Business Layer* manages all IoT system activities and services. Using data provided by AL, it creates a business model, graphs, flowcharts, etc.; it is related to design, analysis, implementation, evaluation, monitoring and management (of the lower layers), and developing IoT system related elements. Decisions can be taken following Big Data analysis. Security features are included. Note that the architecture described above is a high-level view only; further structuring can be made and mapping on various existing protocols.

The work of Khan et al. [16] also proposes a five-layer architecture for IoT, which is similar to the previous one.

1) *Perception/Device Layer (PL):* consists of the physical objects and sensors (RFID, 2D-barcode, etc.) and basically deals with the identification and collection of objects specific information by the sensor devices. The collected information is then passed to Network layer for its secure transmission to the information processing system.

2) *Network Layer (NL):* securely transfers the information from sensor devices to the information processing system. The transmission technologies can be 3G, UMTS, Wifi, Bluetooth, infrared, ZigBee, etc.

3) *Middleware Layer (ML):* is responsible for the service management and has links to the database. It processes information received from NL and store it in the database. It takes automatic decision based on the results.

4) *Application Layer (AL):* provides global management of the vertical applications based on the objects information processed in the Middleware layer.

5) *Business Layer:* manages the overall IoT system including the applications and services. It builds business models, graphs, flowcharts etc based on the data received from AL.

III. IoV Generic Layered Architectures

Several IoV architectures have been recently proposed and discussed. A short critical overview and comparison are exposed below.

Bonomi et al. [5] proposed a four - layered architecture for connected vehicles and transportation. The layers are also called *"IoT key verticals"*, suggesting that a given layer includes not only classical layer functions (i.e., L1, L2,) but rather groups of functions, which could be mapped on one or more classical layers. Also, the four layers are rather corresponding to different geo-locations of the subsystems (vehicles, networking infrastructure, cloud data centers, etc.).

The bottom layer (*end points*) represents the vehicles, plus their communication protocols (basically for V2V communication, using the IEEE 802.11a/p).

The layer two (*infrastructure*), represents communication technologies to interconnect the IoV actors (via WiFi, 802.11p, 3G/4G, etc.).

The third layer (*operation*) performs management actions; it verifies and ensures compliance with all applicable policies, to regulate the information management and flow.

The fourth layer is called *services*/*cloud* (public, private or enterprise) based on a defined profile coupled with the possibility of receiving services (voice, enterprise video and data) on demand. Note that this architectural view is a mixed one and does not clearly separate the sets of functions of various levels.

Note the partial similarity of the above architecture to those described in Section II. However, the cloud layer in the IoV is considered as the top layer in Bonomi's case, including the applications and business functions of the previous IoT architectures.

Kayvartya et al. [6] have proposed a comprehensive IoV five-layer architecture, to support an enriched set of vehicular communications, in addition to traditional V2V, V2R/V2I, i.e., *Vehicle-to-Personal* devices (V2P) and *Vehicle-to-Sensors* (V2S). Each IoV communication type can be enabled using a different WAT, e.g., IEEE WAVE for V2V and V2R, Wi-Fi and 4G/LTE for V2I, CarPlay/NCF (*Near Field Communications*) for V2P and WiFi for V2S. The system includes vehicles and *Road Side Units* (RSU), but also other communication devices. Embedding such a large range of devices makes IoV more complex, (compared to VANET), but more powerful and market oriented.

This architecture goes further than only proposing a generic overall model; separation in three architectural planes is defined: *management*, *operation* and *security*. Such a split is important because it allows later to map various existing protocols and functions (e.g., taken from ITS) to be more easily mapped on architectural layers. The network model is composed of three functional entities: *client, connection* and *cloud.* The layers are (see Figure 1): *perception, coordination, artificial intelligence, application* and *business.*

The *perception* layer (PL) functions generally include those of the traditional physical layer but have also some additional functions related to sensing and actuating actions. The PL is instantiated by *sensors* and *actuators* attached to vehicles, RSUs, smart-phones and other personal devices. The PL main task is to gather information on vehicle, traffic environment and devices (including movement–related parameters).

The *coordination* layer (CL) represents a virtual universal network coordination entity for heterogeneous network technologies (WAVE, Wi-Fi, 4G/LTE, etc.). It creates a unified communication structure for the terminal devices.

The *artificial intelligence* layer (AIL) is represented by a generic virtual cloud infrastructure, working as an information processing and management centre. It stores, processes and analyses the information received from the lower layer and then takes decisions. Its major components are: *Vehicular Cloud Computing (VCC), Big Data Analysis (BDA)* and *Expert System.* The AIL should meet the requirement of applications and services working on top of it.

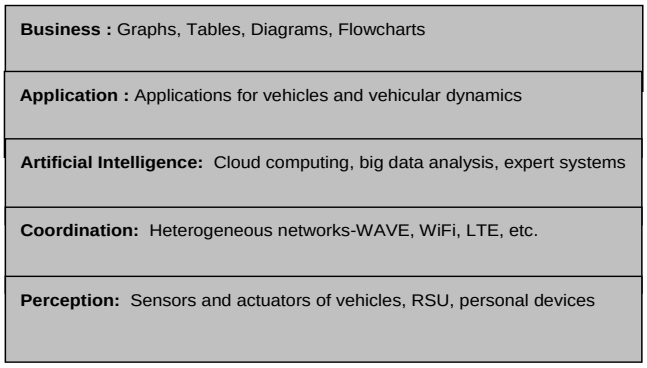


Figure 1. Five-layer IoV architecture (adapted from [6]).

The *application* layer (AL) contains smart applications (e.g., for traffic safety and efficiency, multimedia-based infotainment and web-based utility). The AL includes safety and efficiency applications (VANET legacy) and provides smart services to End Users (EU) based on intelligent analysis done by AIL. The AL efficiently discovers the services provided by AIL and manage their combinations. It also provides EU application usage data to the business layer. Currently, it is recognized that these smart applications constitute a major driving force to further develop IoV.

The *business* layer (BL) includes IoV operational management functions, basically related to business aspects: to foresight strategies for the development of business models based on the application usage data and statistical analysis of the data; analysis tools including graphs, flowcharts, comparison tables, use case diagrams, etc.; decision making - related to economic investment and usage of resources; pricing, overall budget preparation for operation and management; aggregate data management.

The architecture is split in three parallel planes: *operation, management and security.* The work [6] also proposed a possible mapping between the five layers and different protocols already developed in vehicular communications by ITS, VANET, IEEE, etc. The *operation* plane basically contains traditional *data* plane functions but still has some control and management role.

At *perception* layer, current network technologies can be used for access in ITS and VANET (see Figure 2).

The *coordination* layer includes not only TCP/IP transport and network protocols but also different solutions (with no IP usage). Examples are: IEEE 1609.4 along with a *Global Handoff Manager* (GHM-open research) and other protocols proposed at network layer in projects like CALM, WAVE. For instance, in the stack there exist WSMP - Short Message Protocol and FAST -Fast Application and Communication Enabler.

In the *Artificial Intelligence* layer, cloud capabilities are seen as major contributors, working on top of lower sub-layer: CALM Service Layer (CALM-SL) and WAVE-1609.6 service related protocols. The upper sub-layer consists in Vehicular Cloud Computing (VCC) and Big Data Analysis (BDA) related protocols. They can offer cloud services of type "*X as a Service*": Storage (STaaS), Infrastructure (INaaS), Network (NaaS), Cooperation (CaaS), Entertainment (ENaaS), Gateway (GaaS), Picture (PICaaS) and Computing (COMaaS).

Still further research work is necessary, given the current unavailability of enough suitable protocols for VCC and BDA. Another open issue is that VANETs projects, generally, do not have clear definitions of the upper sub-layer, while some IoT projects are recently working towards these.

The *Application* layer includes two sets of applications: *Smart Safety and Efficiency* (SSE) and *Smart Business Oriented* (SBO). The current WAVE resource handler protocol 1609.1 can be used on the top of these applications, to manage the resources among smart applications. The *Business Layer* (BL) in [6] proposes various business models like Insurance (INS), Sale (SAL), Service (SER) and Advertisement (ADV). The set of these functionalities could be further enriched in the future.

The architecture has the merit to integrate in the management and security planes some existing functional blocks and protocols (see Figure 2), already developed in WAVE (P1609.x), CALM and C2C projects.

However, the mentioned 5-layer architecture does not touch some important and recent aspects in developing IoV architecture, e.g., how to distribute computation intelligence between a central cloud and fog/edge units (placed at the network edge) while cloud-fog/edge combination seem to be an efficient and attractive solution for a distributed system like IoV. Also, SDN-like control and NFV implementation possibilities are not discussed in this architecture.

F.Yang et al. work [7] proposes a more comprehensive view on IoV architecture, based on functional requirements and proposed goals, by considering the *driver-vehicle-environment* coordination. IoV is defined as an open converged network system (controllable, manageable, operational, and trustable) based on multi-human, multi-machine, multi-vehicle, and environment coordination. It senses, recognizes, transmits, and computes the large-scale complex static/dynamic information of human, vehicle, network communication and road traffic infrastructure, using advanced ICT technology.

The architecture [7] defines four layers: the *environment sensing and control* layer, *network access and transport* layer, *coordinative computing control* layer, and *application* layer (see Figure 3). The work also summarizes the core technologies of each layer. In the *environment sensing and control* layer, vehicle control and environment sensing technologies are introduced. The network access and transport layer use the current technologies available for vehicle access and communication (access and core networks).

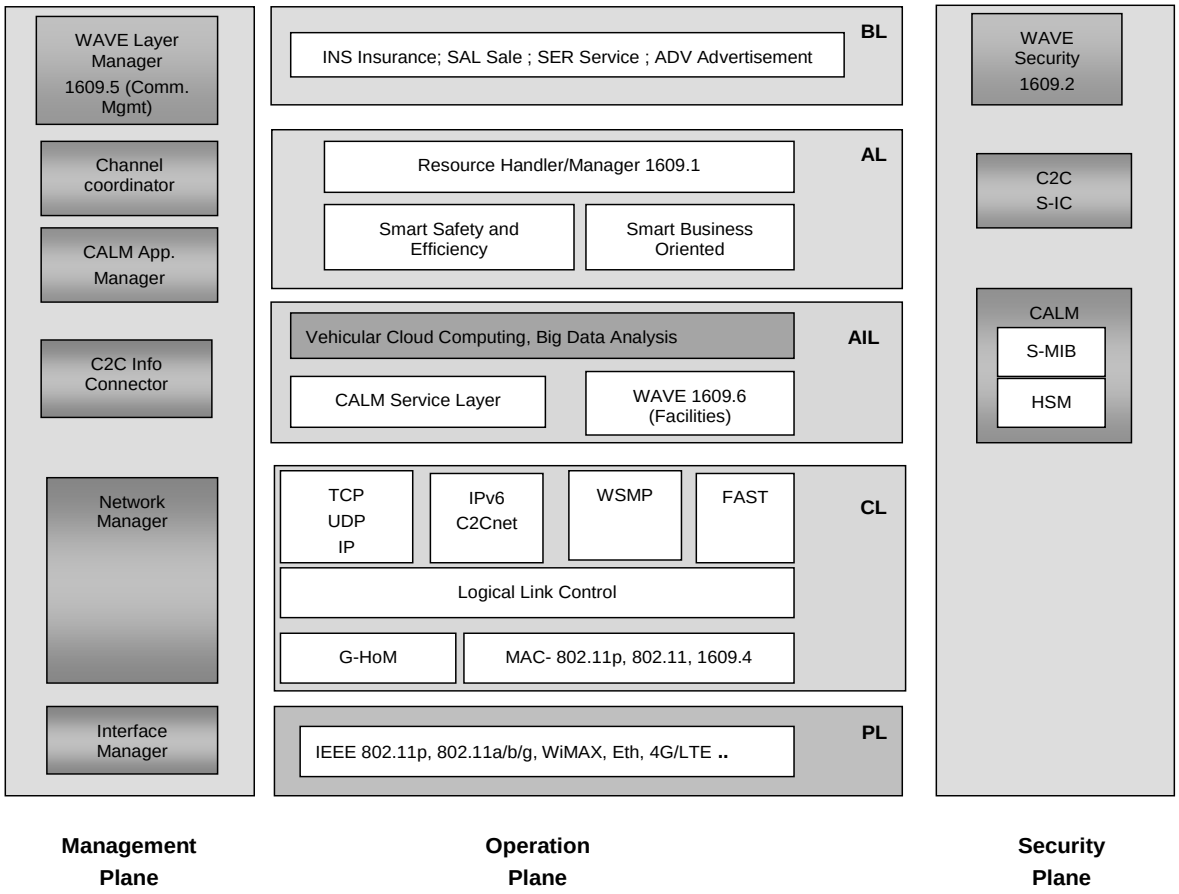


Figure 2. Five-layer IoV architecture mapped on particular protocols (adapted from [6])

PL Perception layer; CL Coordination layer; AIL Artificial Intelligence layer; AL Application layer; BL Business layer; C2C Car to Car; CALM Communication Architecture for Land Mobile; DSRC Dedicated Short Range Communication; WAVE Wireless Access in Vehicular Environment; FAST Fast Application and Communication Enabler; LLC Logical Link Control; G-HoM Global Handoff Manager; WSMP WAVE Short Messages Protocol; BDA Big Data Analysis; VCC Vehicular Cloud Computing; SSE Smart Safety and Efficiency; SBO Smart Business Oriented; INS Insurance; SAL Sale ; SER Service ; ADV Advertisement; HSM Hardware Security Manager; S-IC Security Information Connector; S-MIB Security Management Information Base

A special layer (differently defined w.r.t. other IoV architectural proposals) is introduced, i.e., the *coordinative computing contro*l layer. Here, the coordination among *human-vehicle-environment* in IoV is considered as a main goal. The application layer splits the services in two classes: close and open.

The coordination concept in [7] is based/divided on/in two models and, correspondingly, two objects:

a. *individual coordination model*- dealing with the capabilities of the *pair human-vehicle* and assuring coordinative computing control in the IoV environment. It solves the coordination problems between human and vehicle, and between individual object and swarm object. The in-vehicle network is involved here.

b. *swarm coordination model;* the swarm object consists of all objects of IoV except the individual object. The environment network is involved here.

The *vehicle network environment sensing and control* layer offers the basis for IoV services, including those for autonomous vehicle. The environment sensing is the recognition basis for IoV services, such as services of autonomous vehicle, intelligent traffic, and vehicle information.

From the perspective of vehicles, they sense environment information around these vehicles via autopilot system, traffic jam auxiliary system, and sensor system for achieving auxiliary driving. In terms of environment, this layer monitors and extracts various dynamic information of human, vehicles, and environment through sensing technology. It receives and executes coordinative control instructions and then feedback result to cooperative control. It contributes to the implementations of swarm sensing in swarm model.

The *network access and transport* layer mainly realize the network access, data processing, data analysis, and data transmission, remote monitoring and nodes management. It implements the inter-connection and information exchange, between entities, manages the connectivity resources and

balances information load. When it is the case, it can offer a stable and quality-guaranteed information and communication transport.

The *coordination computing control layer* performs network-wide coordinative computing and control for human-vehicle-environment (data processing, resource allocation and swarm intelligence computing). This layer should include both capabilities to solve the individual model (human-vehicle) related functions and also capabilities for the multi-human and multi-vehicle coordinative computing control and service coordinative management, to support the swarm intelligence computation and various services. This layer should also provide the capability of communication coordinated management.

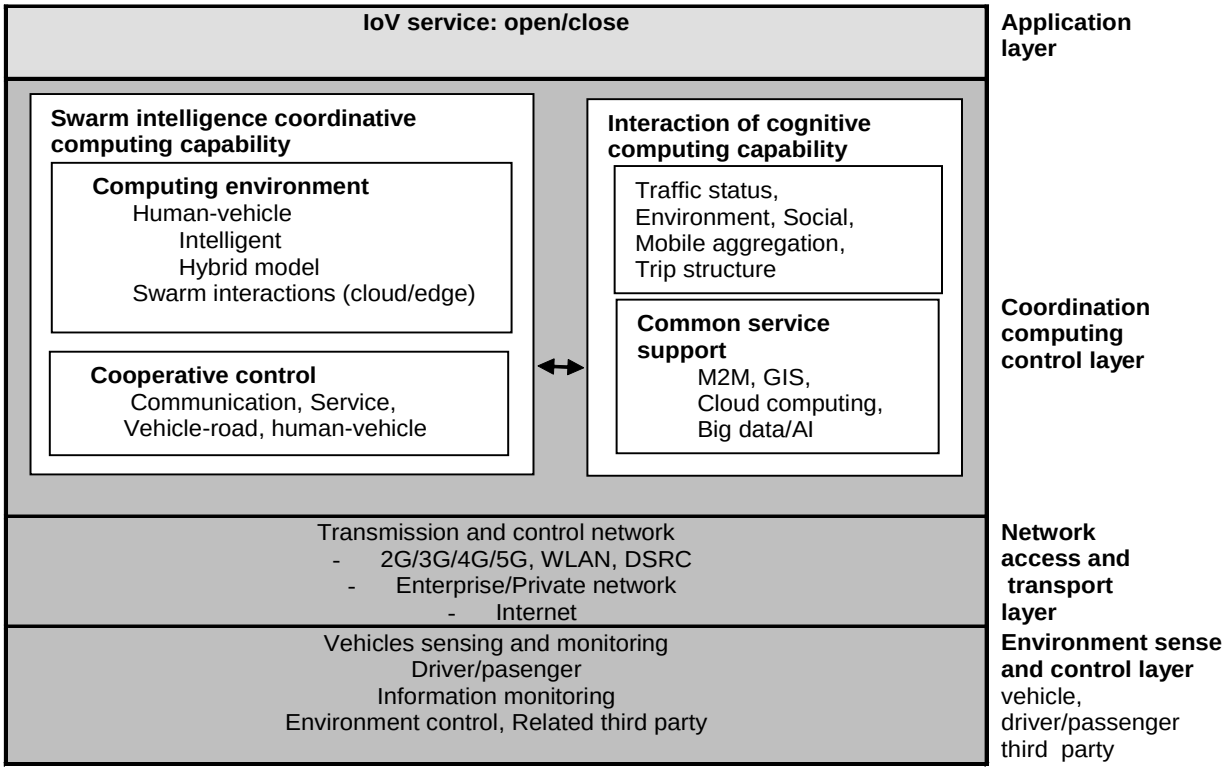
The *application* layer is defined to provide various types of services. It should be open in the sense that could support novel services and business operating modes. The application layer can be classified into closed services (related to the specific industry applications) and open services (i.e., various existing open applications, such as real-time traffic services provided by Internet service providers or to third party providers).

The architecture [7] presents (see Figure 3) in a generic way, the four layers and their internal components. *However, the criteria of splitting the entities/functions of the components included in the coordination computing layer are not very visible, e.g., between the two blocks: swarm intelligent coordinative computing and interaction of cognitive computing capabilities.*

The homogeneity of sub-layers is low in terms of their components. The separation of the overall architecture in different architectural planes is not discussed; therefore, is rather difficult to see the mapping of different already developed functions and protocols (ITS, WAVE, etc.) to the layers of this architecture. This seems to be still an open issue of this architecture.

No consideration about using technologies like SDN, NFV, Fog/edge computing (except a proposal of a virtual vehicle -VV) are mentioned. A refinement of this architecture and more precise structuring would be needed.

Contreras-Castillo et al. [8] propose a seven-layer architecture, supporting the functionalities, interactions, representations and information exchanges among all the devices inside an IoV ecosystem. The authors claim that this architecture (having more than five layers) has as objective to reduce the complexity of each layer and better standardize the interfaces and protocols used in each layer.



F

Figure 3. Four-layer IoV architecture (adapted from [7]).

The interaction model considers the following entities, which can communicate to each other: vehicle (V), person (P), personal device (PD), network infrastructure (I), sensors (S), any device (D) and roadside device (R). Consequently, the communications types might be: V2V, V2R, V2I, V2D, V2P, V2S, D2D.

The network model should support collaboration between multi-users, multi-vehicles, multi-devices (sensors, actuators, mobile devices, access points), multi-communication models (point to point, multi-point, broadcast, geo-cast) and multi-networks (wireless or wire networks with various technologies like WiFi, Bluetooth, WiMAX, 3G, 4G/LTE, etc.

The layers defined in [8] are (bottom-up list):

1 *User interaction* (lowest layer)
2 *Data acquisition*
3 *Data filtering and pre-processing*
4 *Communication*
5 *Control and management*
6 *Business* (highest layer)

An additional layer is named *Security*; however, it is actually a cross layer entity.

Note that *this "layered"- named architecture does not follow the principles of a layered stack architecture* (where each layer traditionally offers some services to the above one). For instance, the Control and management layer and Security layer seem to be rather architectural "planes "and not traditional layers; they have to interact with all other five layers.

The *User interaction* layer contains in-vehicle computing systems including:

a. information-based systems to provide information (e.g., on routes, traffic conditions, car parking availability and warning/advice regarding risks) to components of the driving environment, i.e., the vehicle or the driver;

b. control-based systems to monitor changes in driving habits and experiences and operational elements of the driving task (e.g., adaptive cruise control, speed control, lane keeping and collision avoidance).

It is stated in [8] that designing user interfaces for in-vehicle systems is still raising many new research challenges. Note also that this "layer" actually contains functions of several layers defined in other architectures (e.g., some structured in a similar way as classic TCP/IP stack).

The *Data acquisition* layer has tasks covering all three traditional architectural planes (data, control and management). Generally, it has functions similar to traditional Layer 2. It gathers data (for safety, traffic information, infotainment), from a given area of interest, from all the sources (vehicle's internal sensors, GPS, inter-vehicle communication, Wireless Sensor Networks (WSN), or devices such as cellular phones, sensors and actuators, traffic lights and road signals located on streets and highways. Intra- and inter-vehicular interactions are within the scope of this layer. Various access technologies and associated protocols are supposed to perform the tasks. For intra-vehicle communication, the proposals are: Bluetooth (2.4 GHz), ZigBee (868 MHz, 915 MHz and 2.4 GHz), Wi-Fi HaLow (Low power, long range Wi-Fi, 900 MHz), Ultra-wideband (3.1–10.6 GHz), with data rates up to 480 Mbps and coverage distances up 1000m. For inter-vehicles communication technologies can be: IEEE WAVE/DSRC with IEEE 802.11p for PHY and MAC layers and the IEEE 1609 family for upper layers; 4G/LTE (1700 and 2100 MHz).

The *Data filtering* and *pre-processing* layer is necessary, given that IoV, may generate huge amounts of data, while not all are relevant for all entities of the system. This layer analyses and filters the collected information, to avoid the dissemination of irrelevant information and therefore reduces the overall network traffic. Examples of protocols to be used in this layer are:

- *Xtensible Messaging and Presentation Protocol* (XMPP)
- *Constrained Application Protocol* (CoAP)
- *HTTP Representational State Transfer* (HTTP REST)
- *Message Queuing Telematic Transport* (MQTT)
- *Lightweight Local Automation Protocol* (LLAP).

Several data filtering approaches are referenced in [8], but novel intelligent and efficient data mining techniques are considered to be necessary.

The *Communication* layer performs both data and control function at the networking level, given the set of protocols suggested as: 6LoWPAN, IPv4, IPv6, Routing Protocol for Low Power and Lossy Networks (RPL), etc. This layer should select the best network to send the information, based on several selection parameters (e.g., congestion level, QoS level capabilities over the different available networks, information relevance, privacy and security, cost, etc.). Apparently, the traditional networking functions are split between this layer and Acquisition layer.

The *Control and management* layer is the global coordinator that manages different network service providers within the IoV environment. Its functions are: to manage the data exchange among the various services; to manage the information generated by devices: in-vehicle or around sensors, roadside infrastructure and user devices in the environment; apply different suitable policies (e.g., traffic management and engineering, packet inspection, etc.). *It is not yet clear what intra and inter-domain management tasks has this entity*. The protocols proposed for this layer are: *CALM Service Layer*, WAVE 1609.6, TR-069, *Open Mobile Alliance Device Management* (OMA-DM).

The *Business* layer processes information using various types of cloud computing infrastructures available locally and remotely. Typical functions are: storing, processing and analysing info received from the other layers; making decisions based on data statistical analysis and identifying strategies that help in applying business models based on the usage of data in applications and the statistical analysis. (tools such as graphs, flowchart, critical analysis, etc.).

The *Security* layer (despite of its name - "layer") is an architectural plane, which communicates directly with the rest of the layers. It implements security functions (data

authentication, integrity, non-repudiation and confidentiality, access control, availability, etc.) to exchange data among sensors, actuators, user's devices through secure networks and service providers. The protocols envisaged are similar to those presented in Figure 2.

The work [8] proposes a split of the architecture in two planes: operational – containing six layers and the security plane, aiming to define the structure to include some of the current protocols in the different layers.

Considering the protocols proposed in [8] to be mapped on different layers it is apparent that the *Acquisition* layer is playing the role of access - given that access technology protocols are proposed there: Wi-Fi, 2G/3G/4G//LTE, Bluetooth, IEEE 1609/WAVE, IEEE 802.11p, WiMAX, etc. On the other side the *Communication* layer includes networking protocols like IPv4/IPv6RPL, ROLL, etc. Therefore, the two layers could have been merged as a *Access and Core Network* layer.

The cloud services are located at business level (as vehicular cloud computing) while we believe that a more natural placement could be as in Figure 2, i.e., under application layer.

Some mixture of "layers" and "plane" notions is apparent; there is a lack of enough orthogonality of different "layers". The architecture does not touch the integration of SDN/NFV approach.

Table I shows a comparison of the layered architectures exposed in this section.

TABLE I. LAYERED ARCHITECTURE COMPARISON

Layered Architecture	Criteria of comparison						
	No. of (macro) layers	***Target domain***	***Split in architectural planes***	***Mapping of protocols on architectural stack***	***Cloud computing included***	***SDN/NFV approach introduced***	***Edge/Fog computing approach introduced***
Bonomi et al. [5]	4	IoT/IoV	No	No	Yes (highest layer)	No	Yes
Kayvartya et al. [6]	5	IoV	Yes	Yes	Yes (middle layer)	No	No
F.Yang et al. [7]	4	IoV	No	No	Yes (middle layer)	No	Edge-only summary
Contreras-Castillo et al. [8]	6+1	IoV	Partial	Yes	Yes (highest layer)	No	No

IV. SDN CONTROLLED IOV ARCHITECTURES

Recent works emphasize the benefit of using novel technologies like SDN, NFV, cloud/fog/edge in the context of IoV. This section shortly presents samples of related work dedicated to VANET/IoV with SDN control.

Y.Lu et al. [17] apply SDN control to VANET, to get more flexibility, programmability and support for new services. The architectural components are:

- SDN controller
- SDN wireless nodes and
- SDN-enabled RSUs.

The SDN controller is a single entity performing the overall control of the system. The SDN wireless nodes are vehicles, considered as architectural Data plane elements (equivalent to SDN - forwarders). The SDN RSUs are also treated as Data plane elements, but they are stationary. The benefits of the approach are proved by simulation, while considering some specific use cases (e.g., routing). *However, a complete layered functional IoV architecture is not discussed.*

K.Zeng et al. [18] propose an IoV architecture called *software-defined heterogeneous vehicular network* (SERVICE), based on Cloud-RAN technology [15], able to support the dynamic nature of heterogeneous VANET functions and various applications. A multi-layer Cloud-RAN multi-domain is introduced, where resources can be exploited as needed for vehicle users. The system is hierarchically organized (three levels of clouds are defined: remote, local and micro clouds) and virtualization techniques (offering flexibility) are considered for implementation. The high-level design of the soft-defined HetVNET is presented. The SDN control is organized on two levels (one primary controller and several secondary controllers; each one of the latter controls a given service area). *A definition of a complete layered functional IoV architecture is not in the paper scope.*

A Fog-SDN architecture called FSDN is proposed for advanced VANET by Truong et al. [12], for V2V, V2I and Vehicle-to-Base Station communications. The Fog computing brings more capabilities for delay-sensitive and location-aware services. The SDN components (hierarchically top-down listed) are:

- *SDN Controller* (it controls the overall network behavior via OpenFlow –interfaces; it also performs Orchestration and Resource Management activities for the Fog nodes);
- *SDN RSU Controller* (RSUC) (controlled by the central SDN controller; each RSUC controls a cluster

of RSUs connected to it through broadband connections. The RSUC can forward data, and store local road system information or perform emergency services. From Fog perspective RSUCs are fog devices);

- *SDN RSU* (it is also a Fog device);
- *SDN Wireless Nodes* (vehicles acting as end-users and forwarding elements, equipped with OBU);

The system also contains *Cellular Base Station* (BS) performing traditional functions (they are also SDN-controlled via OpenFlow protocol and can also offer Fog services). *This study does not discuss a full functional layered IoV architecture*.

Kai et al. [19] present an overview of Fog–SDN solution for VANET and discuss several scenarios and issues. It is shown that a mixed architecture Fog-SDN (similar to that proposed in [12]) can be powerful and flexible enough, to serve future needs of IoV. *Again, we note that this study does not discuss a full functional layered IoV architecture.*

Chen et al. [20] discuss an IoV architecture and solutions based on SDN control. An SDN switched network is considered as a core network, controlled by SDN controllers. The vehicles are placed at the edges connected to the core via wireless data and control paths. The architectural planes are similar to those defined in SDN: data plane, control plane and application plane. To these a knowledge plane is added. *However, a full functional layered architecture with mapping of different protocols on this architecture is not discussed. Also, a fog/edge approach is missing.*

V. A SDN-FOG ENABLED IOV FUNCTIONAL ARCHITECTURE

This section proposes a layered functional IoV architecture of a heterogeneous network including SDN control and Fog computing capabilities. We propose a possible infrastructure (Figure 4), *which could be a horizontal extension of that proposed in* [12] for large network configurations (based on definition of regional service areas). Also, an enrichment of the five-layered architecture of [6] is proposed to introduce the functions of SDN control and also Fog computing.

The Data plane includes: mobile units (vehicles) equipped with OBUs; advanced RSUs, which could have enough resources (computing, storage) as to play also Fog node role (F-RSU) or could be regular RSU like in traditional VANETs; base stations (BS) of type WiMAX/3G/4G-LTE. Note that the BSes could also have fog-node capabilities (F-BS notation is used for such cases) A fixed network (partial mesh) can interconnect the RSUs.

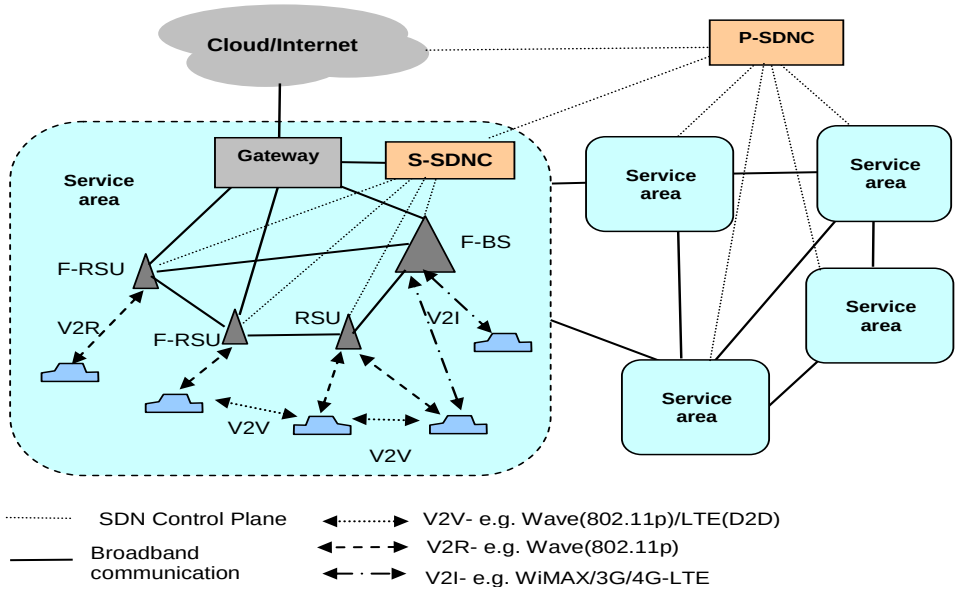


Figure 4. Generic IoV system architecture proposed in this paper.
F-BS - Fog-capable Base Station; F-RSU Fog-capable Remote Side Unit; P-SDNC- Primary SDN Controller; S-SDNC Secondary-SDN Controller; D2D- device to device communication

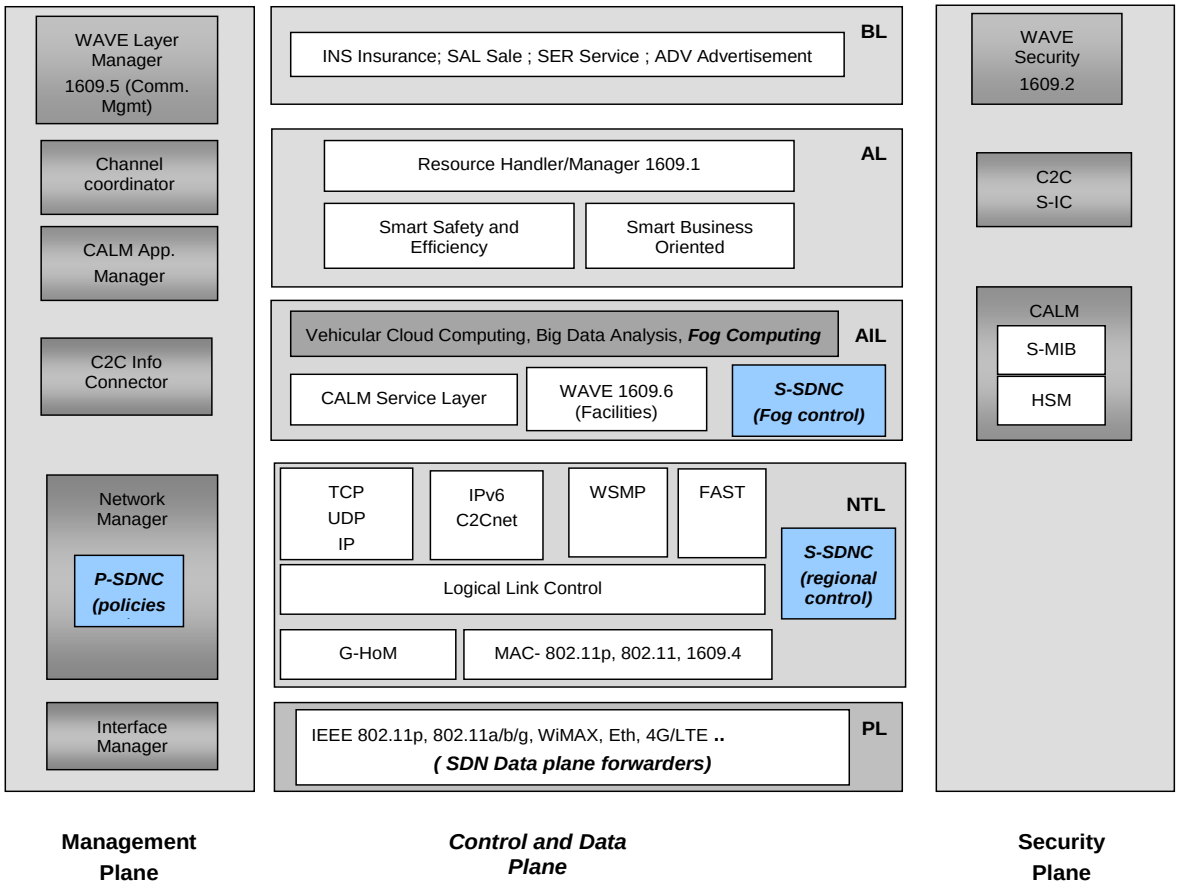


Figure 5. Functional IoV five-layer architecture enriched with SDN control and Fog computing capabilities (extended architecture of [6])

The SDN Data plane contains the forwarding nodes and can be geographically organized in several *service areas*.

The SDN Control plane is organized on two hierarchical levels: primary SDN controller (P-SDNC) controlling the overall behavior of the network and secondary controllers (S-SDNC), one for each service area. The S-SDNC can also contain the resource management functions of the Fog infrastructure. The P-SDNC is logically connected to each S-SDNC via the Control plane overlay-type or physical links. The SDN south interfaces between the controllers and the lower level can be supported by OpenFlow protocol or some other similar protocol. This infrastructure is enough general as to be considered as a candidate or IoV.

Figure 5 shows a proposal to enrich the layered functional architecture introduced by Kaiwartya et al. in [6], by adding SDN and Fog functionalities, supposing that the infrastructure is that of Figure 4. The second layer of the architecture is renamed in Network and Transport Layer (NTL), showing in a more explicit way the role of this layer. The Operation Plane is renamed in Control and Data Plane.

The functionalities of the P-SDNC can be embedded in the management plane, given that its role is to govern the overall network behavior (e.g., some overall policies can be coordinated by this module). The regional SDN control is placed naturally at NTL level as to control the SDN forwarders and also the functions of Fog nodes located in the access area. Additionally, S-SDNC functions can be included in the AIL, to serve this layer needs in terms of Fog AI resource control.

In a complete system the cloud-based services could be split between centralized Cloud computing and Fog nodes. How to manage this in an efficient way is not solved in this paper; it is for further study.

Note that a complete architectural definition for the above proposal is also for further study. While it can be mapped onto five-layer architecture described in Section III, no details are developed here on how the virtualization will be managed and how the off-loading actions are performed in order to preserve the service continuity when the vehicles are moving.

VI. MOBILE EDGE COMPUTING - BASED IOV ARCHITECTURE

Mobile Edge Computing (MEC) technology is an edge-oriented computing technology exposing all advantages of edge computing: low latency/response time, high bandwidth, location and context awareness, reduction in amount of data transferred from a terminal device to a centralized cloud data center and back, reduced round-trip time, etc. [21] [22]. The MEC cloud computing resources and storage spaces are placed at the edge of the vehicular access network (usually in Radio Access Netwok - RAN) and are in close proximity to the mobile terminal.

MEC is a distributed computing environment where applications can benefit from real-time radio and network information and can offer a personalized and contextualized experience to the mobile subscriber. The mobile-broadband experience is more responsive and opens up new monetization opportunities. This creates an ecosystem where new services are developed in and around the Base Station.

The key element is MEC application server, usually integrated in RAN, which can provide computing resources, storage capacity, connectivity, and access to user traffic and radio and network information. MEC offers an open radio network edge platform, supporting multi-service and multi-tenancy. Authorized third-parties may also to make use of the storage and processing capabilities, introducing new businesses on-demand and in a flexible manner.

The main standardization organization involved in MEEC is ETSI, which established in 2014 the Mobile Edge Computing Industry Specification Group. Recently, in 2017, the name MEC has been changed into Multi-access Edge Computing [23] - to better reflect non-cellular operator's requirements and fixed access case.

The general MEC architecture is presented in Figure 6 [22]. The mobile edge host level is the main MEC sub-system consisting of two main parts: the mobile edge host and the mobile edge host level management. The mobile edge host provides the virtualization infrastructure (based on Network Function Virtualisation Infrastructure –NFVI- coming from ETSI Network Function Virtualization -NFV framework) and the mobile edge platform, supporting the execution of mobile edge applications.

MEC has a good perspective as a supportive technology for vehicular communication (V2V, V2I, etc.) and IoV [24] [25]. Vehicles connected to the distributed edges may send/receive information from other vehicles or through the network almost in real-time. The mobility of vehicles is naturally supported by the RAN. However, not many publications in this area exist yet.

K.Zhang, et al., [24] developed a MEC-based model of a vehicular network. Their architecture comprises several levels: *Virtual Computation Resource Pool*- incorporating the network and cloud resources outside the MEC; MEC level – implemented as MEC servers placed in the RAN; RSUs units placed on the roads and mobile unites (vehicles). *This study does not offer a detailed layered architecture, neither aspect of implementation using SDN, NFV technologies, except the implicit use of NFV management and NFVI to realize the MEC architectural stack.*

The main focus of this work is on the computation off-loading process, to preserve the service continuity in the MEC environment. Due to their high mobility, vehicles in transit may pass through several RSUs and MEC servers during the task-off-loading process, and they can off-load their computation task to any MEC servers that they can access. Two methods are possible: selection of the target MEC servers or selecting (for a while) of a new path from the mobile vehicle to the same MEC server (keeping as much as possible the same serving MEC server in order to avoid too frequent moving of virtual machines).

J.Liu et al. [25] propose an SDN-enabled network architecture assisted by MEC, while integrating different types of access technologies.

The architectural components of the overall system are (top-down hierarchical list): Remote Data Center; Backbone network, Regions (MEC server + SDN controller, BS and mobiles organized in VANETs). The MEC servers can inter-communicate via a mesh of fixed network links.

The layered architecture [25] is less elaborated than that proposed in [6]. This one is SDN-like comprising three planes (Data, Control and Application) each including typical functions:

Data Plane (DPl): SDN- "switches" (VANET, BS, Ethernet); lower layer technologies (IEEE 802.11p, LTE/5G, Wire NIC, etc.).

Control Plane (CPl)

- lower sub-layer: Position/Channel sensing, Flow table management, Forwarding strategy;
- upper sublayer: Trajectory prediction, Interface sensing, Radio Resource control, Traffic redirection.

Application Plane (APl) (in the SDN semantics): Topology management, Resource Management, Traffic Offload, SDN controller.

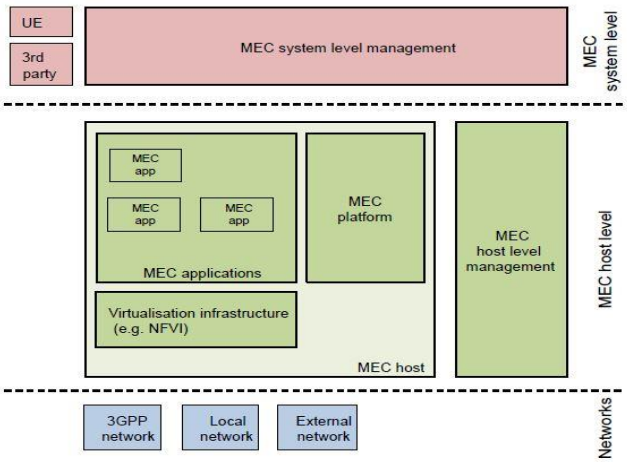


Figure 6. Simplified MEC architecture (ETSI)

The interface between CPl and DPl is based on extended OpenFlow or other similar protocol. *The limitation of this architectural proposal is that no mapping on SDN/NFV/ and fog/edge approach is discussed.*

VII. CONCLUSIONS AND FUTURE WORK

This paper presented a comparative critical view of several IoV architectures proposed in the literature, focused on functional layering aspects.

Among several proposals, we selected a five-layer multiple-plane architecture, considering this model as a good and orthogonal approach, which consistently include the major IoV functionalities and is giving the possibility to clearly define interfaces between layers and architectural planes. Another advantage is that the architecture is consistent with IoT architectural vision. Several examples based on SDN/Fog/Edge approaches are comparatively discussed, mainly from the point of view of layering the architecture and map different protocols on it.

In Section V, a modified Fog-SDN based IoV infrastructure is proposed by the authors, where the associated layered architecture is enriched by considering the additional Fog-based approach and SDN distributed control. This work could be a contribution towards an IoV reference architecture.

Section VI shortly present some MEC-based IoV systems, as an alternative to Fog-based approach. The comparative study of MEC/Fog alternatives also are topics for further work.

Future work should be done to allocate and map different functions of the general functional layered architecture to specific entities of a complete IoV system. This should be done based on their different roles and placement: terminals (vehicles), RSUs, Fog/Edge Nodes, BS, core network, cloud data centers, etc. The virtualization challenges and their impact on the architecture are not yet discussed in this study. This is also subject for further work.

REFERENCES

[1] E. Borcoci, S.G. Obreja, M. Vochin, "Internet of Vehicles Functional Architectures - Comparative Critical Study", The Ninth International Conference on Advances in Future Internet, AFIN 2017, September 10 - 14, 2017 - Rome, Italy, http://www.iaria.org/conferences2017/ProgramAFIN17.html, [Retrieved: December, 2017].

[2] ETSI EN 302 665 V1.1.0 European Standard Telecommunications series, "Intelligent Transport Systems (ITS);Communications Architecture" (2010-07).

[3] Y. Li, "An Overview of the DSRC/WAVE Technology", https://www.researchgate.net/publication/288643779, 2012, [Retrieved: May, 2017].

[4] S. Sultan, M. Moath Al-Doori, A.H. Al-Bayatti, and H. Zedan "A comprehensive survey on vehicular Ad Hoc Network", J.of Network and Computer Applications, Jan. 2014, https://www.researchgate.net/publication/259520963, [Retrieved: December, 2016].

[5] F. Bonomi, "The smart and connected vehicle and the Internet of Things", San José, CA: WSTS, 2013, https://tf.nist.gov/seminars/WSTS/PDFs/1-0_Cisco_FBonomi_ConnectedVehicles.pdf, [Retrieved: December, 2016].

[6] O. Kaiwartya, A.H. Abdullah, Y. Cao, A. Altameem, and M. Prasad, "Internet of Vehicles: Motivation, Layered Architecture, Network Model, Challenges, and Future Aspects" IEEE Access, Special Section on Future Networks, Architectures, Protocols and Applications, Vol. 4, pp.5536-5372, September 2016.

[7] F.Yang, J.Li, T. Lei, S. Wang, "Architecture and key technologies for Internet of Vehicles: a survey", Journal of Communications and Information Networks, Vol.2, No.2, Jun. 2017, DOI: 10.1007/s41650-017-0018-6

[8] J.C. Contreras-Castillo, et al., "A seven-layered model architecture for Internet of Vehicles",Journal of Information and Telecommunications , Vol. 1, No. 1, pp. 4–22, 2017.

[9] Y. Fangchun, W. Shangguang, L. Jinglin, L. Zhihan, and S. Qibo, "An overview of Internet of Vehicles", China Commun., vol. 11, no. 10, pp. 115, October 2014.

[10] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications", IEEE Communications Surveys &Tutorials Vol. 17, No. 4, pp.2347-2376, 2015.

[11] F. Bonomi, R. Milito, J. Zhu, and Sateesh Addepalli, "Fog Computing and Its Role in the Internet of Things", August 2012, http://conferences.sigcomm.org/sigcomm/2012/paper/mcc/p13.pdf, [Retrieved: January, 2017].

[12] N.N. Truong, G.M. Lee, and Y. Ghamri-Doudane, "Software defined networking-based vehicular ad hoc network with fog Computing", Proceedings of the 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM'15), May 2015, Ottawa, Canada. Piscataway, NJ, USA: IEEE, , pp. 1202–1207, 2015.

[13] B. N. Astuto, M. Mendonca, X.N. Nguyen, K. Obraczka, and T. Turletti, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks", Communications Surveys and Tutorials, IEEE Communications Society, (IEEE), 16 (3), pp. 1617 – 1634, 2014.

[14] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network Function Virtualisation: Challenges and Opportunities for Innovations". IEEE Communications Magazine, pp. 90-97, February 2015.

[15] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies", IEEE Wireless Communications, pp.126-135, December 2014.

[16] R. Khan, R. Zaheer, S. Khan, "Future Internet: The Internet of Things Architecture, Possible Applications, and Key Challenges", 10th International Conference, on Frontiers of Information Technology, 2012, https://www.academia.edu/35984783/Future_Internet_The_Internet_of_Things_Architecture_Possible_Applications_and_Key_Challenges?auto=download [Retrieved : Dec.2017]

[17] Y. Lu, M. Gerla, R. Gomes, and E. Cerqueira, "Towards software-defined VANET: Architecture and services", MedHocNet.2014.6849111, https://www.researchgate.net/publication/271472780, [Retrieved:April, 2017].

[18] K. Zeng, L. Hou, H. Meng, Q. Zheng, and N. Lu, "Soft-Defined Heterogeneous Vehicular Network: Architecture and Challenges", IEEE Network, vol. 30, pp.72-79, July/August 2016.

[19] K. Kai, W. Cong, and L. Tao, "Fog computing for vehicular Ad-hoc networks:paradigms, scenarios, and issues", The Journal of China Universities of Posts and

Telecommunications, www.sciencedirect.com/science/journal/10058885, http://jcupt.bupt.edu.cn, 23(2), pp.56–65, April 2016, [Retrieved: January, 2017].

[20] J. Chen, H. Zhou, N. Zhang, P. Yang, L. Gui, and X. Shen, ”Software defined Internet of vehicles: architecture, challenges and solutions”, Journal of Communications and Information Networks Vol. 1, Issue (1): pp.14-26, 2016, DOI: 10.11959/j.issn.2096-1081.2016.002.

[21] M. Patel et al., “Mobile-Edge Computing – Introductory Technical White Paper”, https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/Mobile-edge_Computing_-Introductory_Technical_White_Paper_V1%2018-09-14.pdf/01.01.01_60/gs_MEC001v010101p.pdf [Retrieved: December, 2017].

[22] “Mobile edge computing (MEC); Framework and reference architecture,” ETSI, Sophia Antipolis, France, Mar. 2016. Available: http://www.etsi.org/deliver/etsi_gs/MEC/001_099/003/01.01.01_60/gs_MEC003v010101p.pdf [Retrieved: January , 2018].

[23] T. Taleb,K. Samdanis, B. Mada, H. Flinck, S. Dutta, and Dario Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration", IEEE Comm. Surveys &Tutorials, Vol19, No.3, pp.1657-1681, 2017

[24] K. Zhang, Y. Mao, S. Leng,Y. He, and Y. Zhang, "Mobile-Edge Computing for Vehicular Networks- A Promising Network Paradigm with Predictive Off-Loading", IEEE vehicular technology magazine, pp.36-44, June 2017.

[25] J. Liu, J. Wan, B. Zeng, Q. Wang, H. Song, and M. Qiu, "A Scalable and Quick-Response Software Defined Vehicular Network Assisted by Mobile Edge Computing", IEEE Communications Magazine, pp.94-100, July 2017.

Influence of the Perceived Data Security, Operator Credibility and Provider Trust on Usage Frequency of Internet Services

Erik Massarczyk, Peter Winzer
Faculty of Design – Computer Science – Media
RheinMain University of Applied Sciences
Wiesbaden, Germany
Email: erik.massarczyk@hs-rm.de, peter.winzer@hs-rm.de

Abstract—**An increasing customer usage of Internet services with various devices demands a greater effort on data security, credibility and trust issues as due to extensive connections personal data are spread more widely. However, customers often prefer better services rather than higher data security. Here, the aim of this paper is to examine the positive influence of the perceived data security on the usage frequency of Internet services. The main target is to measure how the user (a) perceived data security, (b) perceived operator credibility and (c) perceived trust influence the usage frequency of Internet services. The named variables are analyzed with an adjusted conceptual model based on elements of the Unified Theory of Acceptance and Use of Technology 2. In general, a significant positive influence of a perceived data security on the usage frequency for specific services could be identified. Yet, the perceived trust in the service providers does not significantly relate to a stronger usage frequency of Internet services. Furthermore, for specific Internet services a positive relationship between the user perception about the operator credibility and the usage frequency of the Internet services could be proven. Consequently, customers have data security concerns and these might hinder them to use several Internet services.**

Keywords-data security; trust; credibility; usage frequency; Internet services.

I. INTRODUCTION

The following analysis illustrates the further approach of the already presented study from us regarding the analysis of influence factors of data security and trust on the actual customer usage of Internet services [1].

The growth of the number of Internet services and of the number of users lead to an increased amount of gained data. Especially services like (a) instant messaging, (b) social media, (c) video (broadcasting/streaming), (d) gaming, and (e) cloud computing are used by more and more people with more different devices [2][3][4]. Consequently, the degree of connection of the people and devices increases quite heavily [2]. Based on the growth of the number of connections and Internet services usages, users are generating more personal data that will be distributed to a greater extent [3].

From the customer point of view, it is difficult to comprehend to which extent personal data is collected, where the personal data is stored and which persons get access to the raised personal data for legal or illegal motives [4][5]. Due to the increased connectivity between the devices, unhindered individual communications and marketing measures, a wide range of information and personal data is disclosed. The data disclosure touches the security and privacy concerns of the customers because the personal information could include critical information and intellectual properties of the users. Furthermore, personal information are countable assets from which enterprises as wells as criminals may benefit [2][6][7]. Nonetheless, each user is responsible which data she or he releases for the usage of the specific Internet services and devices. Obviously, many people are willing to distribute their personal information to get a good performance of the used services. Here, they often do not care about risks of data leakages and data misuse.

The rising number of security incidents shows that criminals more frequently attack enterprises, administrations and private customers to get the personal data because they have identified the values of these personal information and intellectual properties [7]. As a result, customers should care more about possible data privacy and security concerns, while using Internet services.

Consequently, we have examined if the private customers perceive any concerns about their own (a) data privacy and (b) data security, and if they have trust and credibility concerns about their application/software providers and their network operators when they use different Internet services with various devices. This study focuses also on the different conditions and usage opportunities between wired and wireless infrastructures and connections, where different types of data security problems could arise. In this respect, we want to measure the status and the perception of data security, while customers using the following services: (a) email, (b) social media, (c) internet telephony, (d) online shopping, (e) cloud computing, (f) e-learning, (g) instant messaging, (h) online banking, (i) navigation, (j) online administration, (j) video on demand, and (k) internet television. Additionally, we analyze the customer evaluation of the credibility in network operators and the trust of the

providers of the named Internet services. Here, it will be measured how the customers perceive that the network operators keep the private data of the customers (credibility) and to what extent the providers of the Internet services (trust) in general further distribute their personal data. Due to customers use the named Internet services differently in the wired and wireless networks, we separate the results in the two named considerations. On the hand, we consider the perceptions in the fixed/wired infrastructure environment and on the other hand, the results in the mobile/wireless infrastructure environment are presented. In the further consideration of the results and discussion, we present the similarities and differences in the usage frequency of the services between the both networks concerning the influence of data security, trust and credibility issues. The also retrieved perceived importance of data security will not be in the major consideration of this study.

In Section II, (a) the term data security, (b) the challenges, (c) the known literature as well as the used conceptual models and research models will be described. Following this section, the methodology, as well as the theoretical approach for carrying out the analysis, will be briefly explained. In Section IV, the results of the hypothesis tests are presented. Finally, in Section V, a critical discussion of the results takes place and a further view on the ongoing research will be done.

II. Literature Review

A. Data Security

In general, the term "data security" describes the secure management of personal data, secure data transmission and the transparency which institutions or persons have access to the personal customer data [6][8]. The correct implementation of data security usually involves that the customers themselves decide who is entitled to access their data. As mentioned in the introduction, customers often ignore possible risks of sharing information and they are not aware of the amount of data, which they produce and which are the consequences if the personal data would be leaked [9]-[11]. The ignorance shows critical issues in three dimensions. Firstly, customers spread personal data which could be linked to confidential information like bank accounts and credit card numbers [9][10]. Secondly, many companies use and transmit – without permission and knowledge of the customers – private customer information, which the customers disclose during the usage of Internet services [12]. Thirdly, as already mentioned, the number of Internet security incidents – like criminal acts of password capturing, eavesdropping and blackmails – have increased quite heavily during the last couple of years [4][7].

B. Challenges

Yet, the perceptions of (a) data security, (b) trust, (c) credibility, (d) sharing of information and (e) risks differs between the individual customers and depend beside others on factors like demography and culture [9]. Therefore, the user attitudes and beliefs are completely subjective and the people have different perceptions about possible risks and prevention of risks [9]. We assume that most of the customers prefer a good Internet service performance instead of strong security or data protection measures. Here, customers frequently do not care about the consequences of misuse and data leakage. Moreover, the providers of Internet services often do not state information about consequences of misuse and data leakage and do not insert different messages to make sure that the users understand the impacts of their data distribution. To increase the customer caution concerning data disclosure, also the providers should implement several measures, which the customers have to comply with to use the services [13].

Especially these issues motivate us to investigate which factors directly influence the usage frequency of Internet services and the individual perception of data security, credibility and trust.

C. Research Model – Adjusted Model with Elements of the Unified Theory of Acceptance and Use of Technology 2

The main target of this study will be to get an increased comprehension of private customer behaviors, especially in the focus on data security, credibility and trust concerns regarding the acceptance and actual usage of services.

The Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) is the direct expansion of the known UTAUT concepts with the factors hedonic motivation, price, and habit/experience, which allows a broader consideration of critical influence factors on user behavior [14]-[17].

Nevertheless, perceived data security, perceived credibility and perceived trust could not be covered by the existing variables of UTAUT2. Nonetheless, an implementation of external variables as influence factors of the user behavior could be performed. By the approach of Escorbar-Rodriguez and Carvajal-Trujillo, the UTAUT2 model is expanded by external variables trust as well as the further components perceived security and perceived privacy [14][18]. This expansion makes clear that the influence of security measures and perceptions on the behavioral intention to use of an innovation can be investigated [14][18]. Furthermore, this approach motivates us to use the factors perceived data security, perceived credibility and perceived trust as external variables in the own adapted model (see Figure 1) [18]. The single analysis of each relation between the named variables with the usage frequency features here the first step of the upcoming regression analyses. As it can be seen in Fig. 1, we also combine the different variables in one regression analysis to figure out how the different variables also affects each other. Each of the analyses will be prepared for each named Internet service, which were already introduced in the first section of this study. The adapted model keeps therefore only the basic idea of the UTAUT2.

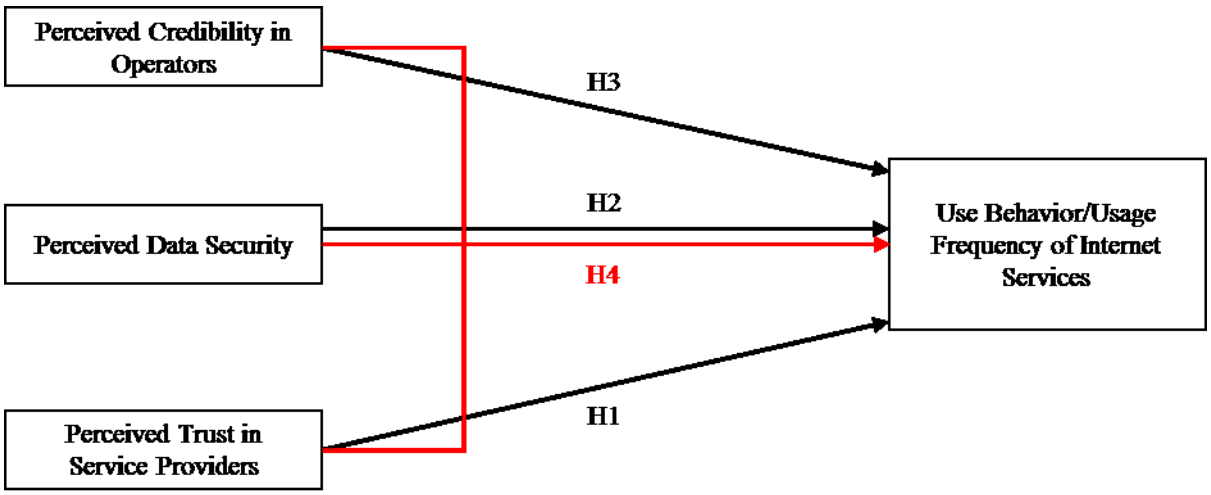


Figure 1. Conceptual Model

As mentioned in the introduction, in comparison to the already presented pre-study, in the further analysis the external variable of perceived credibility of the network operators will be added and further examined [2].

Consequently, we want to measure directly the impact of the perceived data security measures on the actual usage of Internet services (instead of testing the relationship with the behavioral intention to use, as Zhong et al. already did [19]). In other words: The target of investigation is to analyze whether perceived data security, credibility and trust issues lead to a utilization of an Internet service and how the degree of utilization is influenced.

Generally, we estimate that the perception about the increased safety and security of their own data would lead to an increased usage of services. The conceptual model in Fig. 1 bases on the original model of the first presented approach [1]. To support the measurement of the component data security by the customers, the password changing behavior of customers (which is an indicator for the importance of data security) and their relation to the importance of data security will be considered as additional factors. The password changing behavior is therefore relevant to measure if customers with a higher awareness of data security will change their passwords more regularly. We also believe that an outrageous assessment (a) of the proper management of data by network operators and (b) the trustworthiness of service providers would lead to increased use of Internet services.

Perceived credibility describes the users' belief that the used systems would be free of privacy and security threats. In addition, the credibility covers how the customers estimate and perceive the safe storage of their personal information [17].

Customers recognize the behavior of providers if they take care about the personal information and secure transmissions and therefore, if the customers estimate that their personal data are safe [19]-[29]. In case, the operators do not take care about their system, customers perceive the infrastructure as insecure and would avoid to use it.

However, here we definitely want to admit that this point is a pure assumption.

In daily life, people often log into open wireless networks to connect with the Internet and they do not take care or they do not realize the possible threat of data disclosures and linkages. Nevertheless, it could be also the point that the people in the daily life perceive that these infrastructures are secure. Due to these reasons, it is necessary to investigate how the people estimate the security and trustworthiness of their network operators. Furthermore, the customer estimation of the network operators enables a possible assessment of enterprises' trustworthiness and adhere the accepted rules from customer perspective [30]. For this reason, the used survey also includes questions about how the customers perceive the security of the infrastructure and how the network operators use the gained data from the customer. Therefore, we link directly the perceived credibility as positive impact factor for the actual usage of a service, due to the link between the component perceived credibility and a possible behavioral intention to use is already known [18].

Due to the fact that using Internet services (especially mobile services) include security and privacy threats [20], we implement the factor trust. The perception of trust describes how credible the customers perceive the provider [2][18][31][32][33]. The perceived trust of the service providers would be characterized by the fact how the customers estimate the reliability of the service providers in the distribution of personal information to third parties. Based on the assumption that risks and perceived trust directly influence the usage processes [34], the customers would reduce their usage if they expect a loss of privacy and a higher risk in usage [2][35][36][37]. The particular importance of the key factors of risk and trust lies in the fact that these two factors have a major influence on the customer acceptance of innovations (especially mobile payments,

mobile banking and mobile shopping) [19][20][38]-[41]. The trust variable is needed to cover the risk and privacy concerns of the customers and it can be used to figure out how the customers perceive the credible and secure information and experiences of the providers [18][31][32][33]. In addition, trust in a service or in a service provider plays an important role for the customer, since this increases the customer's sense of satisfaction in the service and thus leads to a higher usage frequency [34][42].

Finally, non-existent trust, credibility or the perception of missing security negatively influences customer behavior. An increase in security while using a service would give the customers a more confident secured and satisfied emotion and could possibly imply a stronger usage of this service. Here, previous researches identified that the perceived risk can be seen as one of the key drivers for the estimation of uncertainties in mobile payments, mobile shopping, mobile banking and mobile transactions [35][19][38]-[41], because customers fear possibilities for attacks by mobile transactions and the lack of control. Consequently, customers are paying attention to the products and their providers if they take care about the customers' transactions and personal information security in the usage of mobile payments [19].

Besides mobile banking, mobile payments and mobile shopping, the perceived security issues are even higher in mobile telecommunication networks, due to the fact that mobile networks are shared mediums and different persons can use the same mobile radio cell in the same time. This structure makes the system more vulnerable for attacks within the network [43]. The mobile network operators and providers have to take care about these issues and the introduction of security measures can mitigate uncertainties and risks [44]-[47]. The development of trust in a service is a major aim for customers and providers, because the trust in a service increase the customer convenience and normally leads to a higher performance [34][36][42].

Consequently, the literature conveys the feedback that in several cases trust and the perception of security, risks and uncertainties influence the customer user behavior. All examples and findings demonstrate that possible security issues can significantly negatively influence the intention to use mobile Internet services and therefore, the actual usage frequency could be reduced. Therefore, the authors have set the hypotheses that a better perception in data security leads to an increased usage of services.

Generally, the aim of the analysis is to illustrate how the customers perceive their data security and how they rank the credibility and trust for each Internet service they use. In comparison to the presented relationships between the estimation of perceived security, risks and trust and the usage of mobile banking or mobile payments [21][22], we consider different Internet services for the analysis of the relation between (a) the perceived data security, credibility as well as trust and (b) the usage of services.

Based on these explanations, the hypotheses for this research paper are:

H1: The customer perception of data security has a directly positive effect on the usage of Internet services.

H2: An increased perceived provider trust has a directly positive effect on the usage of Internet services.

H3: An increased perceived network operator credibility has a directly positive effect on the usage of Internet services.

H4: An increased combined perception of trust, credibility and data security has a directly positive effect on the usage of Internet services.

III. METHODOLOGY

The hypotheses are validated based on a current survey. The answers were taken by interviewers in personal interviews, thus ensuring completeness and accuracy of the answers. The respondents were randomly chosen and asked if they wanted to answer the questionnaire. The interviewers were instructed to choose the interviewees as far as possible randomly to make sure to get a sample which represent the demographic characteristics of gender and age of the local population [48][49]. Generally, test persons are asked in December 2016 at public libraries in Wiesbaden (which is a city with approx. 290,000 inhabitants in the middle of Germany) to reach a diversified and representative selection of test persons. In total, the survey includes 290 completed questionnaires. In the first part of the survey, the respondents were asked about their provider and contracts. The second part include questions about the general usage and usage frequency of different Internet services. The subsequent third part of the survey covers the questions on the customer password changing behavior, importance of data security and perceived data security. Questions about perceptions of trust and credibility of providers and operators, as well as on age and gender conclude the survey.

The collected data has been examined based on quantitative research methods with the statistical program SPSS. To evaluate the reliability and validity of the obtained data, Cronbach Alpha was determined and an Exploratory Factor Analysis was performed.

The perceived data security was queried with the question, how the customers perceive their personal data for each specific Internet service in the usage of a fixed and/or mobile Internet access (5-Point-Likert-scale: very secure to very unsecure). For the measurement of the usage frequency of (mobile) Internet services, a 5-Point-Likert-scale (very often to very few) has been used [50]. The perceived credibility (5-Point-Likert-scale: very certain to very uncertain) targets on the question how the customers perceive that the used broadband infrastructure is free of threats. Finally, the trust (5-Point-Likert-scale: very open to very closed infrastructure) is measured by the question whether the users perceive that the Internet service providers spread their personal data without authorization.

As mentioned above, the used approach only keeps elements of the UTAUT2. Therefore, we do not follow the analysis with a Structural Equation Modeling. Instead, we use the ordinary least square regressions to test the significance of each of the named hypotheses [14][15]. In the following combined approach under recognizing and controlling of further variables like importance of data security, perceived credibility and password changing behavior, we use a combined regression analysis.

IV. Data Analysis and Results

A. Result Conditions

The following discussion assumes far predominantly that the participants of the survey answer as private customers, even if it cannot be completely excluded that some of the respondents may also answer from their perspective of personal small enterprises.

We will describe the results of the reliability and validity tests of the overall used hypotheses briefly. After this testing, the regression results of hypotheses will be prioritized to figure out the relationships between (a) perceived data security as well as perceived trust in the service providers and the perceived credibility of network operators and (b) the usage of specific Internet services.

B. Descriptive Results

The results of the second survey presented here extend the original first survey conducted in 2016 and cover 7 additional sets of questions. 55.0% of the respondents are male and the average age of a respondent is between 30 and 39 years. With 48.1%, the group of the 20 and 29-year-olds has the largest share of respondents. Thus, this age group (which is 12.2% of the total population in Germany) is overrepresented in the survey by a factor of four [51]. Based on a study of ARD/ZDF from 2015 the 20 to 29-year-old are nearly 100% Internet users [52].

The over-representation in younger age groups naturally leads to an under-representation of the elder age groups. Consequently, the collected data are not representative. Although there is no representativeness in terms of age and gender, the answers from 290 participants provide much information about the usage behavior of the various communities and thus some conclusions can be drawn on frequency of use and usage preferences. 26.5% of respondents feel confident about their data, but on the contrary, 32.9% of respondents feel more or less insecure about their data. Interestingly, the one third of respondents, who feel insecure in their data security, does not fit at all with the results of the password changing behavior of the customers, since more than 80% of the customers change their passwords much less frequently than once a year: For email accounts 84.3% and for social media accounts 89.2%. Normally, it would be expected that more people change their passwords more regularly if they feel a data insecurity. In this respect, it can be stated as the first conclusion that the perception of the data security does not affect the frequency of the password changes. The reason for this could be that customers are distributing their own data. Therefore, a higher password security increases the overall security, but has no influence the perceived data security. Furthermore, users consider the changing of passwords as cumbersome and not user-friendly.

89.0% of respondents use an anti-virus program, which fit with the quotas of 85.5%, which are also confirmed by studies of the software company McAfee [53]. Most customers associate the use of anti-virus programs with a general increase in data security and often fail to recognize that such programs can only protect the hardware and software from systematic attacks by viruses and malware. However, antivirus programs by their nature do not provide protection against human errors (such as insecure passwords) and thus can only improve data security to a very limited extent.

In average, the customers believe that fixed Internet providers have a little bit safer infrastructure than mobile Internet providers do. Email services are the mostly used services overall (round about 80%). In the fixed infrastructures, about 3/4 of the customers use online shopping, video on demand and online banking (independent from the usage frequency). In the consideration of mobile devices and mobile infrastructures, about 4/5 of the customers use instant messaging.

TABLE I. IMPORTANCE OF DATA SECURITY

Internet Services	Importance of Data Security
Email	54.9% very high importance
Social Media	31.9% very high importance
Online Shopping	53.6% very high importance
Online Banking	75.9% very high importance
Instant Messaging	47.4% very high importance

TABLE II. USAGE FREQUENCY
(Only voting for the highest level of the usage rate)

Internet Services	Usage Frequency
Email	35.1% very frequently
Social Media	43.8% very frequently
Online Shopping	4.9% very frequently
Online Banking	6.9% very frequently
Instant Messaging	63.0% very frequently

TABLE III. TRUST IN SERVICE PROVIDERS

Internet Services	Trust in Data Usage – closed	Trust in Data Usage – open
Email	15.7% very closed	3.9% very open
Social Media	2.1% very closed	21.1% very open
Online Shopping	5.1% very closed	14.1% very open
Online Banking	45.0% very closed	1.7% very open
Instant Messaging	4.6% very closed	14.9% very open

The tables show for the different services: (I) the importance of data security, (II) the usage frequency of Internet services, and (III) confidence in service providers. Interestingly, customers in the services they use very frequently (social media and instant messaging) feel a relatively low data security. Customers also recognize that the providers of these services do not particularly secure the customer data and use it for their own purposes. In opposite, the usage of online banking is relatively rare, but data security is very important to customers in this area, which is, of course, mainly due to the nature of the service and is presumably independent of the channel through which this financial service is provided.

In the consideration of the differences in the perceptions of data security, credibility, trust and usage frequencies of Internet services between women and men, no general diversity can be concluded. However, in the consideration of the difference in means, it can be said that women perceive a lesser degree of data security (women 2.79; men 3.06) and life security (women 3.21; men 3.57) than men do. Despite men feel secure in their life situation, women and men perceive just a neutral data security. This comparison also shows that the general estimation of the data security is less than the general perception about the life security. Therefore, data security worries the respondents and lead to a higher degree of uncertainty.

The view on the data security perceptions of women and men present for the single Internet services presents no difference in means and therefore, we estimate that men and women perceive and treat data security similar. Just in the services online banking and instant messaging, women and men do not perceive similarly in average. Despite men and women estimate secure online banking infrastructures, men estimate a higher data security than women (women 3.57; men 3.98). Contrary, men and women perceive an averagely data security, women have a higher perception of data security for instant messaging services (women 3.13; men 2.67). Interestingly, the credibility of the fixed providers illustrates no difference between women and men. However, the two groups have estimated the credibility of mobile providers quite differently, because women perceive the mobile network operators quite less trustworthy than men.

Table IV illustrates the differences in the usage frequencies between women and men, which presents in general that men use more often Internet services than women do. Due to the significant (below $p<0.05$) F-Ratios about the mark of 1 and 3, the values present a good model fit and describe significant differences in the usage rates of the presented Internet services. The analysis of the means for the other Internet services, which are not included in Table IV, have no significant differences in means.

C. Reliability and Validity

Reliability is a measure of the formal accuracy of surveys/scientific measurements. It is that part of the variance, which can be explained by differences in the characteristic to be measured and not by (measurement) errors. Reliable results must be mainly free of random errors (i.e. reproducibility of results under the same conditions).

The results of the reliability and validity analyses are illustrated in the Tables V and VI. In general, this study includes the following 8 aspects: (1) usage of Internet services (fixed networks), (2) usage of Internet services (mobile networks), (3) usage frequency of Internet services, (4) perceived importance of data security, (5) perceived data security (fixed networks), (6) perceived data security (mobile networks), (7) perceived trust, and (8) perceived credibility.

Generally, all named concepts are examined in the terms of reliability and validity. Following Cronbach, Alpha values must be higher than 0.7 to for a good reliability [54][55][56]. Based on the results in Table V, the collected data for the 7 named aspects are reliable.

TABLE IV. MEAN ANALYSIS FOR USAGE FREQUENCY AND GENDER

Service: Usage Frequency		Mean	F-Ratio	p-significance
Email	female	**3.47**	15.411	**.000**
	male	**4.03**		
Video on Demand	female	**3.35**	4.795	**.029**
	male	**3.69**		
Social Media	female	**3.47**	9.185	**.003**
	male	**4.03**		
Online Banking	female	**2.66**	9.742	**.002**
	male	**3.13**		
Cloud Computing	female	**2.25**	10.611	**.001**
	male	**2.81**		
Instant Messaging	female	**3.95**	6.778	**.010**
	male	**4.38**		
IPTV	female	**3.15**	5.783	**.017**
	male	**3.52**		
Navigation	female	**2.58**	9.064	**.003**
	male	**2.99**		

After the testing of the reliability, the exploratory factor analysis includes the assessment of Kaiser-Meyer-Olkin criterion (KMO), the significance test from Bartlett, and the examination of the cumulative variance to evaluate the validity of the collected data [57]-[61]. Validity considers the consistency of an empirical measurement with the based conceptual/logical measurement concept. To reach a good validity, the concepts should reach significant p values

($p<0.05$) in the Bartlett-Test and KMO values above 0.7 [57]-[61].

Table VI shows good validity scores for the collected data/aspects. The good validity scores are also supported by the results of the cumulative variances higher than 50%, which indicate high explanation rates of the variances of the collected data [58]-[60]. Consequently, the reliability and validity of the collected data are proved. Despite the above-mentioned non-existent representativeness of the collected data, the considered research concepts and scientific questions illustrate that the data could be used for further evaluations.

TABLE V. RELIABILITY ANALYSIS

Research Concepts	Cronbach's Alpha
Usage of Internet Services (fixed networks)	0.780
Usage of Internet Services (mobile networks)	0.784
Usage Frequency of Internet Services	0.803
Perceived Importance of Data Security	0.925
Perceived Data Security (in fixed infrastructures)	0.881
Perceived Data Security (in mobile infrastructures)	0.915
Perceived Trust	0.871
Perceived Credibility	0.772

TABLE VI. VALIDITY ANALYSIS

Research Concepts	KMO	Bartlett-Test	Cumulative Variance
Usage of Internet Services (fixed)	0.825	$p < 0.000$	50.397%
Usage of Internet Services (mobile)	0.804	$p < 0.000$	51.240%
Usage Frequency of Internet Services	0.781	$p < 0.000$	53.724%
Perceived Importance of Data Security	0.901	$p < 0.000$	64.709%
Perceived Data Security (fixed)	0.844	$p < 0.000$	57.791%
Perceived Data Security (mobile)	0.831	$p < 0.000$	62.055%
Perceived Trust	0.827	$p < 0.000$	59.372%
Perceived Credibility	0.709	$p < 0.000$	55.107%

D. Mean Analysis

Next to the reliability and validity analysis and the consideration of the different concepts, as already mentioned in the descriptive results, the survey also covers questions about the usage of anti-virus programs. Due to the nominal coding of the variables of the usage of anti-virus programs, we consider the results based on mean analyses to figure out if the usage of Internet services and the perception of data security differs if the people use anti-virus programs or not. The general view would be that anti-virus programs were normally implemented to create more security for the used systems and that possible threats would be detected and eliminated.

Generally, the (non-)utilization of anti-virus programs does not lead to a significant difference in the perception of data security and the perception of the importance of data security. In addition, if the users utilize an anti-virus program or they refuse to use it, this does not lead to differences in means in the password changing behavior.

The considerations of the decision to use Internet services show normally no differences in mean between people who already use anti-virus programs and people who do not use these services. However, the mean analysis with the help of one factorial analysis of variance (ANOVA) presents a significant difference in means in the usage of email services. Only 63% of the people, who does not use anti-virus programs, utilizes email services. In comparison, 78% of people who have installed anti-virus programs use email services. However, when we add the consideration of the usage frequency under the presented circumstances, there is no significant difference in the usage of email services if the people have an anti-virus program or if they do not have an anti-virus program. Although Internet users want to have a high data security for email services, we cannot finally conclude that the utilization of an anti-virus program will bring a higher security and therefore, the people use more email services. The reason is here that other critical Internet services like online banking or cloud services, which also cover secret details of the private customers, are not influenced by the utilization or non-utilization of anti-virus programs. Therefore, a general impact of anti-virus programs on the perception of data security and the perception of importance of data security cannot be concluded.

The further considerations and analyses of the relations between the named concepts and the conceptual model will be described in the next sub-section.

E. Regression Analyses

As mentioned above, the scope of the study does not allow the testing of all hypotheses.

In the following, at least, the relationship between the factors (a) perceived data security, (b) perceived credibility, (c) perceived trust and the usage frequency of Internet services would be analyzed by means of ordinary least square regressions. The perceived data security is analyzed differently for fixed and mobile Internet services. This differentiation takes account of the fact that the various network/service types have different advantages and disadvantages, and therefore, different utilizations can be expected.

The multiple regression analyses include on the on hand the degree of dependence and on the other hand the grad of the linear relationship (correlation analysis). Independently, if the focus is on the correlation or regression coefficients, in both considerations a 'perfect' relation is expressed by the value 1.000. Nevertheless, correlation/regression coefficients higher than 0.500 symbolize a good interrelation. [60]-[62]

Following the named regression analyses, we combined all possible influence factors of security issues, which have been collected in the survey to analyze their impact on the usage of Internet services.

For this purpose, the perceived data security (= independent variable) is analyzed separately for mobile and fixed broadband infrastructures/services) in relation to the usage frequency of the individual Internet services (= dependent variables); see Table VII.

The r-square values of the individual regressions are quite low, which is mainly due to two causes. On the one hand, only the effects of perceived data security are analyzed for the usage frequency of each service. In each individual case, an r-square for the regression between only an independent variable and a dependent variable is determined. As far as it is assumed, the individual r-squares are not quite as high. On the other hand, the usage frequency of an Internet service does not depend solely on the perceived data security. Based on the estimation of many different influencing factors (some are mentioned in the presented research model), we assume weak regressions, which mean relatively low r-squares.

For the usage of the following services in the fixed and mobile infrastructures, (a) Internet protocol television (IPTV), (b) instant messaging, and (c) online gaming, the customer data security perception does not impact the usage of these services; therefore, the hypothesis H1 cannot be accepted. For the services e-learning and cloud computing, significant positive regression relations could be found for both infrastructures (fixed and mobile). This means that a customer perceives a higher data security in his learning application, he will use the service more frequently. The coefficients of 0.286 (fixed) and 0.370 (mobile) show a quite moderate explanatory rate. As mentioned above, the r-squares of 3.3% (fixed) and 5.7% (mobile) are quite low and describe only a low coefficient of determination. In addition, if customers perceive a higher data security when they use cloud services then they will use them more frequently. Coefficients of 0.330 (fixed) and 0.232 (mobile) and r-squares of 5.8% (fixed) and 2.8% (mobile) shows a moderate explanatory rate and low degree of determination [60]-[62]. For these both services, we do not assume differences in the usage of the services in the both infrastructures and the hypothesis H1 could be accepted.

The analysis of other services (online shopping, online banking, e-mail, social media, and online telephony) shows differences in the results of the regression analyses between mobile or fixed infrastructures. The main reason for differences is the general use of services. Internet users use navigation and social media services twice as frequently by mobile devices in mobile infrastructures in comparison to the fixed-line connections. In contrast, online banking services are used much more frequently via fixed broadband infrastructures.

The perceived data security has only a relatively small (but measurable) influence on the use of navigation services with mobile devices/networks, with regression of 0.161 and r-square of 2.1%. This may be due to the fact that the primary goal of most users of a navigation service is to locate a destination and it is self-evident to them that they may have to make concessions for data security (for example, by authorizing the location).

For fixed networks, positively significant regressions between the perceived data security for (a) emails respectively (b) online banking and the usage of these services could be identified. Despite low r-squares of 5.8% (email) and 5.5% (online banking) and weak regressions, the single coefficients of 0.357 (email) and 0.295 (online banking) represent a moderate regression [60]-[62]. Since emails and, in particular, bank accounts generally contain highly sensitive data from customers, the loss of which can cause considerable damage, customers' need for high data security for these services is particularly high. If the users perceive a better data security for these services, or if the service providers can guarantee their customers a higher data security, they will use these services more frequently.

In addition, email services are often used in professional contact and can contain corresponding confidential information [9][10].

In general, the test of multicollinearities with the Variance Inflation Factor (VIF) shows that all VIF values are below 10 (mostly below 3) and therefore, multicollinearities not exist [57][63][64]. Nonetheless, in some cases, the constants are also significant ($p<0.05$), which could be an indicator for other influence factors or an existing endogeneity. In the further research and examination of the data, we will consider the influence factors and try to figure out which are the indicators for the significant constants.

The relationship of the perceived trust in the service providers (independent variable) and the usage frequency of Internet services (dependent variable) generally show no significant relationship for the specific services. The only exception is online shopping. The positive significant relationship (coefficient = 0.117) shows that customers, who perceive that the shopping providers do not further distribute their personal information, will more frequently use these online shopping platforms.

However, the r-square of 1.7% and the coefficient below 0.200 do not imply a good explanatory rate and the regressive connection seems to be weak [60]-[62]. Generally, the hypothesis H2 about the influence of the customer perception of trust in the service providers of the single specific Internet services on the usage frequency of the specific services cannot be accepted. It seems that trust, as single factor does not have an influence on the customer decision of service usage.

TABLE VII. REGRESSION ANALYSIS – COMPARISON PERCEIVED DATA SECURITY AS INFLUENCE FACTOR FOR USAGE FREQUENCY (single service consideration)

Dependent variables	Independent: Perceived Data Security in Fixed Networks			Independent: Perceived Data Security in Mobile Networks		
	Regression Coefficient B	**Significance**	**R-Square**	**Regression Coefficient B**	**Significance**	**R-Square**
Usage Frequency of Email Services	0.357*	p<0.05	5.8%	No Significance		
Usage Frequency of Cloud Computing Services	0.330*	p<0.05	5.8%	0.232*	p<0.05	2.8%
Usage Frequency of Online Banking Services	0.295*	p<0.05	5.5%	No Significance		
Usage Frequency of E-Learning Services	0.286*	p<0.05	3.3%	0.370*	p<0.05	5.7%
Usage Frequency of Instant Messaging Services	No Significance			No Significance		
Usage Frequency of IPTV Services	No Significance			No Significance		
Usage Frequency of Navigation Services	No Significance			0.161*	p<0.05	2.1%
Usage Frequency of Social Media Services	No Significance			No Significance		
Usage Frequency of Online Gaming Services	No Significance			No Significance		
Usage Frequency of Online Administration Services	0.393*	p<0.05	6.7%	No Significance		
Usage Frequency of Online Shopping Services	No Significance			0.142	p<0.05	1.8%
Usage Frequency of Online Telephony Services	0.228*	p<0.05	2.1%	No Significance		

* The regression presents a significant constant, which could be an indicator for further unconsidered variables or an existing endogeneity, which needs further investigation. Furthermore, the Durban-Watson-Test recognizes a value, which could be an indicator for an existing autocorrelation. To cover the spurious correlations, further investigations must be performed.

A combined regression analysis approach is carried out with the consideration of all of the single data security factors as independent variables in order to analyze the influence the whole impact of perceived data security on the frequency of the user behavior of the specific Internet services. The following variables are controlled: (a) overall perceived data security (in general without any consideration of a single service), (b) perceived importance of data security, (c) perceived credibility of the network operators, and (d) perceived trust in the service providers. The regression analyses for each individual service are carried out separately and shown according to the use of mobile or fixed network services.

The control of the variables that cover security issues (except perceived data security) reveals significant regressive influences of perceived data security on the usage frequency of the specific Internet services (email, cloud computing, online banking and e-learning); see Table VIII.

The control of the variables confirms the results obtained in the first point. When customers use email services over the fixed networks and they feel confident about their data, they will use the data more frequently. Although nearly 80% of the customers use email services over the mobile networks, no significant connection could be found. Despite the non-significance for mobile networks, the regression coefficient of 0.363 for fixed networks shows a moderate explanatory rate [60]-[62]. However, the r-square of 9.7% describes only weak regression with a low coefficient of determination [60]-[62]. The VIF is below 3, so multicollinearities can be excluded [57][63][64]. It seems that customers who experience more data security when using email services will use these services more frequently. This is mainly because customers have stored many confidential information in their email accounts and do not want third parties to have access to these data.

A similar relationship exists for cloud computing: when customers perceive higher data security for cloud computing services, they will use these services more frequently (significantly positive). Despite a moderate regression coefficient of 0.261, the r-square of 8.2% shows a weak regression. The VIF under 3 allows the exclusion of multicollinearities [57][63][64].

The third line of Table VIII represents the influence of the perceived data security on the usage of online banking. It can be identified (for mobile and fixed networks) that users, who have security issues with online banking, do not use online banking. The regression coefficients of 0.218 (fixed) and 0.352 (mobile) describe moderate explanatory rates.

The r-squares of 12.0% (fixed) and 17.7% (mobile) do not imply strong regressions, however, the values are two to three times higher than the r-squares, mentioned above (see Table VIII).

TABLE VIII. REGRESSION ANALYSIS – COMPARISON OF DATA SECURITY AS INFLUENCE FACTOR FOR USAGE FREQUENCY
(combined independent variables consideration on single service consideration)

Dependent variables	Independent: Perceived Data Security in Fixed Networks*			Independent: Perceived Data Security in Mobile Networks*		
	Regression Coefficient B	Significance	R-Square	Regression Coefficient B	Significance	R-Square
Usage Frequency of Email Services	0.363**	p<0.05	9.7%	No Significance		
Usage Frequency of Cloud Computing Services	0.261	p<0.05	8.2%	No Significance		
Usage Frequency of Online Banking Services	0.218	p<0.05	12.0%	0.352	p<0.05	17.7%
Usage Frequency of E-Learning Services	No Significance			0.328	p<0.05	14.5%

* Other independent variables overall perceived data security, perceived importance of data security, perceived credibility of the network operators and perceived trust in the service providers are controlled and implemented.

** The Durban-Watson-Test recognizes a value, which could be an indicator for an existing autocorrelation. To cover the spurious correlations, further investigations must be performed.

These r-squares values describe how much the perceived data security influence the decision how often online banking will be used.

The service e-learning is not used by many customers. Nevertheless, when customers use the service in the mobile environment, the decision to use is influenced by data security issues. The coefficient of 0.328 describes a moderate explanatory rate. The r-square of 14.5% is similar to the results of online banking. Although this regression is classified rather weak, it is still better than most of the regression values found earlier.

For the named services, the hypothesis H1 can be accepted because data security concerns influence the decision to use a service (frequently). However, for the other services (social media, IPTV, online gaming, instant messaging), the hypothesis H1 has to be rejected because an impact of data security issues on the usage of these services could not be (significantly) proved.

Following the considerations of the hypotheses H1 and H2, the hypothesis H3 shall illustrate how the usage frequency of the presented Internet services is affected by the user estimation of the credibility of the network operators for fixed and mobile broadband networks. Table IX visualizes the results for the hypothesis H3, gained from the survey. In this case, only the results, which have shown a significant influence in the model, are presented. For the services email (0.192), navigation (0.184), and online banking (0.210) (see Table IX), only the credibility for the mobile network operators significantly relates to the usage frequency. If the users perceive a higher network credibility, security and trustworthiness, the users will use these services in a higher frequency.

These results match with the results, which were already gained by the analysis for H1. Users have stored personal information in the services email and online banking services and therefore, they are especially looking for a secure access of their own data. For this reason, the users utilize the services in a higher degree when they perceive a higher data security and when the infrastructures are secure, trustworthy and credible.

For the credibility of the fixed network operators, the detected relations cannot be proven. However, all of the three performed regressions, the reached r-squares have values below the 5%. Together with the regression coefficients below 0.3, we have to acknowledge that the explanatory rates of the model are weak. [59]-[61]. However, the decision to use a service is influenced by a couple of different factors. Each of the influence factors show low r-squares when we consider the services solely. The VIF values in the regressions are below 3 and consequently multicollinearities can be excluded [57][63][64]. All regressions have values below 1 in the Durbin-Watson-Test, which means possible autocorrelation issues (like spurious correlations) cannot be excluded for these regressions and further examinations would be necessary.

Despite the critical possible impact factors of autocorrelation and the low values for r-squares and regression coefficients, 75% of the respondents use email and online banking services.

The consideration of the other mentioned Internet services (a) IPTV, (a) online shopping, (c) cloud computing and (d) online gaming shows significant relationships between the credibility of the fixed or mobile network operators and the usage frequency of the named services. Due to the positively significant relations, we expect that the users who perceive a higher trustworthiness and safety in the networks will use these services more frequently than the users without this expectation. The services IPTV (0.276/0.250) and online gaming (0.244/0.256) have coefficients below 0.3 and r-squares below 5% (see Table IX), which means weak regressions and low relationships [59]-[61]. Here, the VIF values are acceptable and multicollinearities can be excluded. However, the two regressions suffer the same problem as the previous considered ones, that the Durbin-Watson-Test show

values around 1. Nonetheless, both services do not face the same problem as the services online banking and email, because both services do not cover the same degree of credible information. Normally, both services give an overview about the actual user behavior, which TV-series or video games the users like and how much time they spend with them. Generally, the usage of these services does not depend on critical information. However, more and more services are combined with stores and online shopping possibilities. Here information about credit cards or personal information may are involved. If we speak about the actual usage of these services, we can fully comprehend that the infrastructures should be secure and transparent, but possible data leakages do not normally lead to the same consequences than if email or online banking accounts have been attacked.

The combination of services with online shopping elements builds the transition to the consideration of the two unconsidered variables, which are also included in Table IX. The services online shopping and cloud computing have r-squares of 7.0% and 8.7% for the analysis of the credibility of the fixed network operators and 5.1% for both analyses for the credibility of mobile network operators. Although the r-squares exceed 5%, we estimate weak regressions. Except the case for the credibility for cloud computing (mobile), the coefficients are below 0.3 and therefore, the credibility influences to a low degree the usage frequency of an Internet service [59]-[61]. Nevertheless, the Durbin-Watson-Test and VIF values are in an acceptable range and so we exclude autocorrelation and multicollinearity issues.

Online shopping and cloud computing also cover critical assets and information like payment information, credit card numbers or enterprise information. Therefore, it can be comprehended that these services have reached slightly better values than the services considered previously. The fact that the credibility of the mobile infrastructure for cloud computing has reached an average rate of explanation shows that the infrastructures used for the two mentioned services should receive the integrity and free attacker entries. Nonetheless, as mentioned above, the results only indicate weak regressive connections and we conclude that other factors are more influencing the usage of the services than the discussed security concerns.

The hypothesis H3 cannot be (fully) rejected or accepted. For the services, which have been presented above, significant weak relationships could be identified and so the hypothesis could be accepted. However, for the other services, which we mentioned in the introduction, relations cannot be concluded and so the hypothesis would be rejected. Consequently, the known linkage between the security patterns of credibility and trust and the behavioral intention, which were already known through the literature [18][21][22], cannot be completely confirmed for the link with actual usage of the service.

To support the previous findings and to expand the results, we will execute another combined regression analysis with the variables, which are used in the first approach above. In the following analysis, we look at perceived credibility with considering the password-changing behavior and perceived importance of data security.

The hypothesis H4 includes the combined analysis of the different security, credibility and trust variables, which have been implemented in the survey. Considering Table X, the dependent variable usage frequency for each specific service depends here directly on the combination of the independent variables.

TABLE IX. COMPARISON PERCEIVED OPERATOR CREDIBILITY INFLUENCE FACTOR FOR USAGE FREQUENCY (single service consideration)

Dependent variables: Usage Frequency	Independent variable: Perceived Credibility of Fixed Network Operators			Independent variable: Perceived Credibility of Mobile Network Operators		
	Regression Coefficient B	Signi-ficance	R-Square	Regression Coefficient B	Signi-ficance	R-Square
Email	No Significance			0.192	p<0.05	2.4%
IPTV	0.276	p<0.05	5.1%	0.250	p<0.05	4.3%
Online Shopping	0.280	p<0.05	7.0%	0.173	p<0.05	2.5%
Cloud Computing	0.281	p<0.05	5.1%	0.361	p<0.05	8.7%
Online Gaming	0.244	p<0.05	3.2%	0.256	p<0.05	3.5%
Navigation	No Significance			0.184	p<0.05	2.8%
Online Banking	No Significance			0.210	p<0.05	3.3%

The implemented independent variables in the model are: (a) perceived data security (single services) in fixed and mobile environments, (b) the general perceived data security, (c) password changing behaviors, (d) perceived importance of data security, (e) perceived credibility of network operators (fixed and mobile), (f) usage of anti-virus programs, and (g) the trust in the providers.

Due to the combined consideration of the variables, higher r-squares can be expected [59]-[61]. Besides the r-squares, the regression coefficients and the F-ratio will be also suitable indicators to present a good feedback about the model fit and the deepness of possible relationships. In the following, we will discuss some of the results, including the examination of H4.

Firstly, email services seem to be directly influenced by the credibility (fixed: -0.289/mobile: 0.289) of the network. However, users who perceive that a mobile infrastructure is free of threats use email services more frequently. Nevertheless, in the consideration in the fixed networks, the user perception is reversed. As shown in Table X, the perception of data security (0.296) also influences the decision how often the service is used. If the users perceive a better data security for emails, they will use the services more often. Despite the r-square of 13% is not quite high and the

coefficients do not exceed 0.3 [59]-[61], the significant F-ratio is above the value of 3.03, which describes a good model fit. The significance of the perceived data security supports the results, which were already got and analyzed in the examination of hypothesis H1. In addition, the significance of the credibility of the mobile network operator positively influencing the usage frequency assists the previously results. The inverse value (-0.289) for the credibility of the fixed network operators makes no sense at first sight. At best, it could be interpreted in such a way that most users now access their emails primarily via mobile networks.

Secondly, the results of cloud computing (0.236) and online banking (0.280) are mostly influenced by the importance of data security. For both services, the data security is important and therefore, when the users perceive this data security they will use this service in a higher frequency. Online banking (0.214) is also influenced by the perceived data security in mobile environments. Therefore, users who perceive a higher data security while using online banking in mobile networks will enhance their usage of this service. Both regressions have r-squares of nearly 20% and symbolize medium explanatory rates. Together with the consideration of the significant F-ratio and F-values above 3, we find a good model fit and that the significant independent variables could well describe the dependent variable usage frequency.

Bearing in mind that the regression coefficients describe a lower degree of relationship. For cloud computing and online banking, we already find several connections between security variables and the usage frequency of these services and all of them are positive. This leads to the conclusion that the users estimate for these services a secure infrastructure, service area and data keeping. If these criteria are fulfilled or at least the customer perceive this status, the users will use the services in a higher frequency.

Considering all results of Table X, we find several influence factors for the usage frequency of Internet services. There is no general influence factor, which influences the usage rate of all services. However, we have to admit that we did not expect this one influence factor at all, because the Internet services are quite different. Nevertheless, a factor, which would be an explanatory factor for the most Internet services, would have been a good result. Concluding, we expect that we cannot generally accept the hypothesis H4, because there is no general influence factor. Otherwise, we see in the model that the variables of the perceived importance of data security (influence factor of 5 services) and operator credibility (influence factor of 4 services) are the major impact variables. For this reason, the hypothesis H4 will not be fully rejected. Several impacts from data security, credibility and trust factors on the development of the usage frequency can be inferred.

Despite the high importance of data security, on the hand, people do not perceive the security of their personal data. This means, the users do not behave in the right way to increase their own personal data security. Although the user estimation of the trust about the service providers how they distribute the data of the users has no influence on the usage frequency of a service, we assume that users have some concerns about service providers about the distribution of personal data. Furthermore, the perception about the credibility and trustworthiness of the networks could possibly influence the customer usage rate of several Internet services. Here, we estimate that most Internet users expect that the Internet infrastructure and the Internet services are free of threats and risks and therefore, these services would be use.

Consequently, our estimation is here that the most Internet users expect that the operators of infrastructure and providers of services take care about the data security of the users. This means the most users do not feel responsible for the security of their own data.

Surely, we do not find general security patterns, which directly relate to the possible user behavior of Internet users. Nonetheless, for specific Internet services several security impact factors and indicators could be found and therefore, it can be assumed that the user behavior and the usage frequency of Internet services are not free of security issues.

V. Conclusions And Future Work

In the journal article, we have analyzed indicators, which influence the decision and usage frequency of Internet services. The focus of the publication is on the effects of perceived data security and perceived trust in the use decision and the usage frequency of Internet services. In the first step, the influence of perceived data security or perceived trust on the usage frequency of certain internet services was examined.

To support the results so far and to expand the results, we have conducted two further analyses. The third one includes the consideration of the impact of perceived credibility on the usage frequency of Internet services. Finally, a combined regression analysis, focusing on the impact of the data security perceived and additional factors by customers on the use of the services were conducted.

It could not be proved in general that security concerns and especially concerns in data security and trust in service providers lead to a reduced or an increased usage of the services. Nonetheless, some evidences and implications for specific services like email, online banking and e-learning exist. Customers, who perceive that their data will be safe, use the service more frequently than customers, who feel uncertain. The main question is why only some of the used services are influenced. We are in the opinion that these developments directly depend on the nature of the service. For example, bank accounts and emails usually contain confidential information, the losses of which can have serious consequences for customers. In contrast, the use of services, such as IPTV merely reveals some information to individual preferences or behaviors. However, most people do not appreciate this information as so critical.

TABLE X. REGRESSION ANALYSIS – COMPARISON OF MODEL FACTORS AS INFLUENCE FACTOR FOR USAGE FREQUENCY
(combined independent variables consideration on single service consideration)

Dependent variable*	Independent variables** Significant parameters	Coefficients	Significance	R-Square	F-Ratio	
					Significance	F-value
Email	Perceived Credibility of Fixed Operator	-0.289	p<0.05	13.0%	p<0.05	3.313
	Perceived Credibility of Mobile Operator	0.289	p<0.05			
	Perceived Data Security (fixed)	0.296	p<0.05			
Cloud Computing	Perceived Importance of Data Security	0.236	p<0.05	20.6%	p<0.05	3.389
Online Banking	Perceived Importance of Data Security	0.28	p<0.05	19.7%	p<0.05	3.458
	Perceived Data Security (mobile)	0.214	p<0.05			
E-Learning	Perceived Importance of Data Security	0.274	p<0.05	18.8%	p<0.05	2.089
Instant Messaging	Perceived Importance of Data Security	0.223	p<0.05	13.2%	p<0.05	2.930
	Password Changing Rate for Emails	-0.227	p<0.05			
IPTV	Password Changing Rate for Emails	-0.241	p<0.05	8.6%	p<0.05	1.472
Navigation	Perceived Credibility of Fixed Operator	-0.197	p<0.05	14.8%	p<0.05	3.068
	Perceived Credibility of Mobile Operator	0.244	p<0.05			
	Perceived Importance of Data Security	0.227	p<0.05			
	Password Changing Rate for Emails	0.219	p<0.05			
Social Media	Usage of Anti-virus programs	0.753	p<0.05	7.7%	p<0.05	1.478
Online Gaming	No Significance					
Online Administration	Perceived Importance of Data Security	0.216	p<0.05	17.8%	p<0.05	1.862
	Perceived Data Security (fixed)	0.333	p<0.05			
Online Shopping	Perceived Credibility of Fixed Operator	0.244	p<0.05	9.5%	p<0.05	1.964
Online Telephony	Perceived Credibility of Mobile Operator	-0.247	p<0.05	13.0%	p<0.05	1.453

* Usage frequency of the single services.
** Perceived data security (single services), password changing behaviors, perceived importance of data security, perceived credibility of network operators, anti-virus usage, perceived data security (general), perceived provider trust.

The second investigation focuses on the perceived trust in service providers. It examines how the transfer of customer data to third parties is evaluated. Interestingly, no evidences for the influence of the perceived trust on the usage of Internet services could be found. It must be predicted that data distributions by the service providers do not affect the user's decision to use a service. This non-existing relation could be explained by the fact that the most people focus on the performance and usability of the Internet services instead of the security, which is mentioned in the second section of this study. Furthermore, it must be assumed that the most people are not aware about these distributions of their data. Therefore, the rejection of this hypothesis is not surprisingly.

The following investigations of the hypotheses H3 and H4 can be well associated with results of the first examination. In general, a full relationship between the factors of the operator credibility in hypothesis H3, the security patterns in the hypothesis H4 and the influence on the usage frequency of Internet services cannot be fully underlined. We show that in some services like online shopping, cloud computing, IPTV and online gaming, that customer perceive the credibility of the networks is free of threats and therefore, it can be directly linked to the usage frequency of the service. In a lower degree, the linkages can also be accepted for the use of email, navigation and online banking services. It can be concluded that the security of the infrastructure is for the use of a couple of services a relevant issue.

In the final hypothesis H4, the usage frequency of an Internet service were directly linked to all of the different security patterns, which were included in the survey. The results present a quite divers field of impact factors for the different Internet services. As mentioned in the section before, there is not one influence factor for all the different Intern services; however, we do did not expect these kind of result. Nonetheless, two security patterns (importance of data security and operator credibility) are influence factors for 50% of the examined Internet services. For this reason, we assume that these two patterns are the major influence factor for user behavior regarding the concerns about data security. As a consequence, these two patterns are more significant than the other factors and so further investigations would be necessary.

Generally, the presented results cannot fully describe the influence of data security, credibility and trust issues on the usage of specific Internet services. However, a general comprehension of the influence of data security behaviors and the trust and credibility of providers and operators would be deepened and for specific Internet services, a relation could be proved. One intension of this study is to present the influence of the providers' reputation on the decision to use an Internet service more frequently, because researchers found out that reputation and credibility of a network and system positively rises the trust in a new application and system like mobile banking [39][44][65][66].

The focus on the assessment of the single influence of data security (and also trust and credibility) on the actual customer usage of Internet services could lead to the problem that the presented approach might not lead to the aimed results, due to the small number of considered concepts.

Consequently, we have presented several differences also in the usage of Internet services and perceptions about data security concerns between the use of the Internet services in mobile and fixed environments. Generally, there does not exist the big difference, however for several services we could find that data security is more important in the mobile than the fixed environment. We estimate here that the most users use these services more and more in the mobile networks and therefore, the perceptions and estimations target more on the mobile environment.

Finally, parts of the differences cannot be explained with the illustrated regression analyses and the considerations of the descriptive statistics, because as already mentioned above, the decision to use an Internet service is quite subjective. Determinants like cultural values, traditions, age, job, obligatory reasons and necessity directly affect usage frequency of an Internet service. Over the time, people normally gain experience with the services. Due to increasing experience, users will develop an increased confidence in the services [17][45][46][47][67]. Consequently, further analyses have to be geared on the analysis of the named factors culture, experience and age and therefore, next surveys and analyses need to test if these factors influence the actual usage of Internet services.

REFERENCES

[1] E. Massarczyk and P. Winzer, "Influence of Perceived Data Security and Trust on the Usage of Internet Services," In S. Böhm, L. Berntzen, and F. Volk (Eds.), The Tenth International Conference on Advances in Human oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2017, IARIA) [08. - 12. October 2017, Athens]. Conference Proceedings and Thinkmind Library (ISSN: 2308-3492, ISBN: 978-1-61208-592-0)

[2] E. Massarczyk and P. Winzer, "Influence of the Perception of Data Security and Security Importance on Customer Usage of Internet Services," International Journal On Advances in Internet Technology, Thinkmind Library (ISSN: 1942-2652), volume 10, numbers 1 and 2, 2017, pp. 1-22

[3] International Telecommunication Union (ITU), "ICT Facts & Figures – The world in 2015," May 2015, https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf, [retrieved: 09.2017]

[4] P. W. Dowd and J. T. McHenry, "Network Security: It's Time to Take It Seriously," Computer (1998), vol. 31, issue 9, IEEE Xplore Digital Library, Sept. 1998, pp. 24-28.

[5] D. Desai, "Law and Technology – Beyond Location: Data Security in the 21st Century," Magazine Communications of the ACM (2013), vol. 56, issue 1, ACM, Jan. 2013, pp. 34-36.

[6] F. S. Ferraz and C. A. Guimarães Ferraz, "Smart City Security Issues: Depicting Information Security Issues in the Role of a urban environment," 7th International Conference on Utility and Cloud Computing, IEEE/ACM, 2014, pp. 842-846.

[7] Kaspersky Lab, "Damage Control: The Cost of Security Breaches," IT Security Risks Special Report Series, https://media.kaspersky.com/pdf/it-risks-survey-report-cost-of-security-breaches.pdf., 2015, [retrieved 09.2017]

[8] D. Nayak, N. Rajendran, D. B. Phatak, and V. P. Gulati, "Security Issues in Mobile Data Networks," Vehicular Technology Conference (VTC 2004), vol. 5, IEEE Xplore Digital Library, Sept. 2004, pp. 3229-3233.

[9] S. Dhawan, K. Singh, and S. Goel, "Impact of Privacy Attitude. Concern and Awareness on Use of Online Social Networking," 5th International Conference - Confluence The Next Generation Information Technology Summit 2013, IEEE Xplore Digital Library, Sept. 2014, pp. 14-17.

[10] D. Malandrino, V. Scarano, and R. Spinelli, "How Increased Awareness Can Impact Attitudes and Behaviors Toward Online Privacy Protection," International Conference on Social Computing, IEEE Xplore Digital Library, Sept. 2013, pp. 57-62.

[11] A. N. Joinson. U.-D. Reips, T. Buchanan, and C. B. Paine Schofield, "Privacy, Trust, and Self-Disclosure Online, " Human-Computer Interaction, vol. 25, no. 1, 2010, pp. 1-24.

[12] B. Krishnamurthy, K. Naryshkin, and C. Wills, "Privacy Leakage vs. Protection Measures: the Growing Disconnect," in Web 2.0 Security and Privacy Workshop, 2011, pp. 1-10.

[13] Q. Tan and F. Pivot, "Big Data Privacy: Changing Perception of Privacy, International Conference on Smart City/SocialCom/SustainCom, IEEE, 2015, pp. 860-865.

[14] V. Venkatesh, J. Y. L. Thong, and X. Xin, "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology," MIS Quarterly, vol. 36, issue 1, 2012, pp. 157-178.

[15] F.-T. Lin, H.-Y. Wu, and T. T. Nguyet Nga, "Adoption of Internet Banking: An Empirical Study in Vietnam," 10th International Conference on e-Business Engnineering, IEEE Xplore Digital Library, 2013, pp. 282-287.

[16] V. Venkatesh, J. Y. L. Thong, and X. Xin, "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology," MIS Quarterly, vol. 36, issue 1, 2012, pp. 157-178.

[17] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," MIS Quarterly, vol. 27, issue 3, 2003, pp. 425-478.

[18] T. Escorbar-Rodriguez and E. Carvajal-Trujillo, "Online Purchasing Tickets for Low Cost Carriers: An Application of the Unified Theory of Acceptance and Use of Technology (UTAUT) Model," Tourism Management, vol. 43, 2014, pp. 70-88.

[19] J. Zhong, A. Dhir, M. Nieminen, M. Hämäläinen, and J. Laine, "Exploring Consumer Adoption of Mobile Payments in China," Academic Mind Trek'13, 2013, pp. 318-325.

[20] Y. S. Wang, Y. M. Wang, H. H. Lin, and T. I. Tang, "Determinants of User Acceptance of Internet Banking: an Empirical Study," International Journal of Service Industry Management, vol. 14, 2003, pp. 501-519.

[21] A. Zmijewska, E. Lawrence, R., and R. Steele, "Towards Understanding of Factors Influencing User Acceptance of Mobile Payment Systems," In: Proceedings of the IADIS WWW/Internet, Madrid, Spain, 2004.

[22] T. Dahlberg and A. Öörni, "Understanding Changes in Consumer Payment Habits – Do Mobile Payments and Electronic Invoices Attract Consumers?," In: 40th Annual Hawaii International Conference on System Sciences (HICSS), 2007, p. 50.

[23] P. G. Schierz, O. Schilke, and B. W. Wirtz, "Understanding Consumer Acceptance of Mobile Payment Services: An Empirical Analysis," Electronic Commerce Research and Applications, vol. 9, issue 3, 2010, pp. 209-216.

[24] C. Kim, W. Tao, N. Shin, and K. S. Kim, "An Empirical Study of Customers' Perceptions of Security and Trust in E-Payment Systems," Electronic Commerce Research and Applications, vol. 9, issue 1, 2010, pp. 84-95.

[25] K. Yang, "Exploring Factors Affecting the Adoption of Mobile Commerce in Singapore," Telematics and Informatics, vol. 22 issue 3, 2005, pp. 257-277.

[26] J. Cheong, M. Cheol, and J. Hwang, "Mobile Payment Adoption in Korea," In: ITS 15th biennial conference, Berlin, Germany, 2002.

[27] T. Dahlberg, N. Mallat, and A. Öörni, "Consumer Acceptance of Mobile Payment Solutions," In: G.M. Giaglis (ed.), mBusiness 2003 – The Second International Conference on Mobile Business, Vienna, 2003, pp. 211-218.

[28] N. Mallat, "Exploring Consumer Adoption of Mobile Payments – a Qualitative Study," Mobility Roundtable, Helsinki, Finland, vol. 16, issue 4, 2006, pp. 413-432.

[29] K. Pousttchi and M. Zenker, "Current Mobile Payment Procedures on the German Market from the view of Customer Requirements," In: 14th International Workshop on Database and Expert Systems Applications, 2003, pp. 870-874.

[30] D. Gefen and D. David, "Reflections on the Dimensions of Trust and Trustworthiness among Online Consumers", Database for Advances in Information Systems, vol. 33, issue 3, 2002, pp. 38-53.

[31] R. De Sena Abrahao, S. N. Moriguchi, and D. F. Andrade, "Intention of Adoption of Mobile Payment: An Analysis in the Light of the Unified Theory of Acceptance and Use of Technology (UTAUT)," Innovation and Management Review, vol. 13, 2016, pp. 221-230.

[32] D. H. McKnight, L. L. Cummings, and N. L. Chervany, "Initial Trust Formation in New Organizational Relationships," Academy of Management Review, vol. 23, 1998, pp. 473-490.

[33] D. H. McKnight, V. Choudhury, and C. Kacmar, "The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: a Trust Building Model," The Journal of Strategic Information Systems , vol. 11, 2002, pp. 297-323.

[34] T. Zhou, "Understanding Mobile Internet Continuance Usage from the Perspectives of UTAUT and Flow," Information Development vol. 27, 2011, pp. 207-218.

[35] T. Zhou, Y. Lu, and B. Wang, "Integrating TTF and UTAUT to Explain Mobile Banking User Adoption," Computers in Human Behavior, vol. 26, 2010, 760-767.

[36] T. Zhou, "An Empirical Examination of Initial Trust in Mobile Banking," Information Development, vol. 21, issue 5. 2011, pp. 527-540.

[37] A. Y. L. Chong, "Understanding Mobile Commerce Continuance Intentions: An Empirical Analysis of Chinese Consumers," Journal of Computer Information Systems, 2013.

[38] L.-D. Chen, "A Model of Consumer Acceptance of Mobile Payment," International Journal of Mobile Communications, vol. 6, issue 1, 2008, pp. 32-52.

[39] M. A. Mahfuz, L. Khanam, and W. Hu, "The Influence of Culture on M-Banking Technology Adoption: An Integrative Approach of UTAUT2 and ITM," 2016 Proceedings of PICMET'16: Technology Management for Social Innovation, 2016, pp. 70-88.

[40] X. Luo, H. Li, J. Zhang, and J. P. Shim, "Examining Multi-dimensional Trust and Multi-faceted Risk in Initial Acceptance of Emerging Technologies: an Empirical Study of Mobile Banking Services," Decision Support Systems, vol. 49, issue 2, 2010, pp. 222-234.

[41] H.-P. Lu, and P. Y.-J. Su, "Factors Affecting Purchase Intention on Mobile Shopping Websites," Internet Research, vol. 19, issue 4, 2009, pp. 442-458.

[42] T. Oliveira, M. Faria, M. A. Thomas, and A. Popovic, "Extending the Understanding of Mobile Banking Adoption: When UTAUT meets TTF and ITM," International Journal of Information Management, vol. 34, 2014, pp. 689-703.

[43] G. Kim, B. Shin, and H. G. Lee, "Understanding dynamics between Initial Trust and Usage Intentions of Mobile Banking", Information Systems Journal, vol. 19, issue 3, 2009, pp. 283-311.

[44] Y.-H. Chen and S. Barnes, "Initial trust and online buyer behavior", Industrial Management & Data Systems, vol. 107 issue 1, 2007, pp. 21-36.

[45] Y. Lu, S. Yang, P. Y. K. Chau, and Y. Cao, "Dynamics between the Trust Transfer Process and Intention to Use Mobile Payment Services: A Cross-Environment Perspective." Information & Management, vol. 48, issue 8, 2011, pp. 393-403.

[46] Y. Lu, Z. Deng, and B. Wang, "Exploring Factors Affecting Chinese Consumers' Usage of Short Message Service for

Personal Communication", Information Systems Journal, vol. 20, issue 2, 2010, pp. 183-208.

[47] Y. M. Shin, S. C. Lee, B. Shin, and H. G. Lee, "Examining Influencing Factors of Post-Adoption Usage of Mobile Internet: Focus on the User Perception of Supplier-Side Attributes", Information Systems Frontier, vol. 12, issue 5, 2010, pp. 595-606.

[48] J. Bortz and N. Döring, "Research Methods and Evaluations," [German] "Forschungsmethoden und Evaluation; für Human- und Sozialwissenschaftler," Heidelberg, Springer, vol. 4, 2009.

[49] M. Kaya, "Data Collection Procedure", [German] "Verfahren der Datenerhebung," in Albers, S./Klapper, D./Konradt, U./Walter, A./Wolf, J. (Hrsg.): Methodik der empirischen Forschung, Wiesbaden, Gabler, vol. 3, 2013, pp. 49-64.

[50] R. Likert, "A Technique for the Measurement of Attitudes," Archives of Psychology, 1932, pp. 199-224.

[51] Destatis, Statistisches Bundesamt, "Population," [German] "Bevölkerung," [Online] https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Bevoelkerungsstand/Tabellen_/lrbev01.html, 2015, [retrieved 09.2017]

[52] Statista, "Internet Users in Germany from 2001 to 2015," [German] "Anteil der Internetnutzer in Deutschland in den Jahren 2001 bis 2015," [Online]http://de.statista.com/statistik/daten/studie/13070/umfrage/entwicklung-der-internetnutzung-in-deutschland-seit-2001/, 2015, [retrieved: 09.2017]

[53] Statista, "Customers without Anti-Virus Protection," [German], "Anteil der Verbraucher ohne aktives Antivirenprogramm in ausgewählten Ländern weltweit," https://de.statista.com/statistik/daten/studie/226942/umfrage/anteil-der-verbraucher-ohne-aktives-antivirenprogramm/, 2017, [retrieved: 09.2017]

[54] L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, vol. 16, 1951, pp. 297-334.

[55] C. Fornell and D. Larcker, "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," Journal of Marketing Research, vol. 18, issue 1, 1981, pp. 39-50.

[56] R. Hossiep, "Cronbachs Alpha," [German] "Cronbachs Alpha," In Wirtz, M. A. (editor): Dorsch – Lexikon der Psychologie, vol. 17. Verlag Hans Huber, Bern, 2014.

[57] J. F. J. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, "Multivariate Data Analysis," Macmillan, New York, NY, Macmillan, vol. 3, 1995.

[58] S. Fromm, "Data Analysis with SPSS Part 1," [German] "Datenanalyse mit SPSS für Fortgeschrittene," Arbeitsbuch, vol. 2, VS Verlag für Sozialwissenschaften, GWV Fachverlage, Wiesbaden, 2008.

[59] S. Fromm, "Data Analysis with SPSS Part 2," [German] "Datenanalyse mit SPSS für Fortgeschrittene 2: Multivariate Verfahren für Querschnittsdaten," Lehrbuch, vol. 1, VS Verlag für Sozialwissenschaften, Springer, Wiesbaden, 2010.

[60] N. M. Schöneck and W. Voß, "Research Project," [German] "Das Forschungsprojekt – Planung, Durchführung und Auswertung einer quantitativen Studie," vol. 2. Springer Wiesbaden, 2013

[61] A. Field, "Discovering Statistics Using SPSS," Sage Publications Ltd., vol. 4, 2013.

[62] F. Brosius, "SPSS 8 Professional Statistics in Windows," [German] "SPSS 8 Professionelle Statistik unter Windows," Kapitel 21 Korrelation, International Thomson Publishing, vol. 1, 1998.

[63] D. Lin, D. P. Foster, and L. H. Ungar, "VIF Regression: A Fast Regression Algorithm for Large Data," Journal of the American Statistical Association, vol. 106, issue 493, 2009, pp. 232-247.

[64] S. Petter, D. W. Straub, and A. Rai, "Specifying Formative Constructs in Information Systems Research," MIS Quarterly, vol. 31, issue 4, 2007, pp. 623-656.

[65] C. Flavian, M. Guinaliu, and E. Torres, "The Influence of Corporate Image on Consumer Trust – a Comparative Analysis in Traditional Versus Internet Banking", Internet Research, vol. 15 issue 4, 2005, pp. 447-470.

[66] M. A. Fuller, M. A. Serva, and J. Benamati, "Seeing is Believing: the Transitory Influence of Reputation Information on E-Commerce Trust and Decision Making", Decision Sciences, vol. 38, issue 4, 2007, pp. 675-699.

[67] S. S. Kim and N. K. Malhotra, "A Longitudinal Model of Continued IS Use: An Integrative View of Four Mechanisms Underlying Post-Adoption Phenomena," Management Science vol. 51, issue 5, 2005, pp. 741-755.

A Survey and Comparison Analysis of Reference Architectures for the Cloud Computing and Internet-of-things Context

Hongyu Pei Breivold
Industrial Internet-of-things
ABB Corporate Research
Västerås, Sweden
hongyu.pei-breivold@se.abb.com

Abstract— **Increased connectivity and emerging autonomous cloud and Internet-of-things technologies are motivating the transformation of the traditional product-focused development to cloud-based solutions and service-oriented business model in many companies. In line with this, several reference architectures for the cloud computing and Internet-of-things have been developed by various research initiatives and industry vendors. Although some of these reference architectures have continued their development tracks in parallel and have different focus, they also have similarities in many perspectives, which may result in confusion in understanding and applying appropriate reference architectures for specific use cases. The aim and main output of this study is therefore to survey these existing Internet-of-things reference architectures using the research method of systematic mapping study, clarify their characteristics, and analyze in-depth how these reference architectures address various perspectives, including technology, process, quality and key system concerns, business and people. Based on the analysis we discuss motivating factors of the reference architectures, the coverage of business architecture and customer context in the reference architectures, as well as the impacts of the reference architectures to research community and practice. In addition, we present several other relevant activities and initiatives related to the architecture context for cloud computing and Internet-of-things.**

Keywords-reference architecture model; survey; comaprison analysis; industrial internet-of-things; smart industrial automation.

I. INTRODUCTION

This article is an extended version of a conference paper [1] published at ICSEA 2017 (the twelfth International Conference on Software Engineering Advances), one of the IARIA conferences. This article has extended the previously published conference paper with more details on the applied research method and a much more in-depth comparison analysis of the reference architectures.

The German Federal Ministry of Education and Research defines Industrie4.0 [2] as the flexibility to enable machines and plants to adapt their behavior to changing orders and operating conditions through self-optimization and reconfiguration. Consequently, future smart factories require systems to have the ability to perceive information, derive findings and insights, and change their behavior accordingly, and store knowledge gained from experiences. Many organizations start to see the potential opportunities of the Internet-of-things and its impacts on providing solutions that could offer operational advantages [3].

Within EU, Cloud Computing and Internet-of-things are listed as hot topics. The European Commission has outlined a European Cloud Computing Strategy [4], and urged for the need to ensure Europe being at the forefront of the development of Cloud Computing and Internet-of-things to benefit on both demand and supply side through wide-spread cloud use and cloud provision. In line with this, there has been a number of EU-funded research initiatives and activities on Cloud Computing and Internet-of-things, covering various aspects such as communication, hardware technology, identification and network discovery, security, interoperability, standardization, etc. Some examples are IERC – European Research Cluster on the Internet of Things [5], Industrial Internet Consortium [6], Industri4.0 [7], and the creation of the Alliance for Internet of Things Innovation (AIOTI) [8] by the European Commission [9], which initiates the development and future deployment of the Internet-of-things technology in Europe. There has also been a number of EU-funded research activities in Internet-of-things implementation and adoption, addressing various domains and use cases in smart cities, smart energy and smart grid, healthcare, food and water tracking, logistics and retail, and transportation [10].

Successful adoption of cloud computing and Internet-of-things requires guidance around planning and integrating relevant technologies into the existing services and applications. Both industry and academia that want to implement cloud-based solutions seek for more information about best practices for migrating and adopting cloud computing and Internet-of-things concepts. According to [11], "Defining a cloud reference architecture is an essential step towards achieving higher levels of cloud maturity. Cloud reference architecture addresses the concerns of the key stakeholders by defining the architecture capabilities and roadmap aligned with the business goals and architecture vision". Study [12] holds similar viewpoints. According to [12], in order to effectively build cloud-based enterprise solutions, there is a need for the definition of a systematic architecture that provides templates and guidelines and can be used as a reference for the architects or software engineers within the software development lifecycle. Therefore, several

reference architectures have been developed and evolved. According to [13], a reference architecture incorporates the vision and strategy for the future. With high level of abstraction, a reference architecture provides a common structure and guidance for dealing with core aspects of developing, using and analyzing systems and solutions that can be tailored to different use cases and specific needs from multiple organizations.

There are several well-known reference architectures for the Internet-of-things that have been developed over the years. Some examples are Reference Architecture Model for Industrie 4.0 (RAMI4.0) [14], Industrial Internet Reference Architecture (IIRA) [15], IoT Architectural Reference Model (IoT-ARM) [16], etc. However, only a subset of the available IoT reference architectures have been reviewed in literature, for instance, study [17] compares two major architectures IoT-ARM and IIRA, whereas study [18] analyzes the IoT architectural reference model (IoT-ARM) and the architecture proposed by WSO2. To our knowledge no detailed survey and analysis of a comprehensive coverage of IoT reference architectures has been published previously to describe the wide spectrum of reference architectures that are available for the cloud computing and Internet-of-things context. The main objective of our research is therefore to systematically select and review published literature, and also include the state-of-the-practice results from various research initiatives and activities within the cloud computing and Internet-of-things area in order to present a holistic overview of the existing reference architectures for the cloud computing and Internet-of-things context.

We have noticed that although some of the reference architectures in this survey have continued their development tracks in parallel and have different focus, they also have similarities in many perspectives, which may result in confusion in understanding and applying appropriate reference architectures for specific use cases. Consequently, we have defined the following research questions:

1) What reference architectures have been reported in the cloud computing and Internet-of-things context?

2) What are the major perspectives covered in these reference architectures?

3) What are the main characteristics for each reference architecture with respect to the different perspectives?

4) What are the impacts of the reference architectures to research community and practice?

In this paper, we present a survey of the existing reference architectures for the Internet-of-things, identify the main perspectives covered in these reference architectures, and analyze the characteristics of these reference architectures from these identified perspectives, such as technology, process, quality and key system concerns, business and people.

The remainder of the paper is structured as follows. Section II describes the research method used for this study. Section III presents an overview of the existing reference architectures for the cloud computing and Internet-of-things. Section IV describes some relevant organized Internet-of-things initiatives and activities. Section V gives an in-depth comparison analysis of the surveyed reference architectures from different perspectives, including technology, process, quality and key system concerns, business and people. Section VI discusses some principle findings of the surveyed reference architectures, including the main motivating factors of the reference architectures, the coverage of business architecture and customer context in the reference architectures, as well as the reference architectures' impacts on research and practice. Section VII concludes the paper.

II. RESEARCH METHOD

The inclusion of the reference architectures in this survey is based on the results from a mapping study [19] as well as additional state-of-the-practice information from various research initiatives and activities within the cloud computing and Internet-of-things area, and covers therefore a collection of the existing reference architectures available. The systematic mapping study is a formalized and repeatable process to document relevant knowledge on a specific subject area for obtaining all available research information related to a research area. The mapping study includes several steps:

(1) definition of research questions;
(2) conduct search for primary studies;
(3) screen papers for relevance using inclusion and exclusion criteria defined; and
(4) classify keywords of abstracts and synthesis of the data extracted. These steps are detailed in [19].

As pointed in [20], the evolution of cloud computing and Internet-of-things has been mainly industry-driven. There are a variety of tools, platforms and infrastructures developed by different business vendors, providing frameworks with various hardware and software capabilities that are used in many industrial cloud-based solutions. Therefore, in addition to searching in scientific databases, we also searched on web sites about different cloud providers, white papers published in industrial communities, such as ARC Advisory Group [21], which performs technology market research for industry. Two major relevant postings from ARC include Operational Technology Viewpoints [22], which provides insights on emerging technologies, practices, and processes for enhancing industrial operations, and Industrial IoT/Industrie 4.0 Viewpoints [23] on digitizing industry and infrastructure.

III. REFERENCE ARCHITECTURES FOR CLOUD OMPUTING AND THE INTERNET OF THINGS

This section summarizes the existing well-known reference architectures for cloud computing and the Internet-of-things.

A. *Reference Architecture Model for Industrie 4.0 (RAMI4.0)*

RAMI 4.0 [14] is a reference architecture for smart factories. It was initiated in Germany, and is driven by major companies in industry sectors. RAMI 4.0 addresses the Industrie4.0 [7] problem space from three dimensions, i.e., it

is hierarchically structured to manage both vertical integration within the factory, as well as horizontal integration extending beyond individual factory locations, in combination with lifecycle and value streams of manufacturing applications for all the factories and all the parties involved, from engineering through component suppliers to the customers. This reference architecture aims to address four aspects, including horizontal integration through value networks, vertical integration within a factory, lifecycle management and end-to-end engineering, and human beings orchestrating the value stream. In RAMI4.0, the term Industrie4.0 is used to stand for the fourth industrial revolution in the organization and control of the entire value stream along the life cycle of a product. All relevant information is available in real-time through the networking of all instances, e.g., people, objects and systems involved in value creation. By connecting these instances, the value stream are derived from data at all times to create dynamic, self-organized, cross-organizational, real-time optimized value networks based on a range of criteria, such as costs, availability and consumption of resources.

B. Industrial Internet Reference Architecture (IIRA)

IIRA [15] is a standard-based reference architecture developed by the Industrial Internet Consortium [6] for industrial internet systems, which are large end-to-end systems integrating industrial control systems with enterprise systems, business processes and analytics solutions. In this context, the term industrial internet is used to represent Internet-of-things, machines, computers and people, enabling intelligent industrial operations using advanced data analytics for transformational business outcomes. It embodies the convergence of the global industrial ecosystem, advanced computing and manufacturing, pervasive sensing and ubiquitous network connectivity.

This reference architecture is based on ISO/IEC/IEEE 42010:2011 [24] and adopts the general concepts in the specification, such as concern, stakeholder, and viewpoint. The term concern refers to any topic of interest pertaining to the system. The various concerns of an industrial internet system are classified into four viewpoints, i.e., business, usage, functional and implementation. The business viewpoint addresses the concerns of the identification of stakeholders and their business vision, values and objectives. The usage viewpoint addresses the concerns of expected system usage and capabilities. The functional viewpoint focuses on the functional components in an industrial internet system, their interrelation and structure, the interfaces and interactions between them and with external environment. The implementation viewpoint focuses on the technologies needed to implement functional components, communication schemes and lifecycle procedures. Some key system characteristics addressed in IIRA to ensure the core functions of industrial systems over time include safety, security and resilience.

C. IoT Architectural Reference Model (IoT-ARM)

IoT-ARM [16], developed within the European project IoT-A, is an architectural reference model that aims to connect vertically closed systems, architectures and application areas for creating open systems and integrated environments and platforms. In this model, Internet-of-things is treated as an umbrella term for interconnected technologies, devices, objects and services. This reference model consists of several sub-models, of which a primary and mandatory model is the IoT domain model, describing all the concepts and their relations that are relevant in the Internet-of-things, such as devices, IoT services, and virtual entities. All the other models, such as the IoT information model, functional model, communication model, IoT trust, security and privacy model, together with the IoT reference architecture are based on the concepts introduced in the domain model. The IoT reference architecture adopts the definition of architectural views and perspectives from [25], though excludes use case specific views to ensure IoT-specific needs and application-independence in the reference architecture. The key architectural views of the Internet-of-things reference architecture include IoT functional view, IoT information view, IoT deployment and operational view. The architectural perspectives of the Internet-of-things reference architecture tackle non-functional requirements, including evolution and interoperability, availability and resilience, trust, security and privacy, and performance and scalability.

D. IEEE Standard for an Architectural Framework for Internet of Things (P2413)

The P2413 standard [26] provides an architectural framework that aims to capture the commonalities, interactions and relationships across multiple domains and common architecture elements. It includes descriptions of various Internet-of-things domains, definitions of IoT domain abstractions, and identification of commonalities between different IoT domains. It also provides a blueprint for data abstraction and trust that includes protection, security, privacy, and safety. Similar to the Industrial Internet Reference Architecture, P2413 leverages existing applicable standards and follows the recommendations for architecture descriptions defined in ISO/IEC/IEEE 42010 [24]. According to [26], this standard provides a reference architecture that builds upon the reference model. The reference architecture covers the definition of basic architectural building blocks and their ability to be integrated into multi-tiered systems. The reference architecture also addresses how to document and mitigate architecture divergence. In this standard, things, apps and services can be integrated into what would be abstracted as a “thing”. Information exchange could be horizontal or vertical, or both.

E. Arrowhead Framework

The Arrowhead framework [27] was developed within an European research project in automation, which aims to facilitate collaborative automation by networked devices for five business domains, i.e., production (manufacturing, process, and energy), smart buildings and infrastructures, electro-mobility, energy production and virtual markets of energy. This framework is based on service-oriented

architecture to enable the Industrial Internet-of-things. The loosely coupled and discovery properties of service-oriented architecture improve the interoperability between devices and the integration of services provided by these devices. The concept of local clouds with well-defined isolation from the open Internet is used to support some key requirements of automation systems, such as real-time, security and safety, scalability and engineering simplicity. The dynamic characteristic of Internet of things is key in this framework. On the one hand, things come and go, and they may have limited bandwidth or energy supply. On the other hand, the integration of IoT systems needs to be dynamic depending on the demand and availability. There are three core components in the local cloud services, i.e., service registry, authorization, and orchestration. In order to be Arrowhead compliant, the applications within the network should register the services they provide within the service registry component. The authorization component manages the access rules for specific services, and the orchestration component manages connection rules for specific services to allow dynamic reconfiguration of the service consumer and service provider endpoints [28].

F. WSO2 IoT Reference Architecture

Based on the projects deployed with customers to support Internet-of-things capabilities, the company WSO2 has proposed a reference architecture [29] that aims to support integration between systems and devices. Their definition of the Internet-of-things is the set of devices and systems that interconnect real-world sensors and actuators to the Internet. The WSO2 reference architecture consists of five layers, including:

(1) device layer, in which each device has a unique identifier and is directly or indirectly attached to the Internet;
(2) communication layer, which supports the connectivity of the devices with multiple protocols for communication between the devices and the cloud;
(3) aggregation/bus layer, which aggregates communications from multiple devices, brokers communications to a specific device, and transform between various protocols;
(4) event processing and analytics layer; which processes and acts upon the events from the bus, and perform data storage; and
(5) client/external communication layer, which enables users to communicate and interact with devices and obtain views into analytics and event processing.

Besides these vertical layers, there are also two cross-cutting layers: (i) device manager, which communicates with and remotely manages devices, and maintain the list of device identities; and (ii) identity and access management for access control.

G. Microsoft Azure IoT Reference Architecture

The Azure Internet-of-things reference architecture [30] is built upon Microsoft Azure platform to connect, store, analyze and operationalize device data to provide deep business insights. This architecture consists of core platforms services and application-level components to facilitate processing needs across three main areas of IoT solutions, i.e., (1) device connectivity; (2) data processing, analytics and management; and (3) presentation and business connectivity. The guiding principles for the architecture include software and hardware heterogeneity to manage diverse scenarios, devices and standards, security and privacy, as well as hyper-scale deployments. The goal of the reference architecture is to connect sensors, devices, and intelligent operations using Microsoft Azure services. The key architecture components to reach this goal include:

(1) device connectivity, which manages different device connectivity options for IoT solutions;
(2) device identity store, which manages all device identity information and allows for device authentication and management;
(3) device registry store, which handles discovery and reference metadata related to provisioned devices;
(4) device provisioning, which allows the system to be aware of the device capabilities and conditions;
(5) device state store, which handles operational data related to the devices;
(6) data flow and stream processing;
(7) solution UX for graphical visualization of device data and analysis results;
(8) App backend, which implements required business logic of an IoT solution;
(9) business systems integration; and
(10) at-rest data analytics.

H. Internet-of-everything Reference Model

The Internet-of-everything reference model [31] is developed by the Architecture Committee of the IoT World Forum hosted by Cisco. This model defines standard terminology and functionality for understanding and developing Internet-of-things solutions, which connect people, process, data and things to enable intelligent interactions between them to achieve relevant and valuable business opportunities. This reference model is composed of seven levels, including:

(1) physical devices and controllers that control multiple devices;
(2) connectivity for reliable and timely information transmission between devices and the network, across networks, and between the network and low-level information processing level;
(3) edge/fog computing that bridges information technology and operational technology, i.e., performing high-volume data analysis and transformation of network data flows into information suitable for storage and higher level processing;
(4) data accumulation that converts event-based data generated by the devices to query-based data consumption for applications to access data when necessary;

(5) data abstraction that renders data and its storage to enable developing simple and performance-enhanced applications;
(6) applications that vary from control application to mobile application or business intelligence and analytics; and
(7) collaboration and processes that involve people and business processes to empower smooth communication and collaboration between people.

I. Intel IoT Platform Reference Architecture

Intel has defined a system architecture specification (SAS), which is a reference architecture for Internet-of-things, i.e., for connecting products and services so that they can be aware of each other and surrounding systems in their ecosystems [32]. There are two versions of reference architectures: version 1.0 for connecting the unconnected, using an Internet-of-Things gateway to securely connect and manage legacy devices that are lack of intelligence and Internet connectivity; version 2.0 for smart and connected things, addressing security and integration capabilities that are essential for real-time and closed-loop control of the data shared between smart things and the cloud. Similar to the Internet-of-things reference architecture proposed by IoT World Forum Architecture Committee, version 2.0 also facilitates the integration of operational technology and information technology. The Intel Internet-of-things reference architecture is a layered architectural framework, comprising of:

(1) communications and connectivity layer, which enables multi-protocol data communication between devices at the edge and between endpoint devices/gateways, the network, and the data center;
(2) data layer with analytics distributed across the cloud, gateways, and smart endpoint devices for optimized time-critical or computation-intensive applications;
(3) management layer for realizing automated discovery and provisioning of endpoint devices;
(4) control layer;
(5) application layer; and
(6) business layer utilizing the application layer to access other layers in the solution.

There is a vertical security layer as well, which handles protection and security management across all layers, spanning endpoint devices, the network, and the cloud.

IV. OTHER INTERNET OF THINGS ACTIVITIES

In addition to the reference architectures presented in the previous section, there are also several other projects, activities and initiatives dedicated in the architecture context for cloud computing and the Internet-of-things, summarized in the following sub-sections.

A. IoT European Research Cluster (IERC)

The objective of IERC initiative [5] is to define a common vision of Internet-of-things technology and address IoT technology research challenges with respect to connected objects, the Web of Things, and the future of the Internet capabilities at the European level, and enable knowledge sharing in the view of global development. According to IERC, Internet-of-things is a dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual things have identities, physical attributes, and virtual personalities and use intelligent interfaces, and are seamlessly integrated into the information network. To facilitate the vision of Internet-of-things business ecosystems implementing smart technologies to drive innovation, a wide range of research and application projects have been set up within the IERC initiative, investigating aspects related to (i) devising disruptive business models, transforming traditional business model to data-driven models where all actors in the value chain are closely interconnected; (ii) trust evaluation and management in Internet-of-things, concerning provision of reliable information and maximizing security, privacy and safety; (iii) the impact and consequences of the fast-paced technology development enabling connected things, services, data and people on society with respect to legal considerations, regulations and policies, such as personal data protection, data ownership; (iv) standards and IoT platforms that support open and dynamic interaction across both dimensions of horizontal IoT domains and vertical application domains, and overcome the fragmentation of closed systems, architectures, and applications. A tightly related Internet-of-things activity to IERC is the Alliance for Internet of Things Innovation (AIOTI) [8], which was initiated by the European Commission to address the challenges of Internet-of-things technology and application deployment, including standardization, interoperability and policy issues that are of common interest among various IoT players.

B. Smart Applicances (SMART)

SMART is an EU-funded study [33] with focus on semantic assets for smart appliance interoperability. Smart appliances are devices used in households capable of communicating with each other and being controlled via Internet. It provides a standardized framework for the smart appliances reference ontology, of which recurring concepts can be used and extended in several domains in addition to residential environments.

C. Architecture and Interfaces for Web-oriented Automation System (WOAS)

The project WOAS [34] is funded by the German Federal Ministry of Economics and Technology as an industrial joint research project with ten German automation companies involved. The aim of this project is to research a new architecture for automation systems based on cloud-based web technologies. The proposed architecture is referred to as a Web-Oriented Automation System (WOAS). A WOAS comprises a system kernel and a configurable number of automation services that implement and realize the required automation functions. The automation service is realized according to the concept of I40 component [14]. The connection of the automation service with distributed automation devices in the network is implemented via

standard industrial interfaces and is also based on the concept of I40 component.

D. Reference Architecture for IoT-based Smart Factory

A research study [35] presents a reference architecture for smart factories and defines the main characteristics of such factories with a focus on sustainable energy management perspective. According to this study, Internet of things relies on both smart objects and smart networks. It is a system in which the physical items are enriched with embedded electronics, such as RFID tags and sensors, and are connected to the Internet. This reference architecture builds upon the interactive relations between smart factories and customers, allowing smart factories to collect and analyze data from products and processes for improved perception of customers' needs and behaviors, as well as better products and services. There are several sets of technologies and perspectives in this reference architecture, including smart machines, smart devices, smart manufacturing processes, smart engineering, manufacturing IT, smart logistics, big data and cloud computing, smart suppliers (i.e., building sustainable relations with suppliers), smart customers' behavior, and smart grid infrastructure for energy management.

V. A COMPARISON ANALYSIS OF THE REFERENCE ARCHITECTURES

The reference architectures described in Section III have certain similarity in technical concepts and architectural principles, but there are also differences in their respective technology approaches and implementations. Therefore, we group particular characteristics that have similar concerns to describe the same or related aspects of these reference architectures. The aspects in the comparison that we have identified and are going to address include:

(1) technology perspective, addressing key concepts and principles used in the reference architecture;
(2) process perspective, addressing the coverage of guidelines and process steps involved when using the reference architecture to generate concrete architectures or migrate existing solutions using the reference architecture;
(3) quality and key system concerns perspective, addressing main quality attributes and system characteristics that a specific reference architecture focuses on; and
(4) business and people perspective, addressing the coverage of value stream aspect and users-centered perspective in a specific reference architecture.

The comparison analysis of the main characteristics of the reference architectures is detailed below, and Table I at the end of the paper summarizes the comparison of the surveyed reference architectures.

A. Technology

The transformation of industrial sectors towards digitalization represents a complex problem space that often requires a solution space with multiple viewpoints or dimensions to describe, understand, and manage this complexity. This is reflected in several reference architectures that are often comprised of multiple models (as in e.g., IoT-ARM), multiple viewpoints (as in e.g., IIRA, P2413), or multiple layers (as in e.g., RAMI4.0, WSO2, Azure IoT, Internet-of-everything model, and Intel IoT). In contrast to these architectures, the Arrowhead framework takes another approach that focuses on cloud integration technologies. It applies service-oriented architecture, and builds local and inter cloud with orchestrated services to enable collaborative automation.

The level of standardization and usage of standards are also different among the reference architectures. For instance, IIRA and P2413 follow the architecture descriptions and general concepts such as concern, stakeholder, and viewpoint as defined in ISO/IEC/IEEE 42010. RAMI4.0 follows and extends other standards, i.e., IEC62264 and IEC61512. IoT-ARM adopts the definition of architectural views and perspectives from [25], utilizing viewpoints with respect to function, information, concurrency, development, deployment and operation to describe a system's internal structure, and the context viewpoint to describe a system's external entities.

Among the other reference architectures, a different approach has been taken instead of using a specific standard. A typical approach is the usage of an established architecture style. For example, the arrowhead framework applies service-oriented architecture to handle interoperability at service level. The WSO2 and Azure IoT reference architectures choose multi-layered architectures to handle interoperability issues within communication layer or device layer using specific protocols and technologies.

B. Process

Most of the reference architectures provide some guidelines and process steps at different levels of detail on the instantiation of the reference architectures. For instance, RAMI4.0 introduces an administration shell that allows the integration of physical things to Industrie4.0. WSO2 provides a mapping into products and capabilities of the WSO2 platform when instantiating the WSO2 reference architecture. Regarding reference architectures IIRA, Azure IoT, Internet-of-everything, and Intel IoT, they all focus on the integration of information technologies (IT) and operational technologies (OT). IIRA provides an implementation viewpoint that describes the technical representation of an industrial internet system and the technologies and system components required for IT/OT integration. The reference architectures Azure IoT, Internet-of-everything, and Intel IoT can be instantiated into concrete solutions and align with domain-specific designs. The reference architecture IoT-ARM provides very detailed guidelines and engineering practices on how to derive concrete architecture from the reference architecture, which specifies transformation rules for translating the abstract models into a concrete architecture. The reference architecture P2413 aims to capture the commonalities across domains and thus provides a basis for the instantiation of concrete domain-specific architectures.

C. Quality and key system concerns

Although the quality attributes addressed in the reference architectures vary, there are a couple of key quality concerns that are common to all, i.e., interoperability and security, which are also two major challenges in the context of cloud computing and Internet-of-things, in which distributed devices and systems from various vendors are connected to exchange data.

Interoperability is defined as "*the capability to exchange information with each other based on common conceptual models and interpretation of information in context*" [15]. Different reference architectures address interoperability differently. For instance, IIRA addresses syntax interoperability in connectivity functional layer, and semantic interoperability in data management. WSO2 uses standard interoperable protocols such as HTTP, MQTT, and AMQP. Similarly, Intel IoT promotes standard-based environment for achieving interoperability. In IoT-ARM, the interoperability is achieved through the design-choice process by identifying and evaluating design choices with respect to their impact to interoperability.

Security is defined as "*the condition of the system operating without allowing unintended or unauthorized access, change or destruction of the system or the data and information it encompasses*" [15]. As for interoperability, various reference architectures address security differently. For instance, IIRA proposes an integrated approach to security by considering end-to-end security capability across all the viewpoints. WSO2 focuses on encryption on devices, identity models, access control, and management of keys and tokens to address security requirements. Intel IoT provides a layered end-to-end security approach for endpoint device, network, and cloud levels. In Azure IoT, security measures are taken across various areas, including device and user identity, authentication and authorization, and data protection, etc.

D. Business and people

Interconnecting things, services, and people, and extracting useful information and knowledge from data analysis will enable intelligent industrial operation, generate new revenue opportunities, and lead to innovative business models. Therefore, the business context, values, and people aspects are essential in reference architectures. Some of the reference architectures have defined specific business viewpoint and people-oriented aspect. For instance, the reference architecture IIRA has a business viewpoint to address the concerns of stakeholders and their business vision, values and objectives. The key capabilities identified in the business viewpoint need to be realized through other viewpoints such as usage, functional and implementation viewpoints. In the reference architecture RAMI4.0, a business process layer is defined to ensure the integrity of the functions in the value stream and map the business models and the overall business process. In contrary to IIRA and RAMI4.0, some other reference architectures do not explicitly include the business perspective. Instead, the business aspect is addressed in the IoT architecture generation process, i.e., generation of requirements and transformation of these requirements into a concrete architecture. The reference architecture IoT-ARM is one such example, in which the definition of business goals sets the scope when generating the concrete architecture. The P2413 reference architecture does not explicitly address the business perspective either, though it explicitly addresses the people aspect by identifying the stakeholders who have an interest in a system and documenting their respective concerns. Some other reference architectures provide technical components that enable business value stream generation. For instance, the Internet-of-everything reference model and Intel IoT reference architecture offer business intelligence and data analytics components to enable smart decision making as a value proposition. Not all reference architectures include the business and people perspective, e.g., the Arrowhead framework and the WSO2 reference architecture.

VI. DISCUSSIONS

The reference architectures described in Section III provide an overview of the existing software architecture research and practice. The following sub-sections discuss the motivating factors of the development of these reference architectures, the scope of cloud computing and Internet-of-things reference architectures, as well as the reference architectures' potential impacts on research and practice.

A. Motivating Factors of IoT Reference Architectures

From surveying the existing reference architectures for Internet-of-things, we have found out several driving forces of the development of these reference architectures, including:

(1) increasing complexity and size of the systems due to the tremendous amount of connected heterogeneous devices both within and across domains;
(2) increased need for shorter time-to-market and rapid development;
(3) new collaborative solutions that require integrated and coordinated information management to ensure improved effectiveness and optimized production processes or process chains in a single plant or across plants;
(4) increasing need to achieve interoperability and compliance between different devices and systems;
(5) increased focus on optimizing the assets in a single physical plant, as well as optimizing operations across asset types, fleets, customers and partners involved in the cloud computing and Internet-of-things value chain for value co-creation.

Many of the above driving forces are also in line with the identified objectives of reference architectures as described in [13].

B. Scope of IoT Reference Architectures

According to [13], a reference architecture should address technical architecture, business architecture and customer context. When surveying the reference architectures, we have found that most of the reference

architectures provide technical solutions, design patterns and tactics. However, the business architecture and customer context are often missing. For instance, some commonly used architecture patterns among these surveyed reference architectures include multitier architecture pattern using edge tier, platform tier and enterprise tier, edge-to-cloud architecture pattern, multi-tier data storage architecture pattern, distributed analytics architecture pattern, gateway or edge connectivity and management architecture pattern, etc. However, the aspects of business models and lifecycle considerations in a business architecture are often missing. In the surveyed reference architectures, RAMI4.0 and IIRA are two reference architectures that explicitly include business architecture aspects. A main characteristics of RAMI4.0 is the combination of lifecycle and value stream with a hierarchically structured approach. IIRA explicitly defines business viewpoint to address business vision, value proposition and objectives. Besides the missing emphasis on business architectures, the customer context that addresses the processes and user considerations in the customer enterprises are often missing as well. This indicates a need for further enhancement of the existing reference architectures with business architecture perspective to help customers realize potential value proposition during the process of concretizing the reference architectures.

C. *Impacts on Research and Practice*

One important aspect of a reference architecture is to provide practices and guidance for generating new concrete architectures [13] in order to be able to identify and close any technical gaps for the implementation of potential use cases. Some reference architectures explicitly address this aspect. For instance, in the reference architecture IIRA, the implementation viewpoint explicitly addresses the technical representation, the technologies and system components required to implement the activities and functions required when generating concrete architectures. Another example is the reference architecture IoT-ARM, which provides best practices and guidance for generating concrete architectures from IoT-ARM. It can also be used to devise system roadmaps that lead to minimum changes between two product generations while guaranteeing system capability and features. Another use of the reference architecture is benchmarking during functional components review process. One example is the reference architecture P2413, which supports system benchmarking, safety and security assessment.

For practitioners in industry, the capability to cope with typical characteristics of legacy systems [36] and address legacy issues is often regarded as one important aspect in a reference architecture. Among the surveyed reference architectures, Arrowhead is one example that addresses explicitly the migration of ISA-95 systems [37] to service-based collaborative automation systems in the cloud. For the reference architectures that do not explicitly take the legacy aspect into consideration, there is a need for identification of further extension and improvement opportunities for the future.

VII. CONCLUSION AND FUTURE WORK

In this paper, we have surveyed existing well-known reference architectures, activities and initiatives for the cloud computing and Internet-of-things. To better understand and apply appropriate reference architectures for specific use cases, we have described the main aspects of these reference architectures, and have made an in-depth comparison analysis of these reference architectures from different perspectives, including technology, process, quality and key system concerns, business and people. The main characteristics covered in these perspectives are described, analyzed and compared among the reference architectures in detail.

Based on the description and comparison analysis, we also discuss the driving forces and motivating factors of these reference architectures, the coverage of technical, business architecture and customer context in the reference architectures, and how they address the generation of concrete architectures, as well as the legacy migration perspective. Although it is difficult to find information on examples of solutions or concrete products implementing each architecture described, we believe that our analysis and discussions would assist practitioners in their choice of reference architectures in practice, and in the meanwhile provide input on how to further improve and develop these reference architectures. For future work, we will examine further the reference architectures to analyze possible options of standardization, and analyze these reference architectures' suitability in different business scenarios and concrete case study contexts.

ACKNOWLEDGMENT

This article is an extended version of the conference paper [1]. Thanks to the IARIA Board for promoting extended versions of selected papers.

Special acknowledgement to the Swedish Foundation for Strategic Research through the project "Internet-of-things and Cloud for Intelligent Manufacturing" (SM16-0025).

REFERENCES

[1] H. Pei-Breivold, "A Survey and Analysis of Reference Architectures for the Internet-of-things", International Conference on Software Engineering Advances, ICSEA, 2017.

[2] The Economist Intelligence Unit, "The Internet of Things Business Index: A Quiet Revolution Gathers Pace," 2013.

[3] Industrial Internet Insights Report for 2015, Accenture .

[4] European Commission, Unleashing the Potential of Cloud Computing in Europe, 2012.

[5] European Research Cluster on the Internet of Things, http://www.internet-of-things-research.eu/, retrieved in April 2018.

[6] Industrial Internet Consortium, http://www.industrialinternetconsortium.org/, retrieved in April 2018.

[7] Industri 4.0, http://www.plattform-i40.de/, retrieved in April 2018.

[8] https://www.aioti.eu/, retrieved in April 2018.

[9] European Commission Internet-of-things, http://ec.europa.eu/digital-agenda/en/internet-things, retrieved in April 2018.

[10] E. Borgia, "The Internet of things vision: key features, applications and open issues", Journal of Computer Communications, 2014.

[11] An Oracle White Paper, "Cloud reference architecture", Oracle Enterprise Transformation Solutions Series, 2012.

[12] J. Liu, L.J. Zhang, B. Hu, and K. He, "CCRA: Cloud computing reference architecture", IEEE International Conference on Services Computing (SCC), 2012.

[13] R. Cloutler, G. Muller, D. Verma, R. Nilchiani, E. Hole, and M. Bone, "The concept of reference architectures", Systems Engineering, vol.13, 2010.

[14] RAMI 4.0, https://www.zvei.org/en/subjects/industry-4-0/the-reference-architectural-model-rami-40-and-the-industrie-40-component/, retrieved in April 2018.

[15] IIRA, http://www.iiconsortium.org/, retrieved in April 2018.

[16] A. Bassi et al, Enabling things to talk – designing IoT solutions with the IoT architectural reference model, ISBN 978-3-642-40402-3, Springer, 2013.

[17] M. Weyrich, and C. Ebert, "Reference architectures for the Internet of things", IEEE Software, vol. 33, issue 1, 2016.

[18] E. Cavalcante, M.P. Alves, and T. Batista, "An analysis of reference architectures for the Internet of things", International Workshop on Exploring Component-based Techniques for Constructing Reference Architectures, 2015.

[19] H. Pei-Breivold, "Internet-of-things and Cloud computing for smart industry: a systematic mapping study", International Conference on Enterprise Systems, 2017.

[20] C. Fehling, F. Leymann, R. Mietzner, and W. Schupeck, "A collection of patterns for cloud types, cloud service models, and cloud-based application architectures", Institute of Architecture of Applica-tion Systems, 2011.

[21] ARC Advisory Group, https://www.arcweb.com/, retrieved in April 2018.

[22] Operational Technology Viewpoints, https://www.arcweb.com/blog/operational-technology-viewpoints, retrieved in April 2018.

[23] Industrial IoT/Industrie 4.0 Viewpoints, https://industrial-iot.com/, retrieved in April 2018.

[24] ISO/IEC/IEEE 42010:2011 Systems and software engineering – architecture description, https://www.iso.org/standard/50508.html, retrieved in April 2018.

[25] E. Woods, and R. Nick, "The system context architectural viewpoint", Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture, WICSA/ECSA, 2009.

[26] IEEE P2413, IEEE Standards Association, http://grouper.ieee.org/groups/2413/, retrieved in April 2018.

[27] Arrowhead Framework, http://www.arrowhead.eu/, retrieved in April 2018.

[28] H. Derhamy, J. Eliasson, J. Delsing, and P. Priller, "A survey of commercial frameworks for the Internet of things", IEEE Conference on Emerging Technologies and Factory Automation, 2015.

[29] P. Fremantle, "A reference architecture for the Internet of things", WSO2 White Paper, version 0.9.0, 2015.

[30] https://azure.microsoft.com/en-au/updates/microsoft-azure-iot-reference-architecture-available/, retrieved in April 2018.

[31] https://www.iotwf.com/resources, retrieved in April 2018.

[32] http://www.intel.com/content/www/us/en/internet-of-things/white-papers/iot-platform-reference-architecture-paper.html, retrieved in April 2018.

[33] https://sites.google.com/site/smartappliancesproject/home, retrieved in April 2018.

[34] R. Langmann, and L. Meyer, "Automation services from the cloud", IEEE International Conference on Remote Engineering and Virtual Instrumentation, 2014.

[35] F. Shrouf, J. Ordieres, and G. Miragliotta, "Smart factories in Industry 4.0: a review of the concept and of energy management approached in production based on the Internet of things paradigm", IEEE International Conference on Industrial Engineering and Engineering Management, 2014.

[36] S. Demeyer, S. Ducasse, and O. M. Nierstrasz, Object-Oriented Reengineering Patterns, ISBN 978-3-9523341-2-6, Morgan Kaufmann, 2003.

[37] International Society of Automation (ISA), ANSI/ISA-95.00.01-2000, Enterprise-Control System Integration - Part 1-5, 2007.

TABLE I. A COMPARISON OF REFERENCE ARCHITECTURES FOR INTERNET OF THINGS

Reference Architectures	Technology	Process	Quality and Key System Concerns	Business and People
RAMI4.0	A key concept is I4.0 component. Service-oriented and layered architecture. Follow and extend IEC62264 and IEC61512 standards. Permit encapsulation of functionalities. Standards compliant.	Address product lifecycle management dimension, horizontal integration across factories and vertical integration within factory. Allow step by step migration to I4.0 components.	Address security for functionality and data, functional safety and safety measures. The I4.0 component possesses the quality of service properties necessary for specific applications.	Address people orchestrating the value stream, and value stream dimension throughout product lifecycle and across factories.
IIRA	Key concepts include concern, stakeholder, and viewpoint. Based on ISO/IEC/IEEE 42010:2011. Standard-based open architecture.	Integration of information technologies and operational technologies.	Address safety, security, trust and privacy, resilience, integrability, interoperability and composability, connectivity.	Business viewpoint to address business vision, value proposition and objectives.
IoT-ARM	Key concepts include aspect-oriented programming, model-driven engineering, views and perspectives. Evolution and interoperability are the main drivers for the reference model and architecture.	Provide guidelines and process steps on how to generate concrete architectures, perform IoT threat analysis, and derive design choices and tactics based on qualitative requirements.	Address evolution and interoperability, performance and scalability, trust, security, privacy, availability and resilience.	Business goals, cost and benefit analysis are used in the architecture generation process. Specification of an IoT business process model to make use cases IoT-ARM compliant.
P2413	Key concepts include concern, stakeholder, and viewpoint. Based on ISO/IEC/IEEE 42010:2011 standard.	Provide guidelines for cross-domain interaction, documenting and migrating architecture divergence.	Address system interoperability, functional compatibility, protection, security, privacy and safety.	People perspective is reflected in the process of identifying stakeholders and their concerns.
Arrowhead	Key concepts include local cloud, global cloud. Automation cloud integration based on service-oriented architecture. Information centric.	Provide maturity levels of legacy system migration to cloud, engineering tools for development, and test support of cloud automation systems.	Address service interoperability and integrability, security, latency, scalability, dynamic/continuous engineering.	Not explicit
WSO2	Influenced by open-source projects and technologies	Not explicit	Address connectivity and communications, device management, data collection, analysis and actuation, scalability, security, and integration.	Not explicit
Azure IoT	Key principles include heterogeneity, security, hyper-scale deployments, and flexibility. Data concepts include device and data model, data streams, and device interaction.	A vendor-specific solution architecture	Not explicit	Business systems integration layer and solution UX are two architecture components relevant to business and people.
Internet-of-everything	A key concept is edge-ware. Multilevel model for IoT;	Integration of information technologies and operational technologies; enablement of legacy applications.	Address interoperability, security, and legacy compatibility.	Application layer covering business intelligence and analytics. Collaboration and processes layer explicitly involves people and processes.
Intel IoT	Building blocks include things, networks, and cloud.	Integration of information technologies and operational technologies.	Address data and device connectivity, security, and interoperability.	Value proposition by smart decision making based on data analytics.

Voices from Venezuela: Examining Blogs to Study the Socio-Politico-Economic Crisis and Consequent Emigration

Esther Ledelle Mead, Muhammad Nihal Hussain, Mohammad Nooman, Samer Al-khateeb, Nitin Agarwal†
†Jerry L. Maulden-Entergy Chair Professor of Information Science
University of Arkansas at Little Rock, Little Rock, United States
{elmead, mnhussain, msnooman, sxalkhateeb, nxagarwal}@ualr.edu

Abstract—**The objective of this research is to continue our journey into determining whether the blogosphere, as a type of social media platform, can be used to disseminate information regarding the socio-political views and concerns of citizens within a specific community. We expand upon our example case of focusing on information relative to the Venezuelan community regarding the current Venezuelan socio-economic crisis. Are Venezuelan blogs being used to discuss socio-political events and the quality of life concerns that are associated with the economic crisis in that community? Are Venezuelans using blogs to discuss possible migration away from the region as a result of these concerns? The Blogtrackers tool was used to analyze almost 30,000 Venezuelan blog posts collected between August 2003 and March 2017. Our analysis showed that the blogosphere is indeed being used as a platform by citizens to discuss these issues. We show how the posting frequency, sentiment, and keyword trends have changed over time relative to the changes in the socio-political landscape of the region and the events surrounding them. Of particular interest is the keyword trend analysis that shows that blogs are being used to discuss issues associated with quality of life factors and interest in migration away from Venezuela as a result of the crisis. We believe that this study can be used as a starting point to show the value of analyzing blogs in helping to identify humanitarian needs, in facilitating assistance efforts, and in drafting policy decisions.**

Keywords- blogosphere; situation awareness; socio-political; social media data analysis; migration; Venezuela; Blogtrackers.

I. INTRODUCTION

This paper is an expanded version of the earlier work by Mead et al. [1], which established the basis for using blog analysis for studying socio-political awareness. Social media platforms, such as Twitter, Facebook, YouTube, and blogs have changed the way citizens express and share their sentiments regarding socio-political situations within their communities, and have created a space for citizen journalism [2]. These new mediums of communication have also allowed for citizen sentiment to be channelized from the online forum to the streets in the form of public debates and protests. Tufekci and Wilson [3], for example, demonstrated that participation in protests, both before and on the first day of the Tahrir Square demonstrations was elicited by information that citizens posted on blogs, Facebook, and Twitter.

A blog is a "personalized webpage, kept by the author in reverse chronological diary form" [4]. Blogs give people a social identity and are a medium for association, self-expression, and dissemination of information [5]. Blogs can be a very effective way to gain an in-depth understanding of issues and events due to the fact that recording, revisiting, and reflecting on the past is possible through previous blog entries. The ability for readers to leave comments in an interactive environment is an important part of blogging.

Blogs serve as an interactive platform for information exchange and discussion, and provide useful information about events [6]. Blogs serve as a way for citizens to gain situational awareness of the socio-political landscape of their environment, and data experts who track and analyze blogs can gain an understanding of the perspectives and intentions that exist among citizens. The Blogtrackers tool [7] was used to analyze a large dataset of Venezuelan blogs to determine whether the blogosphere is being used to disseminate information about issues stemming from the Venezuelan socio-economic crisis, and, if so, how this content is changing over time. This information is particularly helpful for emergency responders, and policy and decision makers leading humanitarian assistance efforts.

The rest of this paper is organized as follows. Section II reviews the currently published literature and related work regarding citizen use of social media platforms relative to the attainment of situational awareness of socio-political issues. Section III explains the methodology used such as the data collection process and the Blogtrackers tool (http://blogtrackers.host.ualr.edu/Blogtrackers/). Section IV provides a discussion of the data analysis results. Section V concludes the study outlining future research directions.

II. RELATED WORK

In this section we discuss the various previously published work related to the analysis of social media data as a means for studying socio-political awareness.

A. Informational Power of Social Media Platforms

Many sources highlight the power of social media platforms such as blogs, Twitter, Facebook and YouTube to be effective tools for allowing citizens to engage in socio-political scenarios such as obtaining information, disseminating information, participating in socio-political discussions, and becoming mobilized to act or to participate in impactful events [8, 9, 10, 11, 12, 13, 14, 15, 16]. Additionally, some sources focus on or add that the effectiveness of social media platforms as informational and motivational tools can be leveraged by organizations for crisis management and emergency response [8, 11, 15, 17, 18, 19, 20, 21]. Much of the literature highlights the power

of social media platforms by using the example of the 2010 protests in Egypt, followed by an Egyptian Revolution in 2011, wherein the public used social media to communicate their dissatisfaction with socio-political issues such as poverty, unemployment, corruption, high prices, repression, and human rights abuse [9, 10, 11, 16, 22, 23].

B. New Media Versus Traditional Media

Many sources refer to social media platforms as "New Media" [13, 16, 24], "Participatory Media" [9, 25], and "Modern Information and Communication Technologies (ICTs)" [13, 16, 22, 23], and draw a contrast with "Traditional Media" platforms such as television, newspaper, and radio (especially, state-controlled traditional media) [9, 11, 13, 14, 16, 22, 24, 25]. Valentini et al. [14] showed that crises tend to be communicated or framed differently within the writings of new media (specifically blogs) than they are within the writings of traditional media. Some sources contend that citizens are increasingly using social media sites for obtaining and disseminating socio-political information due to the perceived or real incompetence of or censoring by state-owned media outlets [16, 22, 24].

C. A Focus on the Blogosphere

Using the possibility of migration due to the Venezuelan economic crisis as a case study, Mead et al. established the basis for using blog analysis for studying socio-political awareness [1]. Before Mead et al., only a small pool of sources focus specifically on blogs as a means for obtaining and disseminating socio-political information [22, 26, 13, 14]. The number of blogs has been said to double every five months; they are easily found and can be accessed freely by anyone with an internet connection [13]. Al-Ani et al. [22] highlighted the power blogs played in mobilizing citizens during the Egyptian revolution of early 2011. Blogs represent a "counter-narrative" to the government-controlled media especially during times of crisis, and provide a means to voice dissent and to challenge authoritative power [22]. Blogs provide a means to potentially develop and maintain a strong sense of community among citizens interested in certain themed topics [26]. Additionally, blogs often-times represent "citizen-based news sources" that challenge traditional media in terms of the ability to form public opinion [14]. The information posted on blogs and the commentary reactions to the posts, therefore, have become of increasing interest to social media researchers.

D. Cybersecurity Issues

Some sources highlight the potential cybersecurity issues related to the use of social media platforms for spreading situational awareness in terms of socio-political issues [27, 25, 11]. Fearn [27] warns that "cybercriminals" are increasingly using programs called "bots" to attack social media users via Twitter, Facebook, and YouTube with negative comments and to spread disinformation or fake news via these platforms. Since the blogosphere is also a social media platform, the potential for blogs to be used nefariously by bots can be argued [25]. Goolsby et al. [25] also argue that social media can be used to broadcast "hoax messages", which can "hide among the stream of natural messages and be accepted…".

E. Methodologies for Analyzing Blog Data

A review of the literature shows a variety of methodologies for conducting data analysis on blog data. Al-Ani et al. [22] utilized the technique of topic modeling on blog data to ascertain how blog topics changed over time between 2004 and 2011. The topic modeling technique revealed time-specific themes in the blog data that could then be compared with public events and citizen actions such as protests [22]. Berendt et al. [26] used text and graph mining to analyze blog data. The text mining revealed the word themes in the data; whereas the graph mining exposed the connections between the bloggers in terms of their use of these words/themes [26]. In addition to longitudinal content analysis, Valentini et al. [14] applied sentiment analysis to blog data wherein they attempted to assign the sentiment values of neutral, positive, or negative to each blog post. Their analysis revealed interesting time-relevant topics and associated sentiments, such as a strong lack of trust in public and private governing entities surrounding political issues [14].

F. Migration Potential Resulting from Socio-Political Crisis?

Google [28] created an interactive data visualization for tracking the interest in migrating to various countries of destination from specific countries of origin. Of particular interest is the data visualization showing a ranking of 1 to 211 based on user search counts about migrating to the United States from within a particular country. For example, Venezuela is "ranked 105 out of 211 countries for Google searches for immigration to the United States from 2014-2015" [28]. According to numerous sources, Venezuelans are protesting against their government due to such reasons as unemployment or low income, lack of access to basic needs such as food and medicine for themselves and their children, and political corruption [29, 11, 30]. Although the Google search data visualizations [28] precludes many of the Venezuelan issues that have been recently reported, one can still ask the question of whether or not there is potential for Venezuelans to migrate away from the many problems that they have been experiencing in their specific socio-political crisis? An analysis of numerous active Venezuelan-specific blogs was conducted to provide insights.

III. METHODOLOGY

In this section we discuss the methodology for this study including how the data set was obtained and the data analysis methods used.

A. Data Collection, Cleaning, and Indexing

Using the Blogtrackers tool, three steps were executed in order to crawl and collect the data from an identified set of Venezuelan blogs: (1) exploring the blog site, (2) crawling the blog site, and (3) cleaning and storing the data in a database for analysis and retrieval (Fig. 1). Hussain et al. provide a detailed explanation of the mechanics of the

Blogtrackers tool in "Analyzing the Voices during European Migrant Crisis in Blogosphere" [31].

1) Exploring the blog site: Several blog sites were identified that specifically discussed issues relevant to Venezuela. Subsequently, each site was explored to determine whether their structure was ideal for use with the Blogtrackers tool. It was also important that the blog continually be focused on Venezuelan topics and contain specific metadata attributes for each post such as author, title, and date.

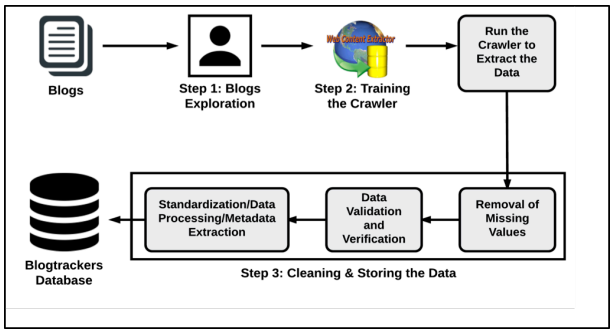


Figure 1. Data collection process.

2) Crawling the data: The Web Content Extractor (WCE) tool [32] (Fig. 2) was then used to collect data from each blog site. Once the crawler is set up, the tool begins from a set of seed URLs—the blog sites' home pages—and advances through each blog post to extract all of the desired attributes.

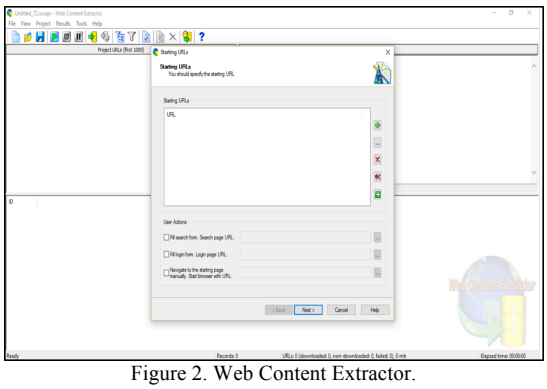


Figure 2. Web Content Extractor.

3) Cleaning and Storing the Data: A three-step cleaning process was used. (1) Clean from within WCE by deleting empty fields and advertisement URLs. (2) Clean with SQL queries to select validated and verified data. (3) Clean with a script to standardize attributes, extract metadata, sentiments, and outbound URLs.

B. Analyzing the Data with Blogtrackers

Blogtrackers is a tool designed to explore the blogosphere and gain insights about events and how these events are perceived in the blogging community [33]. After setting up a Venezuelan blog tracker, five features of the Blogtrackers tool were used to analyze the resultant dataset.

1) Posting Frequency

The "Posting Frequency" feature was utilized to identify any unusual patterns in the blog postings. This aids in detecting real-time events that interested the blogging community. The user can click on any data point on the graph to get a detailed list of the named-entities mentioned in blog posts during that time-period. This feature also displays a list of active bloggers with number of posts. Fig. 3 shows the posting frequency for Venezuelan blogs from 2003 to 2017.

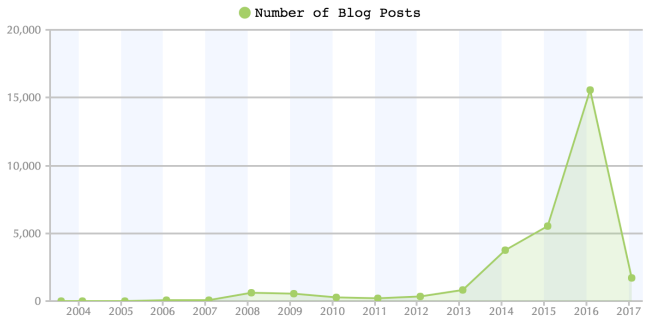


Figure 3. Change in Venezuelan blog posting trends from 2003 to 2017.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

2) Keyword Trends

The "Keyword Trends" feature was used to provide an overall trend for keywords of interest. It helps track changes in topics of interest in the blogging community. The user can select any data point on the trendline to view all the blogs and a network of co-occurring named-entities. Fig. 4 shows the keyword trends related to the ongoing Venezuelan socio-political crisis.

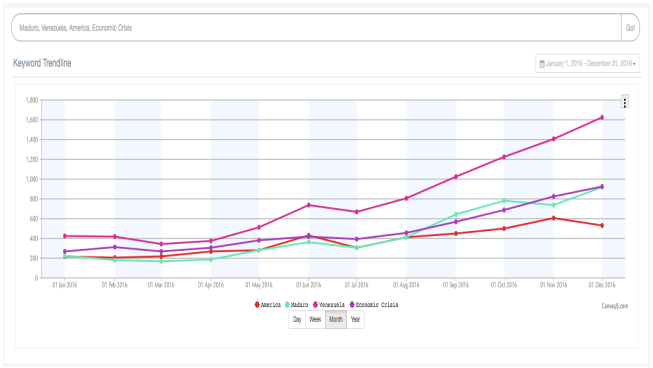


Figure 4. Trends for keywords "Venezuela", "Maduro", "America" and "Economic crisis" for 2016.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

3) Sentiments and Tonality

The "Sentiments and Tonality" feature was used to display the trend of positive and negative sentiments of blogs for a selected time-period (Fig. 5).

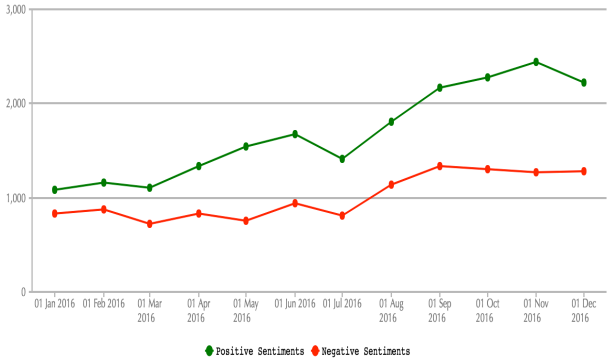


Figure 5. Venezuelan blog sentiment trends for 2016.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

The sentiment and tonality features are defined by Pennebaker et al. [34, 35] and as calculated by the Linguistic Inquiry and Word Count (LIWC) software [36]. Additionally, a data analyst can drill down by clicking on any point of interest and view radar charts (Fig. 6), which display tonality attributes such as personal concerns, time orientation, core drives, and cognitive process.

Figure 6. Tonality of two random Venezuelan blog posts.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

4) *Influence*

The "Influence" feature was used to identify the influence of a blogger or a post. Agarwal et al. discussed the concept of bloggers' *influence* [6]. The "Influence" feature of the Blogtrackers tool can display the influence trends over time for the top 5 influential bloggers (Fig. 7). Clicking on a point on the trend line allows a deeper dive into the data by displaying the most influential posts for that period. Additionally, a user can explore the content themes of active-influential, inactive-influential, active-non influential, and inactive-non influential bloggers.

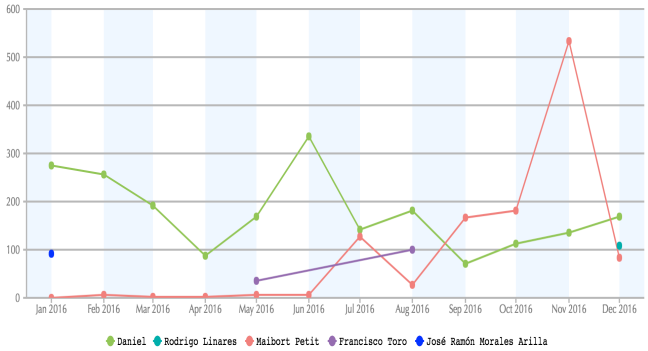


Figure 7. Influence trend for top 5 bloggers for 2016.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

5) *Additional Blog Information*

The "Additional Blog Info" feature was used to provide additional information about a specific blog. A dashboard-like screen is presented to the user, revealing the posting trends and sentiments of the selected blogs, as well as a list of the underlying URLs and domains. Fig. 8 shows this feature being used to look at a specific blog from the database called, "Caracas Chronicles". At a glance, we can see some interesting things about this particular blog; such as, it is most active on Monday's; it was most active during the month of February in 2014, December in 2015, and April in 2016; and there was more negative sentiment in 2016 than in 2015.

Figure 8. "Additional Blog Info" Blogtrackers feature for selected Venezuelan blog.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

IV. DISCUSSION

In this section we describe the database of Venezuelan blog posts that we collected in accordance to our above-mentioned methodology, and we discuss details from our data analysis methods.

A. Venezuelan Blog Database

To assess whether Venezuelan blogs were discussing issues related to the economic crisis in Venezuela, 40 blog sites were identified. The blogs were found using simple manual search techniques on various platforms, such as google.com, blogsearchengine.org, and fastblogfinder.com. The blogs were reviewed by our research team to ensure that they fit the structure required by the WCE and the Blogtrackers tool. A final dataset of 29,493 blog posts was obtained between August 27, 2003 and March 26, 2017. A total of 177,870 links were extracted (120,296 being distinct links) from 13,590 domains and 749,829 entities. The post sentiments were also extracted. Table I provides the language distribution for this dataset.

TABLE I. LANGUAGE STATISTICS

Language	Blogs	Blog Posts
Spanish	23	16,916
English	29	12,490
Italian	3	51
French	3	3
Portuguese	2	2
German	1	2
Breton	1	1
Catalan	1	1
Polish	1	1

B. Posting Frequency of Venezuelan Blogs

Fig. 9 shows a more detailed view of blog posting frequency from January 2015 to March 2017, which indicates a continuous increase. Specifically, blogging activity increased drastically between March 2016 and January 2017.

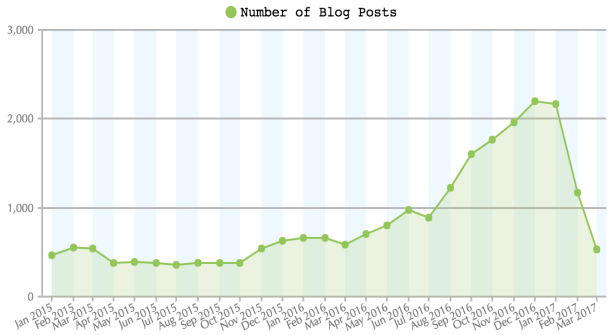


Figure 9. Change in Venezuelan blog posting frequency from January 2015 to May 2017.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions

This increase corresponds to news reports of hundreds of thousands of people beginning to take to the streets in protest during early September of 2016 [37]. Between the beginning of January, 2015 and the beginning of March, 2017, blogging frequency can be said to have notably risen by the beginning of June, 2016 (increasing by about 50%) (Fig. 9).

C. Keyword Trendlines for Venezuelan Blog

To further assess the extent of the impact of the crisis on Venezuelan citizens, the dataset was searched for quality of life keywords such as “need food”, “need water”, “need petrol”, “need medicine”, “high prices”, and “inflation”. The resultant keyword trendlines indicated that the occurrence of these quality of life factors fluctuated over time (Fig. 10). The occurrence of English keywords related to quality of life such as “need food”, “food shortage”, “high prices”, “need water”, “need petrol”, “need medicine”, and “inflation” all began to notably increase beginning around January 1, 2012 (Fig. 10).

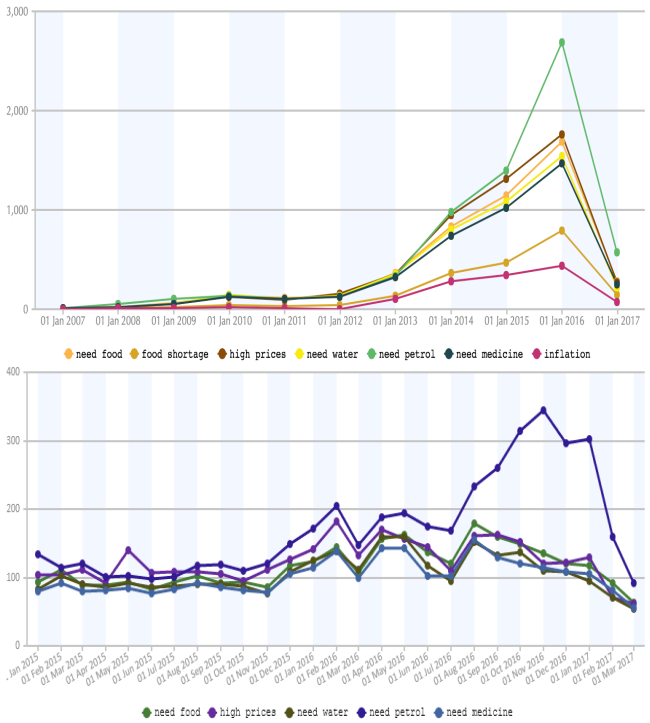


Figure 10. English keyword trendlines for various quality of life keywords over time.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

Trading Economics data shows that beginning January, 2012, crude oil prices began to decline and the Venezuelan monetary exchange rate plummeted [39, 40]. Of the above-mentioned keywords, “need petrol” was the most prevalent since around mid-2013 through January 1, 2017, spiking dramatically in early January of 2016. On February 18, 2016, CNN Money reported that on the previous day, Venezuelan President Nicholas Maduro announced a 6,000% price increase for gasoline/petrol in Venezuela [41]. Additionally, of these keywords, “high prices”, “need food”,

"need water", and "need medicine" were the next highest ranking, respectively. This ranking order of occurrence remained the same from early January 2014 through the end of 2016. Beginning early January 2016, however, the occurrence of all of these keywords began to decline through early January of 2017. A broader look from January 2015 through March 2017 reveals a spike in the occurrence of the keyword "high prices" in early May 2015 and again in early February 2016. The spike in early May 2015 may correspond to incidents of raids on Venezuelan markets by "black-market foot soldiers" beginning in early April [42]. This broader view also shows a spike in the occurrence of "need petrol" in early February 2016. The count declines from the rest of February 2016 to early March 2016, then begins to climb again through early May 2016. There was a drastic increase beginning early July 2016, which peaked in early November 2016. There was steady decline from early January 2017 to March 2017, where the occurrence of the keyword "need petrol" returned to levels closer to the other noted keywords. The keyword "need food" broke out to become the second-highest occurring keyword beginning early July 2016 (recall the NYT report of Venezuelans ransacking stores). It held this position through March 2017, closely followed by "high prices", which exceeded it slightly and briefly in early January 2017. The occurrence of "need medicine" also rose significantly from early July 2016 through early August 2016. The occurrence of "need water" rose to the fourth-highest position from early September 2016 through late October 2016.

Trendlines for the Spanish equivalents for these quality of life keywords were also generated (Fig. 11). For example, similar to the increase in blog posting frequency, the quality of life keywords trends figures each show a large increase in frequency beginning near early September of 2016 [37, 43]. The occurrence of the Spanish equivalents of many of these keywords related to quality of life such as "escasez de alimentos", "necesito agua", "inflacion", "cuidado de la salud", "necesito comida", "altos precios", "necesito medicina", and "necesito gasolina" also saw variation over time (Fig. 11). The occurrence of the Spanish keyword "inflacion" (inflation) did not register on the graphical output for January 1, 2003 to January 1, 2017, and "necesito gasolina" (need gasoline) barely registered. The occurrence of "escasez de alimentos" (food shortage) and "cuidado de la salud" (health care) mirrored each other, beginning to increase in early January 2013, increasing further sharply in early January 2015, peaking in late December 2016, and dropping sharply through early January 2017. "Necesito agua" (need water) consistently held the second-highest occurrence count notably from early January 2014 through early January 2017, with a breakout spike between early January to late December 2015. Also through this early January to late December 2015 timeframe, "altos precios" (high prices), "necesito medicina" (need medicine), and "necesito comida" (need food) held the next highest count rankings, respectively. "Necesito agua" (need water) saw a spike in occurrence count between early January 2008 and early January 2009. A broader look at the occurrence of the Spanish keywords from January 2015 through March 2017 reveals that "cuidado de la salud" (health care) and "escasez de alimentos" (food shortage) consistently mirrored each other and held the top rank for most occurring of the quality of life keywords. "cuidado de la salud" and "escasez de alimentos" occurred at least 50% more of the time than the others from early January to early March 2015, then beginning again from mid-November 2015. Although experiencing a short decline from early June to early July 2016, the count of these two keywords continued to rise through their peak in early December 2016. They began to decline sharply from early January 2017, returning to near January 2015 levels by early March 2017. "Necesito agua" began in a breakout rise in early May 2016, with a mild spike in early June 2016 (corresponding to the NYT's story of ransacking stores), another spike in early December 2016, then a slow decline through early March 2017. "Altos precios", "necesito medicina", "inflacion", and "necesito comida" (in that apparent ranking order) only began to register on the data visualization from around early October 2016 through around early January 2017. "Necesito gasolina" barely registered on the data visualization in this 14-month period except from early November 2016 to around mid-February 2017.

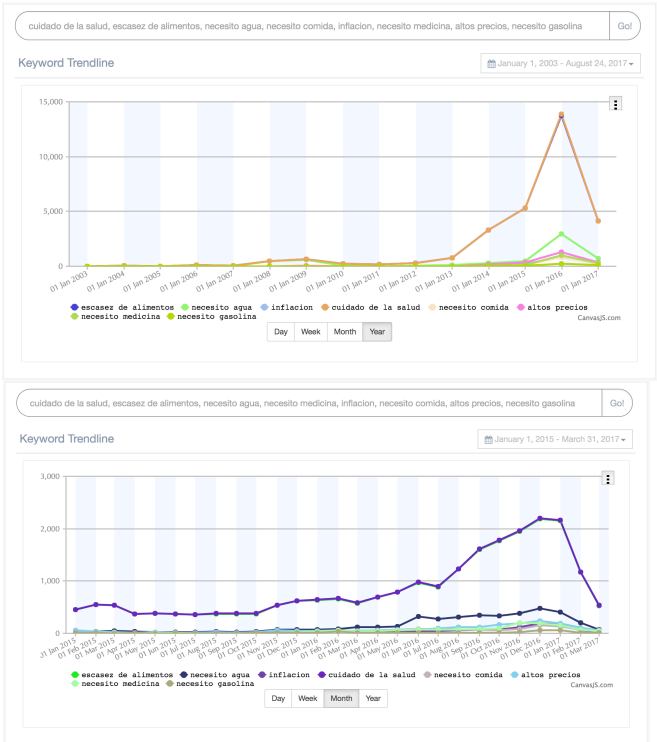


Figure 11. Spanish keyword trendlines for various quality of life keywords over time.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

These quality of life keywords may represent a set of motivational factors leading to an interest in Venezuelan citizens migrating away from Venezuela. To explore this suggestion, we generated several trendlines for keywords such as "immigrate", "migrate", "emigrate", and "leave Venezuela" (Fig. 12).

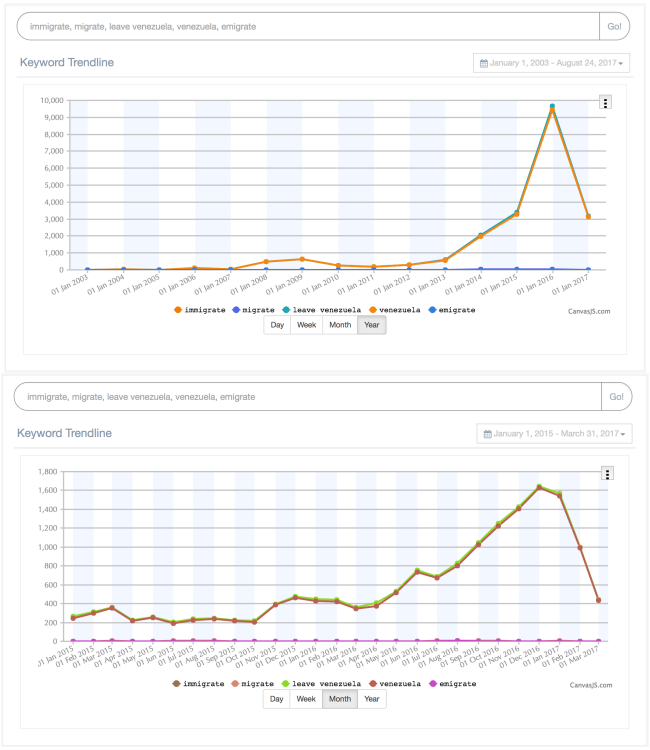


Figure 12. English keyword trendlines for various keywords that possibly indicate migration interest over time.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

The frequency of occurrence of the keyword, "leave Venezuela" seems to drastically increase between the end of 2015 and beginning of 2016. The keyword "leave Venezuela" began to increase dramatically in occurrence beginning early January 2013 through its peak of almost 10,000 posts in late December 2015, with the most notable rise being between early January 2015 and late December 2015. According to a report by United Press International, in 2016, 75% of the Venezuelan population lost an average of 19 pounds due to the effects of the economic crisis that was satirically referred to as the "Maduro Diet" [44]. By early January 2017, occurrences of "leave Venezuela" dropped down to about 3,000 (Fig. 12). The bottom chart of Fig 12 kind of zooms in to show that the peak in occurrence of the keyword "leave Venezuela" actually occurred in early December 2016.

Trendlines for the Spanish equivalents for these migration keywords were also generated (Fig. 13). It was revealed that the keywords "leave Venezuela" and "Venezuela" were almost identical, indicating that every time Venezuela was mentioned it was about leaving or migrating from Venezuela. As with the quality of life indicators, these migration-related keywords trends figures show a large increase in their frequency beginning in 2016, especially near early-September. Of the Spanish equivalents of English keywords that may be indicative of a desire to flee Venezuela, "salir de Venezuela" (leave Venezuela) was the only one that notably registered on the data visualization between early January 2003 and early January 2017 (Fig. 13).

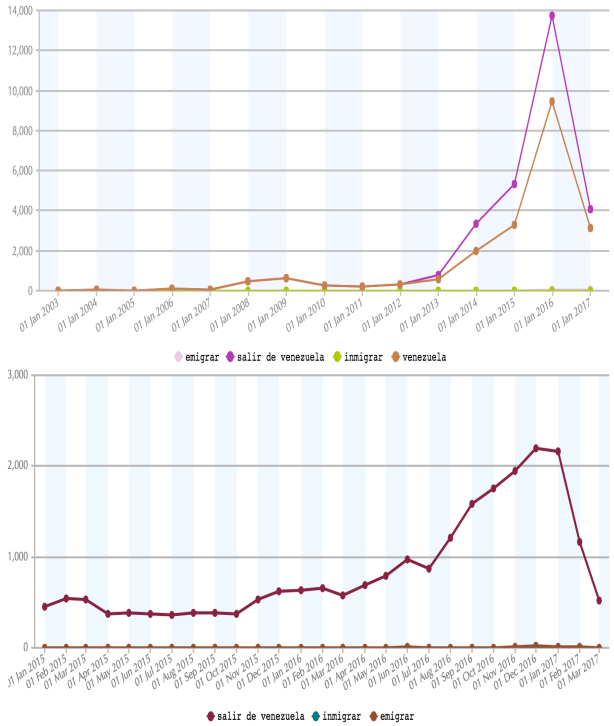


Figure 13. Spanish keyword trendlines for various keywords that possibly indicate migration interest over time.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

The others, "emigrar" (emigrate) and "inmigrar" (immigrate), barely registered. Occurrences of the keyword "salir de Venezuela" began to increase in early January 2012 (recall the Trading Economics data showing the decline in both crude oil prices and the Venezuelan monetary exchange rate), peaking in late December 2016 near 14,000, then beginning to decrease sharply through late December 2016 or early January 2017. According to a report by CNN Money, there was a 160% increase from 2015 to 2016 in Venezuelans seeking U.S. asylum [45]. The most notable increases in the occurrence of the keyword "salir de Venezuela" (each at different rates) were between early January and late December 2013, early January 2014 and early January 2015, and early January to late December 2015. Between early January 2015 and early March 2017, "salir de Venezuela" occurred at least 50% more of the time than did the keywords "inmigrar" (immigrate) and "emigrar" (emigrate). At its peak for this date range, "salir

de Venezuela" occurred over 200% more often than "inmigrar" and "emigrar". There was a slight decline in the occurrence of the keyword "salir de Venezuela" from early June to early July 2016. Subsequently, however, there was a resumption of the sharp rise of the occurrence of "salir de Venezuela" in early July 2016 (again, recall the NYT's story of the ransacking of Venezuelan stores). The keyword "emigrar" (emigrate) began to register slightly more on the data visualization from around early November through around mid-December 2016.

D. Timeline of Sentiments Among the Venezuelan Blogs

Recall that Fig. 5 shows the prevailing sentiment of the blogs for a specific time period. Fig. 14 provides a more detailed view of how the sentiment of the Venezuelan blogs has changed over time. The graphical output appears to show a sharp shift to a prevailing positive sentiment in early January 2016. This could correspond to the Venezuelan leader, Nicholas Maduro, declaring a 60-day economic emergency for the region in order to give himself power "to pay for welfare services and food imports" [43]. Subsequently, however—and although many blog posts were identified as having a negative sentiment—the prevailing sentiment of the blogs has been positive for the past two years. This prevailing positive sentiment seems counterintuitive, as we expected to see a prevailing negative sentiment due to the ongoing Venezuelan economic crisis and continued reports of Venezuelan citizens protesting in the streets. Therefore, we believe that the concept of sentiment with regard to this dataset needs to be addressed further in future work involving a more detailed dataset and a possible revision of how sentiment is calculated. Consequently, for this paper, we can only conclude that the graphical output at this time does not seem to reveal any significant differences over time with regard to the concept of sentiment, and appears to only fluctuate along with the recorded count of blog posts. Regardless, the visualization shows that between early January 2015 and early March 2017, the overall sentiment of the Venezuelan blog posts was predominantly positive (there were more blog posts rated "positive" than there were those rated "negative") (Fig. 14). The difference (spread) between the number of Venezuelan blog posts rated "positive" and the number of blog posts rated "negative" was relatively small (from about 50 to 150) from early January through around mid-November 2015. Although the number of blog posts rated "positive" and the number of blog posts rated "negative" both continued to increase between mid-November 2015 and early January 2017, the spread widened most notably between early April and early June 2016, then again between early September 2016 through early January 2017. Between mid-year 2007 and mid-year 2009, there were more Venezuelan blog posts rated "positive" than there were blog posts rated "negative". Between early January 2010 and early January 2012, the number of blog posts rated "positive" and the number of blog posts rated "negative" were roughly equal. In early January 2013, the spread (difference) between the number of blog posts rated "positive" and the number of blog posts rated "negative" began to widen, with more being "positive". Again, this seems odd in light of the fact that Venezuela entered into a deep recession beginning January 2014 [46].

Although fluctuating in size, this positive-dominated spread continued through at least early January 2017. The bottom chart of Fig. 14 shows the number of blog posts rated as having a "positive" sentiment relative to the number of blog posts rated as having a "negative" sentiment for each day from May 2, 2016 to June 30, 2016.

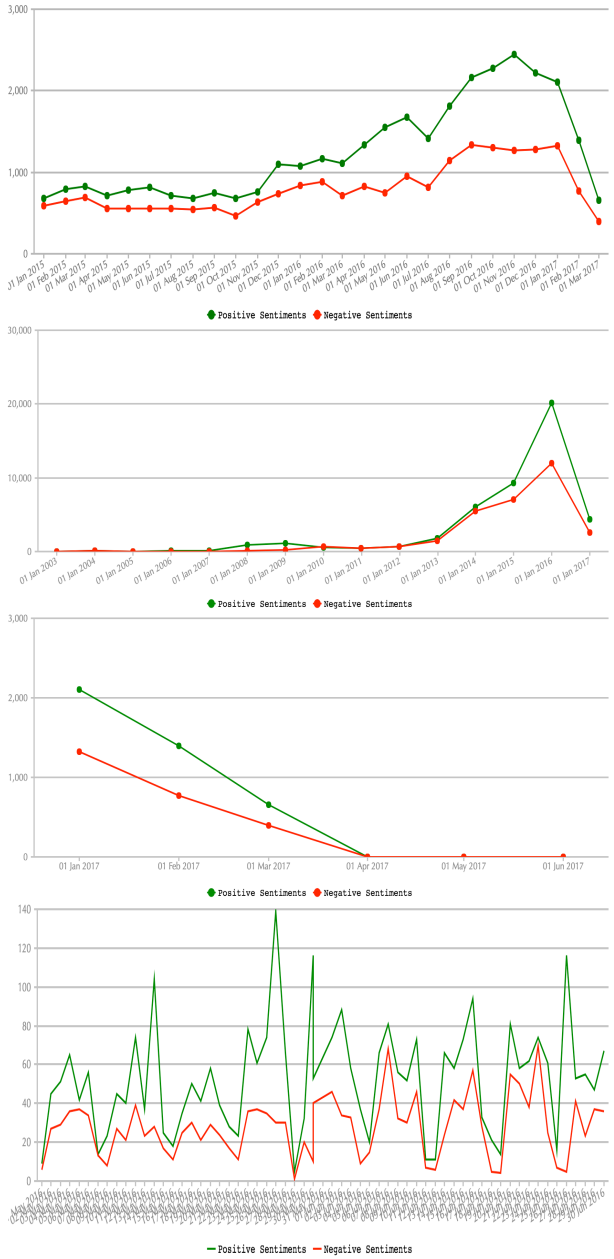


Figure 14. Change in sentiment of Venezuelan blogs over time.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

During this time range, the number of blog posts rated "positive" always exceeded the number of blog posts rated "negative", although only slightly for at least seven dates. On at least four dates, the number of blog posts rated "positive" sharply exceeded the number of blog posts rated "negative"; for example, on May 13, May 26, May 30, and June 26. Reuters reported that on May 13, 2016, Maduro declared a 60-day "state of emergency" [47]. News reports for May 26, 2016 also include topics related to the "state of emergency" declaration; Reuters reported that Venezuelan opposition to Maduro, essentially, warned him against corruption related to contracts with foreign companies [48]. On May 30, 2016, BBC News reported that eleven people were killed in a recession-related shooting in Trujillo state [49]. Finally, on June 26, 2016, NPR broadcasted a story on its "All Things Considered" radio program discussing how Venezuelans are running out of food and medicine [50]. At least one date stands out on the graph where the number of blog posts rated "negative" rose to almost reach the number of blog posts rated "positive"—June 23, 2016. The Washington Post reported that on that date (June 23, 2016) the Organization of American States met and argued for a recall referendum against Maduro [51]. Curiously, the bottom-most chart of Figure 14 seems to show that on May 28, 2016 there were barely any blog posts. This may be due to the system of scheduled electricity blackouts that the Venezuelan government was implementing as reported by The NYT on this date [52].

E. Tonality of Venezuelan Blogs

Recall that Fig. 6 displays the feature of Blogtrackers that shows tonality attributes of individual blog posts. For example, for two random Venezuelan blog posts in our dataset, the predominant personal concerns were "Work" and "Money". The predominant time orientation was "Present focus". The predominant attribute for core drives and needs was "Power" (and to a lesser extent "Achievement"). The predominant cognitive process was not as clear, varying among that of "Differentiation", "Tentativeness", and "Cause". The predominant summary variable was "Analytical Thinking". Finally, the predominant sentiment/emotion for this timeframe was "Anger". We did not analyze the tonality feature further for this paper, but believe that the concept should be examined in future work.

F. Influential Venezuelan Bloggers

Recall that Fig. 7 shows the top 5 influential bloggers for a specified time period. Blogtrackers calculates influence based on the definitions and algorithms developed by Agarwal et al. [25]. Fig. 15 is another example of the feature, using January 2015 to March 2017 as the selected time-period of analysis. One blogger, "Daniel", was consistently more influential than other bloggers [53]. We did not analyze this feature further for this paper, but believe that the concept of influence within the blogosphere should be examined in future work. With fluctuating levels of influence, between late December 2015 and early March 2017, the five most influential authors of the Venezuelan blogs in our database were "Francisco Toro", "mesaredondacontracomunistabg", "Juan Cristobal Nagel", "Manuel Madrid" and "Daniel" (Fig. 15). "Manuel Madrid" only seemed to be active during early December 2015. "Francisco Toro" apparently didn't begin activity until early June 2015, and ceased being active in early January 2017. "Juan Cristobal Nagel" apparently ceased being active in early June 2016 (again, this date corresponds to the NYT's story of Venezuelans ransacking stores). The most consistent and predominantly influential blogger was "Daniel". "Daniel" was notably the most-influential blogger during at least three key date categories: early March 2015, early December 2015, and early June 2016.

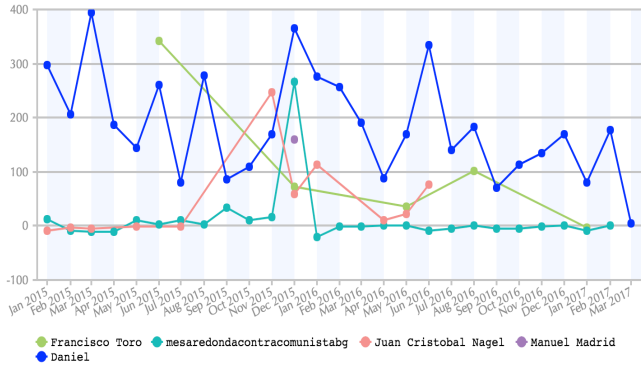


Figure 15. Influence trend of top 5 Venezuelan bloggers from January 2015 to March 2017.
Reader is encouraged to use digital version for better readability of images, labels, legends, and captions.

G. Socio-politico-economic Awareness

One event can be said to have corresponded to numerous spikes (or otherwise noticeable changes in various data categories—early June of 2016. June 19, 2016 is when news reports emerged of Venezuelans ransacking stores out of desperation stemming from food shortages and high prices. This event seemed to spark bloggers to increase their posting frequency. This event also seemed to cause bloggers to increase their use of the keywords "necesito agua" (need water). Additionally, and although the data show a seeming decline from early June to early July of 2016, the use of the keyword "salir de Venezuela" (leave Venezuela) quickly resumed in frequency in early July of 2016; perhaps communicating an increasing level of frustration with the economic conditions. This pattern with the frequency of the keywords "salir de Venezuela" comes on the heals, however, of another data pattern that popped out in the visualizations that show that the frequency of the use of the keywords "cuidado de salud" (healthcare) and "escasez de alimentos" (food shortage) experienced a short decline from early June to early July of 2016. Perhaps this means that instead of blogging about these two topics, the bloggers were discussing the possibility of migration. The data also showed

that the number of blog posts rated as having a positive sentiment began to decline around the time of this event. In fact, the number of blog posts rated as having a negative sentiment rose to almost reach the number of blog posts rated as having a positive sentiment around June 23, 2016. Another characteristic of the data that seemed to correspond to this event is that one of the top five most influential bloggers, "Juan Cristobal Nagel" apparently ceased being active in early June 2016, which, although speculation, causes one to wonder if these harsh economic conditions were behind his exit from the Venezuelan blogosphere. Ultimately, however, these patterns in the data seem to suggest that indeed bloggers, especially the Spanish-language bloggers, are raising awareness of these socio-political events.

V. Conclusion & Future Work

This paper extends the work by Mead et al. that established the basis for using blog analysis for studying socio-political awareness. This approach remains novel in that few researchers have specifically focused on analyzing blogs, and instead focus on other social media platforms, such as Twitter, Facebook and YouTube. As a detailed example, this research showed that Venezuelan blogs are being used to disseminate socio-political information in an attempt to increase awareness of events and sentiments during the Venezuelan economic crisis. Our analysis showed that the frequency and content of blog posts changed over time, reflecting changes in the socio-politico-economic landscape of the region—such as protests and other news events, the documented decline in quality of life factors such as the need for food and medicine, and expressed interest in migration away from Venezuela. The sentiment of the blogs seemed to change over time as well, but the overall sentiment analysis was inconclusive and seemed counterintuitive and therefore the concept needs to be addressed further in future work. For example, the number of blog posts that were rated as having a positive sentiment far exceeded the number of blog posts rated as having a negative sentiment even when the dates revolved around significant events such as on June 26, 2016, which was only a few days after news reports emerged of Venezuelans ransacking stores out of desperation. We believe, however, that blog analysis—with Blogtrackers and other tools—can continue to be used to gauge socio-political awareness of important issues among various populations. This paper continues to develop the stage for future work using Blogtrackers and other natural language processing tools and techniques for blog analysis as a possible approach for anticipating events (e.g., protests, migration, refugee scenarios). Future work may also include further analysis of the concepts of blog tonality and blogger influence. Specifically, further understanding blogger influence is of great importance due to the emergent concepts of fake news, opinion manipulation, and disinformation campaigns. Broadly speaking, however, this particular study sheds a spotlight on the blogosphere's role in assessing situation awareness of a region engulfed in socio-politico-economic crisis. We believe that the information derived from monitoring blogs can provide actionable insights to local emergency responders (i.e., real-time blog monitoring), and humanitarian assistance organizations and policy decision makers (i.e., based on short- and long-term trend analysis).

Acknowledgment

This research is funded in part by the U.S. National Science Foundation (IIS-1636933, IIS-1110868 and ACI-1429160), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2605, N00014-17-1-2675), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059) and the Jerry L. Maulden/Entergy Fund at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

[1] E. L. Mead, M. N. Hussain, M. Nooman, S. Al-khateeb, and N. Agarwal, "Assessing Situation Awareness through Blogosphere: A Case Study on Venezuelan Socio-Political Crisis and the Migrant Influx," The Seventh International Conference on Social Media Technologies, Communication, and Informatics (SOTICS 2017), 2017, Oct. 8, 2017, Athens, Greece. ISSN: 2326-9294

[2] D. Gillmor, "We the Media: The Rise of Citizen Journalists," National Civic Review, vol. 93 (3), pp. 58–63, 2004, doi:10.1002/ncr.62.

[3] Z. Tufekci and C. Wilson, "Social media and the decision to participate in political protest: Observations from Tahir Square," Journal of Communication, vol. 62 (2), pp. 363-379, 2012, doi: 10.1111/j.1460-2466.2012.01629.x.

[4] H. S. Du and C. Wagner, "Learning with weblogs: An empirical investigation," presented at the 2005 Proceedings of the 38th Annual Hawaii international Conference on System Sciences, pp. 7b-7b, Jan. 24, 2005, doi: 10.1109/HICSS.2005.387.

[5] Y. Xie and P. Sharma, "Students' lived experience of using weblogs in a class: An exploratory study," Association for Educational Communications and Technology (AECT 2004), pp. 839-846, 2005, Accession Number: ED485009.

[6] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," Proc. 2008 International Conference on Web Search and Data Mining (WSDM 2008), pp. 207–218, ACM, 2008, ISBN: 9781595939272.

[7] Blogtrackers, "Analyze anything about blogs," University of Arkansas at Little Rock, 2017. [Online]. Available from: http://blogtrackers.host.ualr.edu/ 2018.05.26

[8] D. T. Cain, "Twituational awareness: gaining situational awareness via crowdsourced# disaster epidemiology." Diss. Monterey California: Naval Postgraduate School, Sep. 2013. [Online]. Available from: https://calhoun.nps.edu/bitstream/handle/10945/37594/13Sep_Cain_Daniel.pdf?sequence=1 2018.05.26

[9] A. R. Curtis, "From Arab Spring to Shahbag: The role of social media in terms of national crisis," Journal of Mass Communication Journalism, vol. 5 (2), pp. 1-3, 2015, doi: 10.4172/2165-7912.1000241.

[10] M. Golden Pryor, M. Wulf, W. Alanazi, N. Alhamad, and O. Shomefun, "The Role of Social Media in Transforming Governments and Nations," International Journal of Business & Public Administration, vol. 11 (1), pp. 19-30, Sum. 2014, ISSN: 15474844.

[11] V. Karagiannopoulos, "The role of the internet in political struggles: Some conclusions from Iran and Egypt," New Political Science, vol. 34 (2), pp. 151-171, May 2012, doi: 10.1080/07393148.2012.676394.

[12] J. B. Lim, "Engendering civil resistance: Social media and mob tactics in Malaysia," in International Journal of Cultural Studies, vol. 20 (2), Mar. 2017, doi: 10.1177/1367877916683828.

[13] N. Pang and D. Goh, "Can blogs function as rhetorical publics in Asian democracies? An analysis using the case of Singapore," Telematics and Informatics, vol. 33 (2), pp. 504-513, 2016, doi: 10.1016/j.tele.2015.08.001.

[14] C. Valentini and S. Romenti, "Blogging about crises: The role of online conversations in framing Alitalia's performance during its crisis," Journal of Communication Management, vol. 15 (4), pp. 298-313, 2011, doi: 10.1108/13632541111183398.

[15] L. J. Van Leuven, "Optimizing citizen engagement during emergencies through use of Web 2.0 technologies," Thes. Monterey California Naval Postgraduate School, Mar. 2009. [Online]. Available from: https://calhoun.nps.edu/bitstream/handle/10945/4819/09Mar_Van_Leuven.pdf?sequence=1 2018.05.26

[16] F. H. Abdullah, "The Role of Social Network Platform in Egyptian's Political Upheaval in January 2011," International Journal of Social Sciences & Educational Studies, vol. 3 (2), pp. 94-108, Dec. 2016, ISSN: 2409-1294.

[17] B. Birregah, T. Top, C. Perez, E. Chatelet, N. Matta, M. Lemercier, and H. Snoussi, "Multi-layer crisis mapping: a social media-based approach," Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2012), IEEE, Jun. 2012, pp. 379-384, doi: 10.1109/WETICE.2012.47.

[18] Y. Liu, P. Piyawongwisal, S. Handa, L. Yu, Y. Xu, and A. Samuel, "Going beyond citizen data collection with mapster: a mobile+ cloud real-time citizen science experiment," e-Science Workshops, IEEE Seventh International Conference, (eScienceW 2011), IEEE, Dec. 2011, pp. 1-6, doi: 10.1109/eScienceW.2011.23.

[19] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, and X. Zhang, "Senseplace2: Geotwitter analytics support for situational awareness," IEEE Conference on Visual Analytics Science and Technology (VAST 2011), pp. 181-190. IEEE, 2011, doi: 10.1109/VAST.2011.6102456.

[20] J. Yin, A. Lampert, A. Cameron, R. Robinson, and P. Power, "Using social media to enhance emergency situation awareness," IEEE Intelligent Systems, vol. 27 (6), pp. 52-59, Feb. 2012, doi: 10.1109/MIS.2012.6.

[21] J. Zhu, F. Xiong, D. Piao, L. Liu, and Y. Zhang, "Statistically modeling the effectiveness of disaster information in social media," Global Humanitarian Technology Conference (GHTC 2011), pp. 431-436. IEEE, Oct. 2011, doi: 10.1109/GNTC.2011.48.

[22] B. Al-Ani, G. Mark, J. Chung, and J. Jones, "The Egyptian blogosphere: a counter-narrative of the revolution," Proc. ACM 2012 conference on Computer Supported Cooperative Work, pp. 17-26. ACM, 2012.

[23] S. I. Yuce, N. Agarwal, and R. T. Wigand, "Women's Right to Drive: Spillover of Brokers, Mobilization, and Cyberactivism," International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP 2015), vol. 9021, pp. 232-242, Mar. 2015, doi: 10.1007/978-3-319-16268-3_24.

[24] P. P. Y. Leong, "Political Communication in Malaysia: A Study on the Use of New Media in Politics," JeDEM-eJournal of eDemocracy and Open Government, vol. 7 (1), pp. 46-71, 2015, ISSN: 2075-9517.

[25] R. Goolsby, L. Shanley, and A. Lovell, "On cybersecurity, crowdsourcing, and social cyber-attack," Office of Naval Research, Wilson Center, Commons Lab, policy memo series, vol. 1, 2013. [Online]. Available from: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA580185 2018.05.26

[26] B. Berendt and R. Navigli, "Finding Your Way through Blogspace: Using Semantics for Cross-Domain Blog Analysis," AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006, pp. 1-8.

[27] N. Fearn, "When bots go roque," International Data Group (IDG) Connect. Feb. 17, 2017. [Online]. Available from: http://www.idgconnect.com/blog-abstract/24495/when-bots-rogue 2018.05.26

[28] Google, "Google Trends: Migrant Crisis-A human tragedy," 2015. [Online]. Available from: https://trends.google.com/trends/story/US_cu_iQ-fCFEBAABHhM_en 2018.05.26

[29] Fox News, "Study: Venezuelans lost 19 lbs. on average over past year due to lack of food," Feb. 20, 2017. [Online]. Available from: http://www.foxnews.com/world/2017/02/20/study-venezuelans-lost-19-lb-on-average-over-past-year-due-to-lack-food.html 2018.05.26

[30] L. Melendez, "Encovi 2016: 74% of Venezuelans lost more than 8 kilos of weight last year," Feb. 18, 2017. [Online]. Available from: http://runrun.es/rr-es-plus/297797/encovi-2016-74-de-los-venezolanos-perdio-mas-de-8-kilos-de-peso-el-ano-pasado.html 2018.05.26

[31] M. N. Hussain, K. Bandeli, M. Nooman, S. Al-khateeb, and N. Agarwal, "Analyzing the Voices during European Migrant Crisis in Blogosphere," 2nd International Workshop on Event Analytics using Social Media Data associated with The 11th International AAAI Conference on Web and Social Media (ICWSM 2017), May 15-18, 2017, Montreal, Canada.

[32] Newprosoft, "Web Content Extractor," 2017. [Online]. Available from: http://www.newprosoft.com/web-content-extractor.htm 2018.05.26

[33] N. Agarwal, H. Liu, L. Tang, and S. Y. Philip, "Modeling blogger influence in a community," Social Network Analysis & Mining, vol. 2 (2), pp. 139–162, Jun. 2012, ISSN: 18695450.

[34] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Operator's Manual Linguistic Inquiry and Word Count: LIWC 2007," LIWC.net, Austin, Texas. [Online]. Available from: http://www.depts.ttu.edu/psy/lusi/files/LIWCmanual.pdf 2018.05.26

[35] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Sep 15 2015. [Online]. Available from: http://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_LanguageManual.pdf 2018.05.26

[36] "LIWC - Linguistic Inquiry and Word Count." [Online]. Available from: http://liwc.wpengine.com/ 2018.05.26

[37] N. Casey, "Thousands March in Venezuela to Demand President's Ouster," New York Times, Sep. 1, 2016. [Online]. Available from: https://www.nytimes.com/2016/09/02/world/americas/caracas-venezuela-nicolas-maduro-protests.html?_r=0 2018.05.26

[38] N. Casey, "Venezuelans Ransack Stores as Hunger Grips Nation," New York Times, June 19, 2016. [Online]. Available from: https://www.nytimes.com/2016/06/20/world/americas/venezuelans-ransack-stores-as-hunger-stalks-crumbling-nation.html 2018.05.26

[39] Trading Economics, "Crude Oil 1946-2018," 2018. [Online]. Available: https://tradingeconomics.com/commodity/crude-oil 2018.05.26

[40] Trading Economics, "Venezuelan Bolivar 1988-2018," 2018. [Online]. Available: https://tradingeconomics.com/venezuela/currency 2018.05.26

[41] J. Wattles and P. Gillespie, "Venezuelans face a 6000% hike in gasoline price," CNN Money, February 18, 2016. [Online]. Available: http://money.cnn.com/2016/02/17/news/world/venezuela-gas-price-hike/index.html 2018.05.26

[42] M. Mogollon, "Packs of black-market foot soldiers raid Venezuela markets," LA Times, Apr. 1, 2015. [Online]. Available from: http://www.latimes.com/world/mexico-americas/la-fg-venezuela-shopping-20150401-story.html 2018.05.26

[43] BBC News, "Venezuela economy: Nicolas Maduro declares emergency," Jan. 15, 2016. [Online]. Available from: http://www.bbc.com/news/world-latin-america-35329617 2018.05.26

[44] A. V. Pestano, "Venezuela: 75% of population lost 19 pounds amid crisis," United Press International, February 19, 2017. [Online]. Available: https://www.upi.com/Venezuela-75-of-population-lost-19-pounds-amid-crisis/2441487523377/ 2018.05.26

[45] P. Gillespie, "Thousands of Venezuelans fleeing to the US," CNN Money, May 23, 2017. [Online]. Available: http://money.cnn.com/2017/05/23/news/economy/venezuela-us-asylum-refugees/index.html 2018.05.26

[46] Z. Aleem, "How Venezuela went from a rich democracy to a dictatorship on the brink of collapse," Vox, September 19, 2017. [Online]. Available: https://www.vox.com/world/2017/9/19/16189742/venezuela-maduro-dictator-chavez-collapse 2018.05.26

[47] A. Ulmer and C. Pons, "Venezuela opposition slams 'desperate' Maduro state of emergency," Reuters, May 13, 2016. [Online]. Available: https://www.reuters.com/article/us-venezuela-politics/venezuela-opposition-slams-desperate-maduro-state-of-emergency-idUSKCN0Y501X 2018.05.26

[48] Reuters Staff, "Venezuelan opposition warns foreign contracts need assembly's OK," Reuters, May 26, 2016. [Online]. Available: https://www.reuters.com/article/us-venezuela-politics/venezuelan-opposition-warns-foreign-contracts-need-assemblys-ok-idUSKCN0YH2L9 2018.05.26

[49] BBC News: Latin America, "Venezuela: Eleven shot dead in Trujillo state," BBC News, May 30, 2016. [Online]. Available: http://www.bbc.com/news/world-latin-america-36410563 2018.05.26

[50] NPR: All Things Considered, "Running Out of Food, Medicine and Patience in Venezuela," June 26, 2016. [Online]. Available: https://www.npr.org/2016/06/26/483624346/running-out-of-food-medicine-and-patience-in-venezuela 2018.05.26

[51] C. Morello, "OAS head calls for recall of Maduro to restore democracy in sprialing Venezuela," The Washington Post, June 23, 2016. [Online]. Available: https://www.washingtonpost.com/world/national-security/oas-head-calls-for-recall-of-maduro-to-restore-democracy-in-spiraling-venezuela/2016/06/23/1568edcc-3974-11e6-a254-2b336e293a3c_story.html 2018.05.26

[52] N. Casey and P. Torres, "Venezuela Drifts Into New Territory: Hunger, Blackouts and Government Shutdown," New York Times, May 28, 2016. [Online]. Available: https://www.nytimes.com/2016/05/28/world/americas/venezuela-economic-government-collapse.html 2018.05.26

[53] D. Duquenal, "Venezuela News And Views: A blog about life under, and resisting, a dictatorship". [Online]. Available from: http://daniel-venezuela.blogspot.com 2018.05.26

A Method of Misbehavior Detection with Mutual Vehicle Position Monitoring

Shuntaro Azuma
Computer and Information Science
Graduate School of Science and Engineering
Doshisha University
Kyoto, Japan
email:syuntaro.azuma@nislab.doshisha.ac.jp

Manabu Tsukada
Graduate School of Information Science and Technology
Tokyo University
Tokyo, Japan
email:tsukada@hongo.wide.ad.jp

Kenya Sato
Computer and Information Science
Graduate School of Science and Engineering
Doshisha University
Kyoto, Japan
email:ksato@mail.doshisha.ac.jp

***Abstract*—Due to the development of vehicle-to-vehicle (V2V) communication, safe driving support such as collision prevention and adaptive cruise control has been achieved. In addition, vehicle-to-infrastructure (V2I) communication as well as communication with a cloud server using mobile lines (vehicle-to-cloud communication) have been developed in recent years. These communications are altogether called vehicle-to-everything (V2X) communication. Through V2X communication, a vehicle's peripheral information can be shared with other vehicles on a cloud server. However, the problem of masquerade attacks on the cloud must be addressed. By faking vehicle information on a cloud server, an adversary may deliberately cause traffic congestion and/or accidents. In this research, we proposed a method that detects misbehavior (masquerade data) from aggregated data on a cloud server using V2X communication by utilizing the surrounding vehicle information. We also analyzed possible threats and requirements for data that are sent to cloud servers, and evaluate the proposed method's implementation. Using the proposed method, we detected 93% of the masquerade data, improved the detection rate by 100% by increasing the threshold value of the proposed method, and enhanced the effect of guaranteeing the data's reliability. Furthermore, we evaluated the false positives of the proposed method and its execution processing time, examining the method's feasibility.**

Keywords–vehicle security; V2X communication; detecting masqueraded data.

I. INTRODUCTION

This paper is based on "A Method of Detecting Camouflage Data with Mutual Vehicle Position Monitoring" published in VEHICULAR 2017[1]. This paper consists of 6 sections. We describe the background of this research in section II. In section III, we compare existing research and analyze the threat of vehicle spoofing to clarify the novelty and goal of this research. We show our proposed method in section IV, and we describe its evaluation and consideration in section V. Considering the evaluation, we offer insight to our future work in section VI, and summarize this research in the final section VII.

II. BACKGROUND

In recent years, research on autonomous driving and vehicle-to-vehicle (V2V) communication are being conducted in the Intelligent Transport Systems (ITS) field. In addition to providing V2V communication using the Vehicular Ad hoc Network (VANET), vehicles can engage in vehicle-to-infrastructure (V2I) communication with roadside aircraftst and vehicle-to-pedestrian (V2P) communication with tablets owned by pedestrians. Vehicles can also do vehicle-to-cloud (V2C) communication with cloud servers using mobile lines. These communication methods listed above are generally referred to as vehicle-to-everything (V2X) communication. When vehicles perform V2X communication, cloud servers can collect various kinds of information, and we can create a Local Dynamic Map (LDM) [2] for cooperative driving from the collective management of road and vehicle information. This type of communication sometimes is referred as probe information systems [3] or floating car data (FCD) [4]. In addition, various systems and services can be provided, which includes simplification of management tasks such as summarizing operation results, analyzing operation trends, summing up tasks, and simplifying the input of daily reports.

On the other hand, in a system using a cloud server, masquerade data transfer to a cloud influences a system. Attacks against safe driving support services using a cloud pose a threat because the intentional transfer of masquerade data to a cloud are on the rise. Attackers can block roads or cause traffic congestion by sending fake traffic accident information to a cloud server. Various masquerade techniques of vehicle disguise have been identified, such as faking driving and position information as well as a vehicle 's condition. In this research, we focus on masqueraded position information among all of the data received by a cloud server from vehicles, and we attempt to detect them by mutually monitoring the position information of vehicles using V2X communication.

III. RELATED WORK

There are previous work researching the detection of malicious vehicles in V2X communication [5] [6], but in reality, the definition of a malicious vehicle is ambiguous. In this section, we analyze attacks on vehicle communication and clarify what kind of malicious vehicles to be solved in this research.

TABLE I. THREATS ANALYSIS ABOUT TRANSMISSION DATA

THREAT		REQUIREMENT	COUNTERMEASURE
Eavesdropping		Confidentiality	Encryption
Falsification		Completeness	Encryption
Spoofing	Vehicle impersonation	Node reliability	PKI
	Data masquerade	Date reliability	**Target of this research**

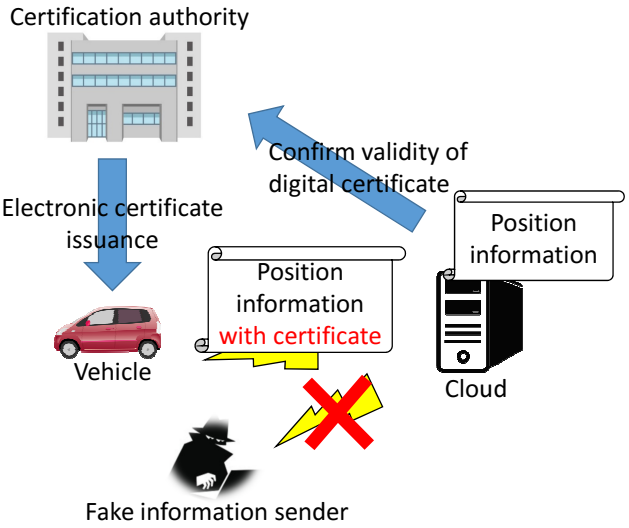


Figure 1. PKI to adapt to vehicles

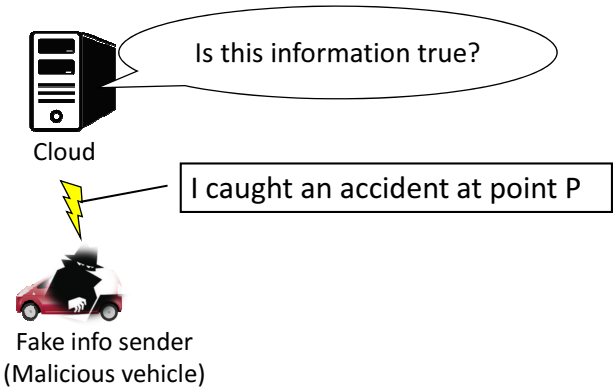


Figure 2. Problem of settling by this research

A. Threat Analysis of Transmission Data

Table I shows the threat analysis of data transmitted to a cloud server. These threats include eavesdropping attacks, falsifications, and spoofing. Spoofing attacks are divided into vehicle impersonation and data masquerade. Vehicle impersonation means that attackers pretend to be other vehicles. For example, even though one vehicle does not have any trouble, an attacker pretends to be another vehicle and then calls the police lying that it had an accident. An example of data masquerade is when a vehicle's own position information or status is masked.

Security requirements regarding these threats include confidentiality, completeness, node reliability, and data reliability. To supply confidentiality and completeness, data encryption is proposed and can be done by a secret key or an ID base cipher. Node reliability identifies vehicles that are pretending to be other vehicles. The Public Key Infrastructure (PKI) method, which is adapted by the vehicles, is one good resolution because certificates guarantee vehicles. Data reliability prevents attackers from masquerading data. However, this is not effective for all spoofing acts.

B. Difference Between Node and Data Reliability

Node reliability means that a cloud server trusts a particular vehicle and believes that it is not pretending to be a different vehicle. The previous section showed that the PKI method can be adapted by vehicles to resolve the problem. A cloud may be able to verify the electronic certification and confirm the transmitter's information by the mechanism shown in Figure 1.

On the other hand, this research focuses on data masquerade, as described in Figure 2. Since data encryption and PKI do not confirm whether the received data are masqueraded, data masquerade is inherently different from node reliability which can be resolved by these methods. We will propose a method that can handle such example, which guarantees the reliability of the data.

IV. PROPOSAL

In this section, we will propose a method to detect misbehavior from data transmitted to a cloud.

A. Outline

Vehicles can use V2X communication. When they send their position information to a cloud server, they also send other information in addition to their position. In this research, a cloud server detects masqueraded data from transmitted data by using the relay base station information in V2C communication and peripheral vehicles in V2V communication. We will explain separately them to simplify our proposing method.

B. Presuppositions

1) A safe channel has been secured by relationships of mutual trust among all vehicles and cloud servers
2) Vehicles and cloud servers have been mutually certified beforehand.
3) Relationships between cloud servers and base stations have been built.

C. Definition of Terminology in Proposed Method

- Vehicle ID

This ID is used by vehicles in V2V communication, and this is a different public ID for each vehicle.

- V2C Vehicle ID

This ID is used for a unique key in V2C communication. This secret ID is not available to others. V2C Vehicle ID and Vehicle ID is uniquely related.

- Via Base Station (BS) ID

This ID is used in V2C communication, and this is a different ID for each base station.

- Peripheral Vehicle (PV) ID

This ID is a received vehicle ID from other vehicles in V2V communication.

D. Use of Base Station's Information in V2C Communication

When sending position information in V2C communication, vehicles attach V2CVehicleID to their position information, and send it to a cloud. A relay base station on the V2C communication attaches its own ViaBSID to information which was sent from vehicles and encapsulates it. V2CVehicleIDs of all vehicles are registered in cloud servers, and cloud servers can be known from which vehicles inquiry when they confirm these IDs. A possible communication range covered by

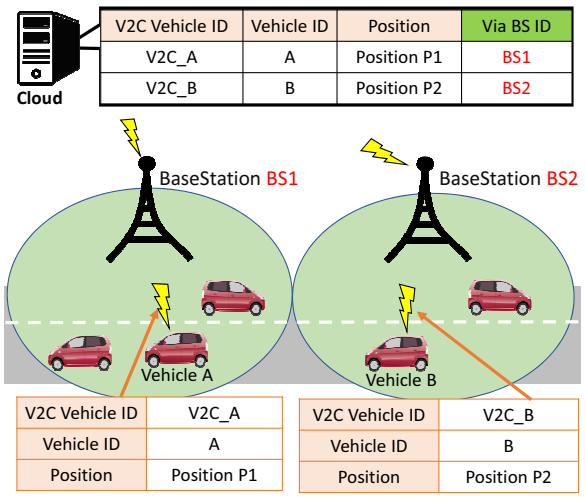


Figure 3. Use example of base station's information in V2C communication

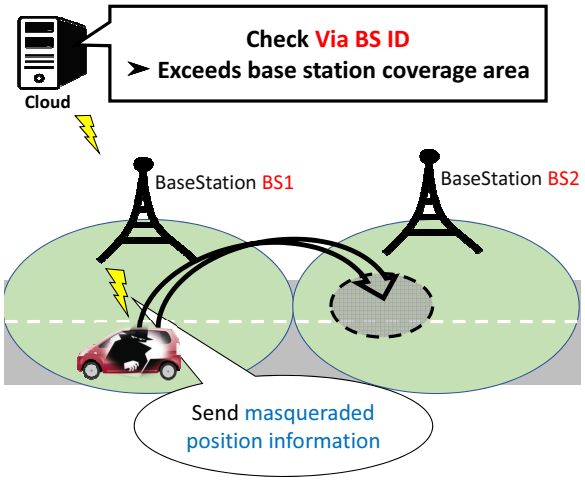


Figure 4. Advantage of using base station's information

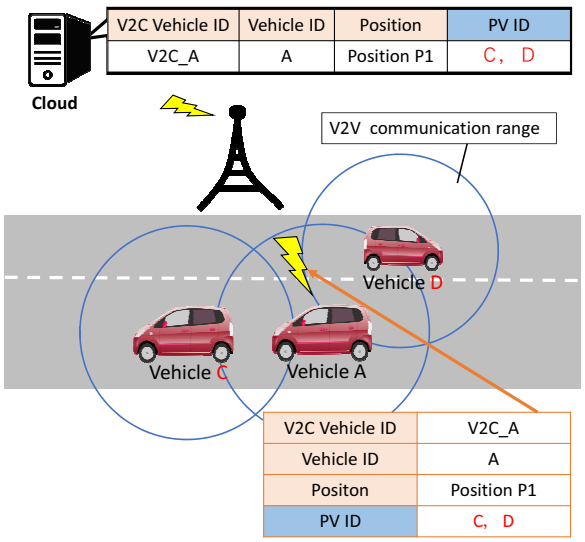


Figure 5. Use example of peripheral vehicle information in V2V communication

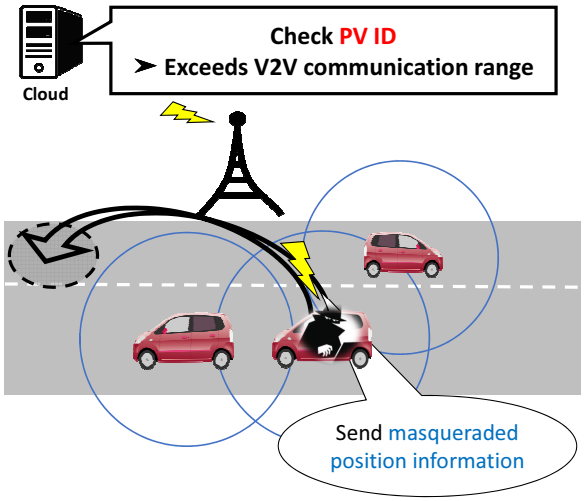


Figure 6. Advantage of using peripheral vehicle information

base station's area and ViaBSIDs are also registered in cloud servers.

Figure 3 indicates an example of base station's information in V2C communication. Vehicles possess own V2CVehicleID; base stations also possess own ViaBSIDs. V2CVehicleIDs shall be V2C_A or V2C_B, and ViaBSIDs shall be BS1 or BS2 to explain simply. When vehicles perform V2C communication, a cloud can obtain not only VehicleID or vehicle's position information but also ViaBSID and V2CVehicleID.

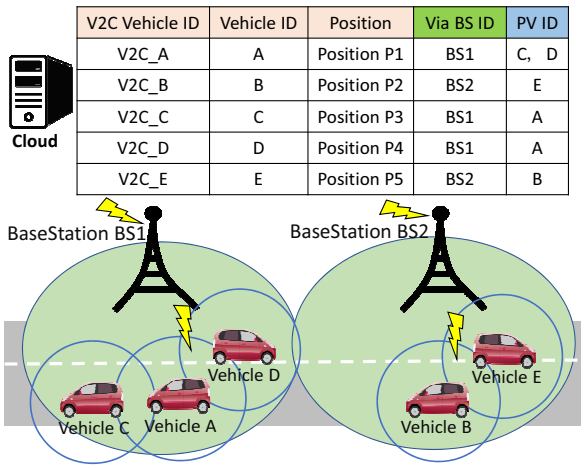
Figure 4 shows a countermeasure example of position data masquerade. We can detect masqueraded position information toward another base station using relay base station's information in V2C communication.

E. Use of Peripheral Vehicle's Information in V2V Communication

Vehicles exchange VehicleIDs with nearby vehicles in V2V communication. We define that peripheral vehicles are traveling vehicles within V2V communication coverage area, and PVID is received vehicleID from a peripheral vehicle. In our proposed method, only VehicleID is exchanged in V2V communication.

When vehicles send their position information to a cloud, these information include V2CVehicleID, VehicleID, and received PVIDs in V2V communication. PVIDs show a guarantee that nearby vehicles exist in V2V communication coverage area. Figure 5 shows an example of peripheral vehicle's information in V2V communication. Vehicle A communicates with the vehicle C and D which are traveling in V2V communication coverage area, and acquires those vehicle IDs. Vehicle A handles acquired VehicleIDs as PVIDs, and a cloud use PVIDs to check Vehicle A's position information with peripheral Vehicles C and D.

Figure 6 shows a countermeasure example of position data masquerade. We assume that a malicious vehicle masquerades its own position information. A cloud confirms PVIDs sent from a vehicle and compares received position information with peripheral vehicle's positions which are relevant to PVIDs. When a cloud finds that transmitted position information is outside V2V communication coverage with peripheral vehicles, the cloud determines that the received position information has been masqueraded. However when this information



V2C Vehicle ID	Vehicle ID	Position	Via BS ID	PV ID
V2C_A	A	Position P1	BS1	C, D
V2C_B	B	Position P2	BS2	E
V2C_C	C	Position P3	BS1	A
V2C_D	D	Position P4	BS1	A
V2C_E	E	Position P5	BS2	B

Figure 7. Use example of peripheral vehicle information in V2X communication

does not exceed the coverage area, the cloud trusts the received position information. Vehicles acquire peripheral vehicle information in V2V communication and mutually monitor them. This helps cloud servers detect masqueraded data.

F. Detection Method of Misbehavior With V2X Communication

Our proposed method is a combination of two described above by using V2X communication (Figure 7). A cloud receives not only position information or VehicleID but also peripheral vehicle's and relay base station's information. Masqueraded data can be detected through these information, as described in Figure 8.

V2CVehicleID is used in the first step on Figure 8. Cloud servers confirm whether received data is sent from vehicles or not. Second, cloud servers compare ViaBSID with received position information to confirm whether a sending vehicle exists in relay base station's coverage area. When receives position information exceeds this area, we assume that it can't be consistent and that received information was regarded as masqueraded data. This step helps detect data masquerade toward other base station's coverage area. At the third or fourth step, cloud servers detect masqueraded data by using PVIDs. Third, cloud servers search vehicles corresponding to sending vehicle's PVID, and firth, they compare received position information with peripheral vehicle's position corresponding to PVID. If the distance between two vehicle's position exceeds V2V communication coverage, we assume that it can't be consistent and that received information was regarded as masqueraded data. This operation is performed a predetermined number of times. In proposed method, a predetermined number of times means the number of PVIDs which is necessary for cloud servers to trust. This is a so-called threshold value. By setting threshold, we can assure more reliable data.

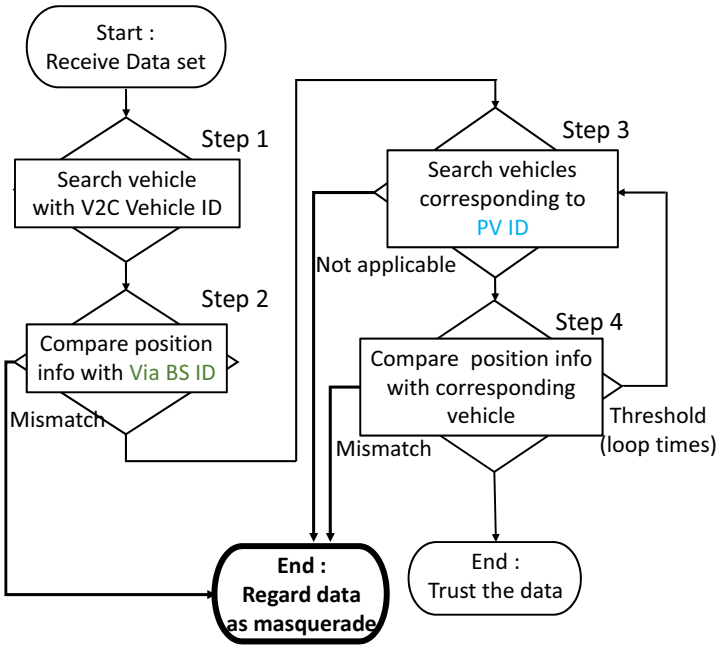


Figure 8. Misbehavior detection procedure to a vehicle send data

V. EVALUATION AND CONSIDERATION

To evaluate the usefulness of our proposed misbehavior detection, we will calculate the evaluation. Then we will consider the practicality of our proposal from the obtained evaluation.

A. Simulator

In this paper we use Scenargie [7] as a simulator to evaluate the performance of our proposed method. Scenargie is a network simulator developed by Space-Time Engineering (STE). By combining expansion modules such as LTE, V2V communication and multi-agent, we can construct a realistic simulation. In addition, since communication systems and evaluation scenarios are becoming more complicated, this ingenious simulation has greatly reduced the effort required to create scenarios.

B. Evaluation Model

For a evaluation environment, we use one square kilometer Manhattan model and use simulation parameters shown in Table II. We set the number of vehicles to 158 [cars] and the range to 1 [km^2] because the average car density in Japan is 158 [cars/km^2]. ITU-R P.1411 model is a radio wave propagation scheme that considers road map information, and radio waves are attenuated based on the shape of the road, so we compared with a two-ray model which includes direct waves and reflected waves from the ground, this model is close to reality. Figure 9 shows one scene when the simulator is active.

TABLE II. SIMULATION PARAMETER

Simulator	Scenargie2.0	
Vehicle number	158 [cars] (five of the send masquerade positions.)	
Area	1000 [m] × 1000 [m]	
Communication mode	ARIB STD T109	LTE
Use frequency band	700 [Mhz]	2.5 [GHz]
Communication interval	100 [ms]	1.0 [s]
Radio spread model	ITU-R P.1411	LTE-Macro
Base station ground clearance	1.5 [m]	

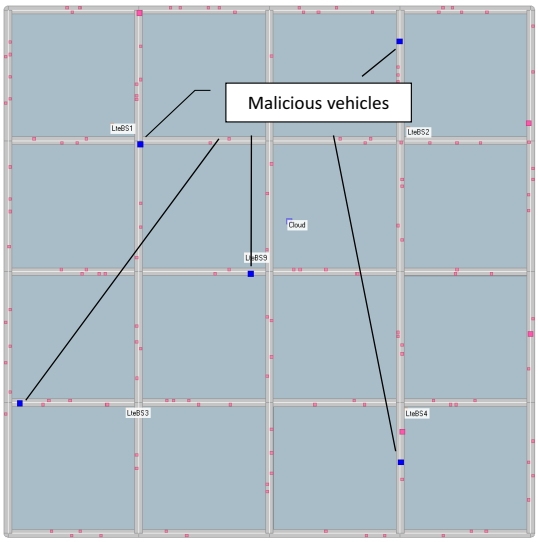


Figure 9. One scene when the simulator is running

C. Evaluation of Misbehavior Detection

Figure 10 shows per-threshold detection rates of masqueraded data from data aggregated in a cloud server. Masqueraded data transmitted to a cloud could be detected at 100% by increasing the proposed method's threshold. However, when the threshold was low, we could not completely detect all of masqueraded data. The reason is shown in Figure 11 and 12. Figure 11 shows an example of misbehavior in V2V communication coverage with peripheral vehicles. A malicious vehicle masquerades its own position (position 1) to position 2 in V2V communication coverage. In this case, since peripheral vehicles guarantee masqueraded information from a malicious vehicle, a misbehavior becomes possible. Figure 12 shows a collusion between malicious vehicles. Since they mutually guarantee masqueraded position information, cloud server trusts these information. However, these problems can be addressed by increasing the prescribed number of times (threshold values) in Figure 8. By increasing the threshold values, we can create the situation shown in Figure 13, and it is possible to limit masquerading by malicious vehicles.

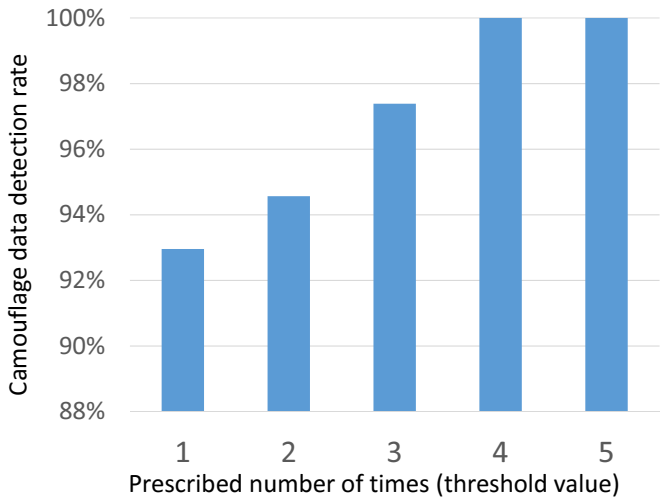


Figure 10. Misbehavior detection rates in a cloud received data

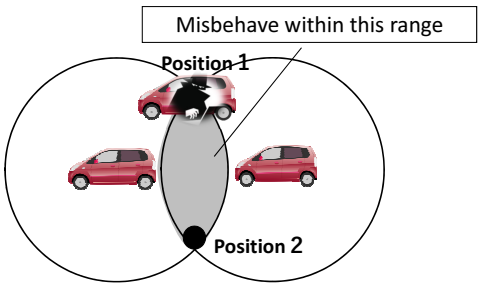


Figure 11. Misbehavior in V2V communication coverage with peripheral vehicles

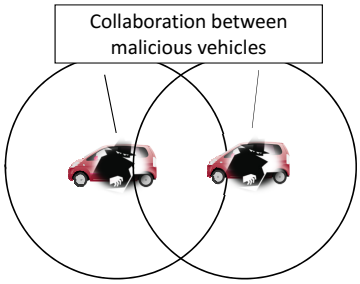


Figure 12. Collusion between malicious vehicles

D. Evaluation of Misdetection Rate

Figure 14 shows false detection rates (false positives) of our proposed method, which is based on the average vehicle density in Japan. The method's threshold is the number of PVIDs which is necessary for cloud servers to trust. In the previous section, we found that an increase in the threshold improves the detection rates of masqueraded data. Here, we will calculate the false positive detection rates (false positive), regarding whether cloud servers trust information on vehicles that are not misbehaving.

In the simulation environment shown in Table V, Figure 14 shows that the false positives when all 158 cars are not misbehaving. By increasing the threshold value, false positive rates improved. Increasing the threshold value under Japanese average vehicle density, cloud servers erroneously detects normal communication as abnormal.

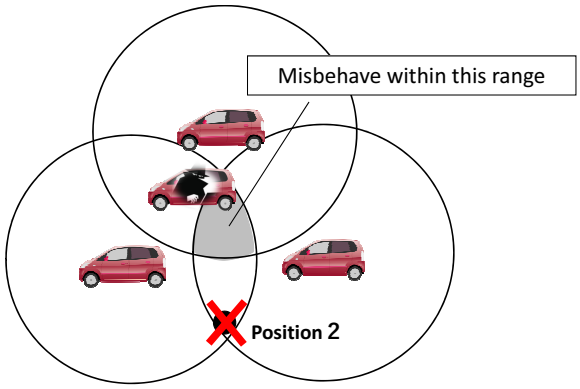


Figure 13. Restriction on misbehavior accompanying an increase in information on peripheral vehicles

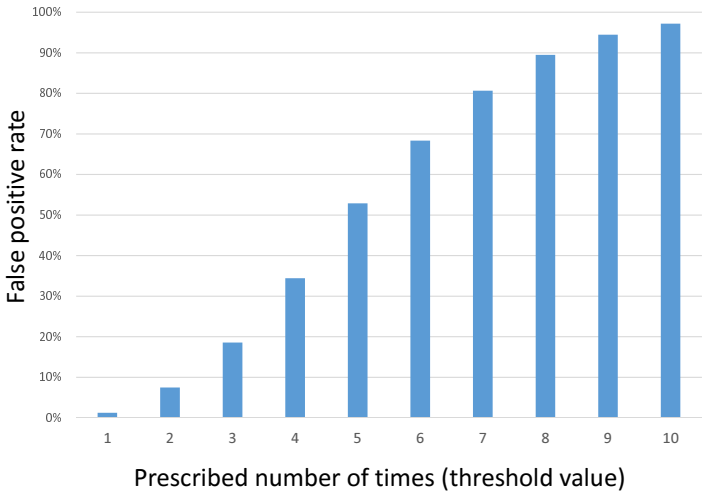


Figure 14. False positives by threshold values under Japanese average vehicle density (158[cars/km^2]) environment

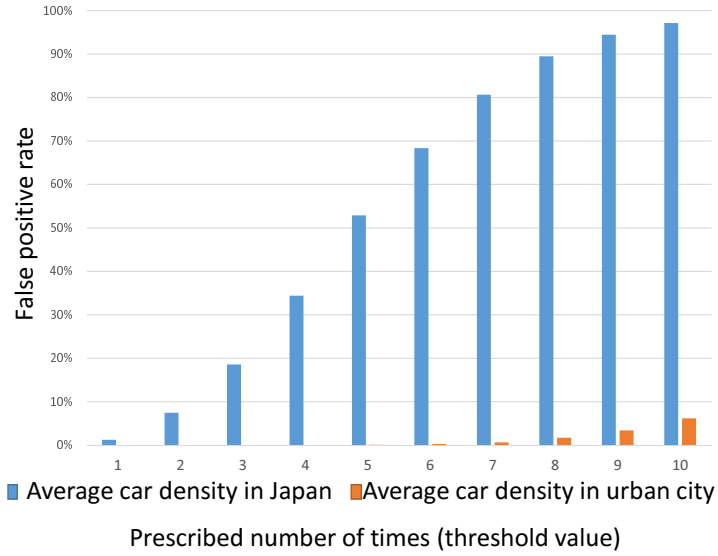


Figure 15. False positive comparison with urban area's average vehicle density (1128 [cars/km^2]) environment

Then, the false positive rates under the average vehicle density environment in urban city (Osaka), which has the highest average car density in Japan, are shown in Figure 15. In a high vehicle density area, since vehicles can acquire a lot of peripheral vehicle information in V2V communication, even if the threshold is increased, an increase of the false detection rate can be suppressed. Therefore, we found that our proposed method is more effective in areas with high vehicle density than lower density. Actually the influence of masquerading vehicle information is great in areas with high vehicle density. Malicious act such as faking driving and position information may cause the large accident in higher density than lower. Our proposed method can guarantee transmitted information from vehicle to cloud servers, and it works better in areas with a larger number of peripheral vehicles than in areas with fewer peripheral vehicles. Our proposed method is useful in traffic congestion zones where data masquerade has a huge impact. Table III and IV show precision, recall, and F-measure in our proposed method. Even looking at this table, we can say the same above.

TABLE III. F-MEASURE UNDER 158[cars/km^2] ENVIRONMENT

Threshold	1	2	3	4	5	6	7	8	9	10
Precision	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Recall	0.99	0.93	0.81	0.66	0.47	0.32	0.19	0.11	0.056	0.029
F-measure	0.99	0.96	0.90	0.79	0.64	0.48	0.32	0.19	0.11	0.056

TABLE IV. F-MEASURE UNDER 1128[cars/km^2] ENVIRONMENT

Threshold	1	2	3	4	5	6	7	8	9	10
Precision	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Recall	1.0	1.0	1.0	1.0	1.0	1.0	0.99	0.98	0.97	0.94
F-measure	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99	0.98	0.97

E. Comparison of ARIB STD-T109 and IEEE 802.11P

ARIB STD-T109 is a V2V communication's standard used in Japan. In a previous evaluation we used this standard based on V2V communication in the simulator. IEEE802.11p is a Dedicated Short Range Communication (DSRC) standard for wireless vehicular networks in the United States and Europe [8] [9]. IEEE802.11p, which is the standard for transports and the network layers, is standardized as the part of IEEE 1609 [10] family and defines the architecture and security physical layer access etc. for DSRC. The main differences between ARIB STD-T109 and WAVE are the frequency band and the number of channels. We will confirm that what kind of difference using ARIB STD-T109 will make with IEEE802.11p in our proposed method. The simulation environment using IEEE802.11p is shown in Table V.

F. Differences Among Misbehavior Detection Rates

Figure 16 shows the per-threshold detection rates of masqueraded data from data aggregated on a cloud server. Unlike Figure 10, it is the result of using IEEE802.11p. Compared to using T109, the detection rates improve with IEEE802.11p. When the threshold value is 3, the detection rate is 100%. To determine this difference, our proposed method must determine how much vehicles communicate with peripheral vehicles. If there are many communication targets around a vehicle, cloud servers can trust vehicles and exclude malicious vehicles. T109 has a lower frequency than IEEE802.11p. However since IEEE802.11p has strong propagation strength, vehicles can be communicated to more peripheral vehicles, leading to better results.

G. Differences Among Misdetection Rates

Figure 17 shows the false detection rates (false positives) of the proposed method based on the average car density in Japan when we used IEEE802.11p. It shows considerably better results than Figure 14. As stated above, this is related

TABLE V. SIMULATION PARAMETER

Simulator	Scenargie2.0	
Vehicle number	158 [cars] (five of the send masquerade positions.)	
Area	1000 [m] × 1000 [m]	
Communication mode	IEEE802.11p	LTE
Use frequency band	5.9 [Ghz]	2.5 [GHz]
Communication interval	100 [ms]	1.0 [s]
Radio spread model	ITU-R P.1411	LTE-Macro
Base station ground clearance	1.5 [m]	

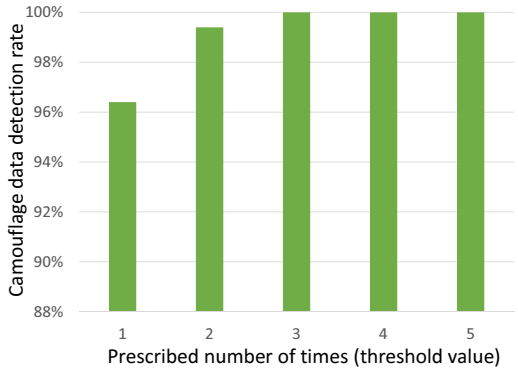


Figure 16. Misbehavior detection rates in a cloud received data when using IEEE802.11p

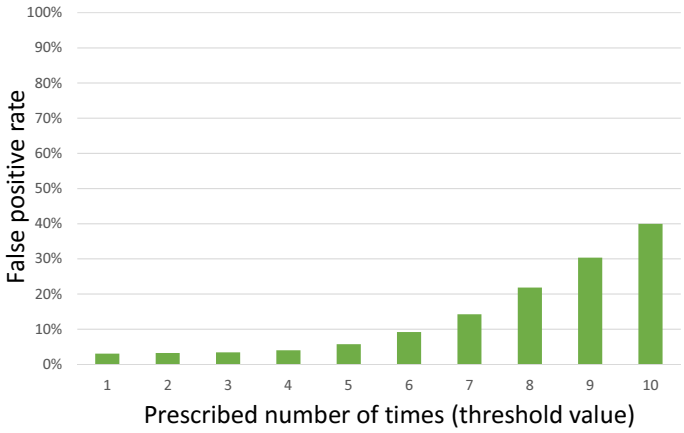


Figure 17. False positives by threshold value under Japanese average vehicle density (158[cars/km^2]) environment when using IEEE802.11p

to the fact that vehicles using IEEE802.11p can communicate with peripheral vehicles, and our proposed method works well.

Figure 18 indicates the false positives under an average vehicle density environment in an urban city (Osaka), which has the highest average car density in Japan. The graph shows almost the same result as Figure 17. When T109 was used, the

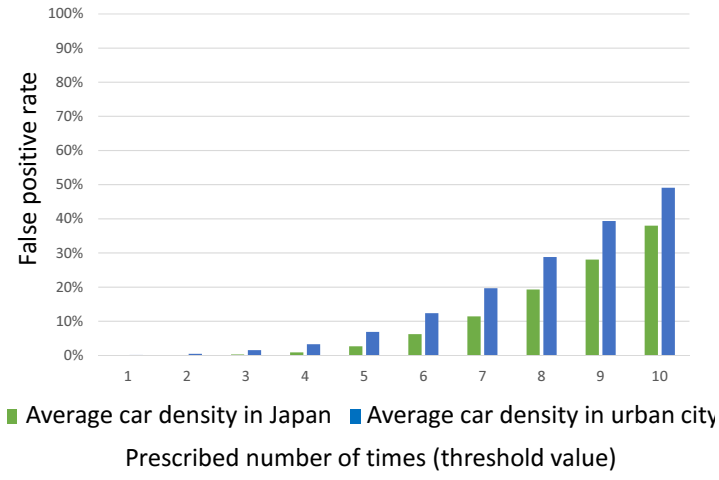


Figure 18. False positive comparison with urban area's average vehicle density (1128 [cars/km^2]) environment when using IEEE802.11p

results were significantly different depending on the car density because the number of peripheral vehicles (that can communicate) increased due to the greater vehicle density. However, since there is only slight difference when IEEE802.11p is used, we believe that it was probably communicating with vehicles that can already communicate under an average Japanese car density environment. This means that the results did not change even under an average Japanese vehicle density environment. When using T109, a suitable threshold must be set for various vehicle densities, but since there is no change in the decreasing positives when using IEEE802.11p, a certain threshold may be satisfied under any vehicle density environment.

H. Evaluation When Increasing Malicious Vehicles

We believe that we can improve our proposed system by increasing the number of malicious vehicles and setting the threshold value to 5. When the number of malicious vehicles is increased, the false positives are shown in Figure 19, and 20. In both T109 and IEEE802.11p, the amount of the false positives did not change even when the malicious vehicles are increased because five regular vehicles were included in the communication targets (peripheral vehicles). Therefore, even ian increase in the number of malicious vehicles increases did not affect the false positives of regular vehicles. Unfortunately, this result is not good.

In our proposed method, all communicating vehicles are regarded as peripheral vehicles that can guarantee their own position information. Even if there are malicious vehicles in the position information, no problem occurs as long as a threshold number of regular vehicles exists. For example, we assume that a certain vehicle communicates with 20 peripheral vehicles. When 15 out of 20 units are malicious, should this vehicle be trusted by the cloud? Although the present system is supposed to trusted ism, more than half of the surrounding vehicles are probably malicious. In other words, even though a malicious vehicle may be correctly identified, we can accept a majority of the opinions. This influence increases the number of vehicles that are identified as malicious.

I. Measurement of Processing Time in Our Proposal

In the evaluation environment shown in Table VI, the processing time necessary for the detection of masqueraded data is evaluated by Table VII and Table VIII, based on the flowchart of Figure 8. It shows the processing time for one vehicle. By using BSIDs, masqueraded data can be detected at the beginning of processing by the proposed method, and

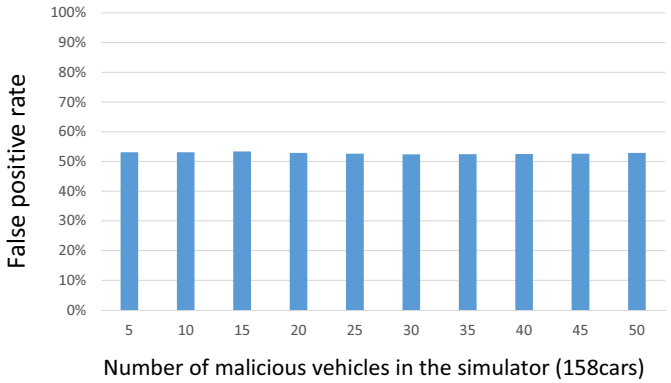


Figure 19. False positives by number of malicious vehicles under Japanese average vehicle density (158[cars/km^2]) environment when using T109

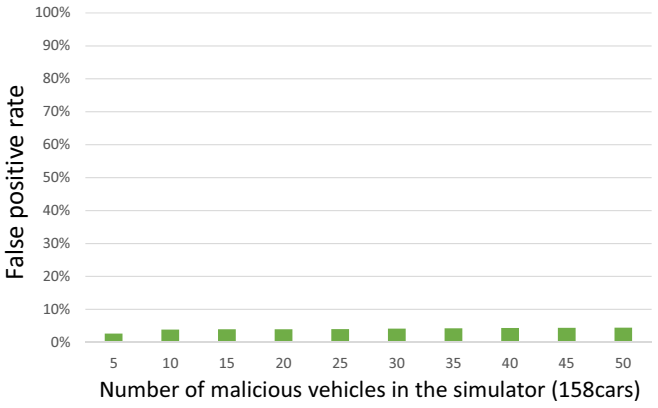


Figure 20. False positives by number of malicious vehicles under Japanese average vehicle density (158[cars/km^2]) environment when using IEEE802.11p

TABLE VI. ENVIRONMENT IN THE PROCESSING TIME MEASUREMENT

OS	macOS Sierra
Processor	1.6GHz Intel corei5
Memory	8GB 1600MHzDDR3
Script	Python
Data base	MySQL, posgreSQL

the processing time becomes relatively fast. In the detection method using PVIDs, the processing time is different for each threshold. By increasing the threshold value, the detection procedure of masqueraded data by PVID is repeated. Even during the repetition, since the processing time changes depending on whether comparative data can be found relatively early or in the final stage, a range was set for the processing time up to Step 4. A case where no masqueraded data is not detected is defined as normal termination, and the upper limit of the processing time at that threshold is indicated. As the threshold of our proposed method increases, the processing time required for normal termination increases. Furthermore,

TABLE VII. PROCESSING TIME OF UNJUST MEASURE TO A VEHICLE OF SEND DATA ON MYSQL

Threshold	Detected in step2	Detected in Figure 8's step4	Usual end
1	0.10[ms]	0.31[ms]	0.31[ms]
2	0.10	[0.31,0.53]	0.53
3	0.10	[0.31,0.76]	0.76
4	0.10	[0.31,0.96]	0.96
5	0.10	[0.31,1.2]	1.2
6	0.10	[0.31,1.4]	1.4
7	0.10	[0.31,1.6]	1.6
8	0.10	[0.31,1.8]	1.8
9	0.10	[0.31,2.0]	2.0

TABLE VIII. PROCESSING TIME OF UNJUST MEASURE TO A VEHICLE OF SEND DATA ON POSGRESQL

Threshold	Detected in step2	Detected in Figure 8's step4	Usual end
1	0.15[ms]	0.52[ms]	0.52[ms]
2	0.15	[0.52,0.88]	0.88
3	0.15	[0.52,1.2]	1.2
4	0.15	[0.52,1.6]	1.6
5	0.15	[0.52,2.0]	2.0
6	0.15	[0.52,2.3]	2.3
7	0.15	[0.52,2.7]	2.7
8	0.15	[0.52,3.0]	3.0
9	0.15	[0.52,3.4]	3.4

we confirmed that the processing time varies depending on the type of database. As a result of calculating the evaluation under the same condition, we found that MySQL is faster in processing time than posgreSQL.

J. Overall Processing Time

In previous subsection V-I, we showed the processing time in our proposed method, especially the processing time in Figure 8's step 4. If we operate our system in reality, the overall processing time will be as follows.

$$T_{all} = T_1 + T_2 + T_3 \quad (1)$$

T_{all} : Overall processing time
T_1 : Delay time of V2C communication
T_2 : Database access time
T_3 : Processing time in my proposal

Table VII or VIII which is calculated in the previous subsection corresponds to T_2, T_3. When we consider the total processing time, we need to consider the delay time of V2C communication. We must determine the threshold values based on the V2C communication delay and the allowable range of the delay times of safe driving support systems.

VI. FUTURE WORK

In this section, considering the evaluation result of the previous section we describe what kind of research we will do in the future.

A. Determination of the Appropriate Threshold

The threshold value we set is the number of loops in Figure 8, that is, the number of peripheral vehicles required for a cloud to trust. If we do not decide the appropriate value, our proposed system will not be realistic. Considering the false positive problem, we devise two methods of determining thresholds.

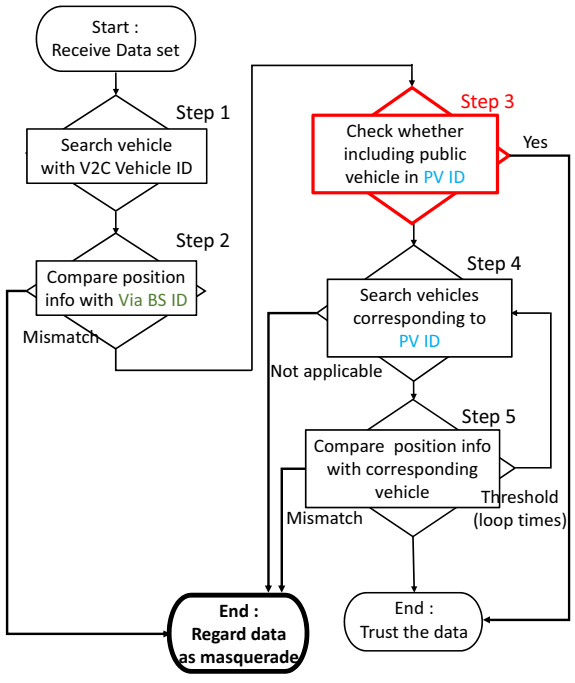


Figure 21. New flowchart introducing vehicle weighting

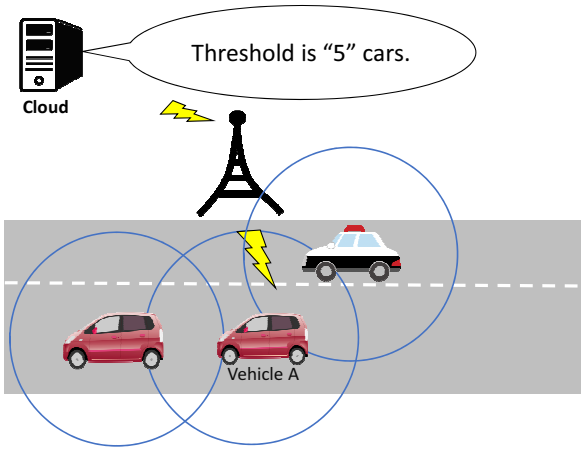


Figure 22. Example when introducing vehicle weighting

1) We weight the public vehicles and trust more on the cloud even for vehicles below the threshold.
2) We dynamicaly determination of threshold value with consideration of vehicle density.

1)This good result (Figure 15, 18) only applies in the urban area. We need to take another measure under the environment of Japanese average car density. And also we must consider the lesser nighttime of car streets and the lower density environment. Figure 21 shows our new countermeasure to the false positive. We give weight to public vehicles such as police vehicles and buses than normal vehicles. Even if the vehicle communicating with the public vehicle (that is, the vehicle including the public vehicle in peripheral vehicle information) does not exceed the threshold value, this one is trusted by a cloud. We consider the environment such a Figure 22. In the case the threshold required for the cloud to trust is 5. Vehicle A has only three peripheral vehicles. But because there are a police vehicle in them, a cloud trusts vehicle A. We think that this method will reduce the false positive if public vehicles are running even in low vehicle density areas.

2)Based on the results (Figure 14,15), we calculate vehicle density for each base station and change the threshold value for each base station. Figure 23 shows the overall picture.

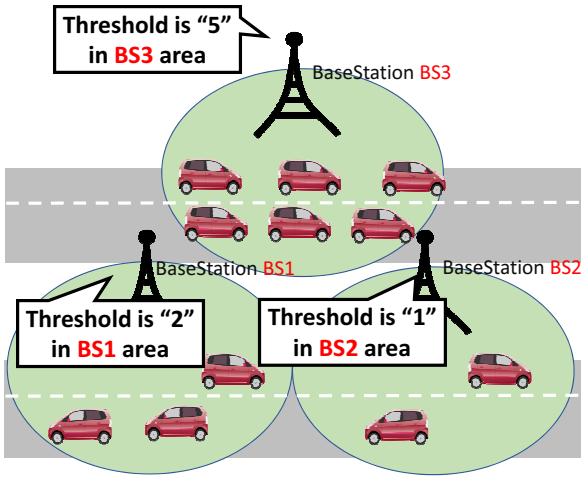


Figure 23. Examples when introducing dynamic threshold determination

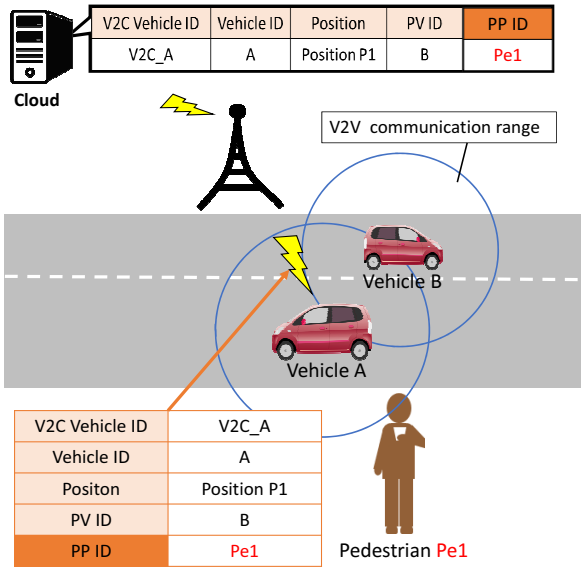


Figure 24. Use example of peripheral pedestrian information in V2P communication

Since vehicle information is transmitted to the cloud via a base station, we think that we can calculate the car density inside the base station.

B. Increase in Communication Target

In this research vehicles only communicated with a cloud and other vehicles. There is vehicle-to-pedestrian (V2P) or vehicle-to-device (V2D) communication in V2X communication. If pedestrians can access a LDM with their smartphones, vehicles can mutually monitor the position information by communicating with the device of the pedestrian. Research that pedestrians can easily access LDM exists [11], so we can realize it by using just information exchanged with V2P communication. Figure 24 indicates the example of peripheral pedestrian information in V2P communication. Vehicle A communicating with pedestrian Pe1, vehicle A gets the peripheral pedestrian ID (PPID). Then vehicle A send these information to a cloud. A cloud can judge whether to trust also from information other than peripheral vehicle information. We think that we can further improve our proposed system.

However there is a problem that the amount of information transmitted by vehicles to a cloud increases, and another problem is how to set a threshold value. People can possess multiple smartphones simultaneously, so the importance of PPID is lower than PVID. We also need to consider those who use malicious behavior with smartphones.

VII. CONCLUSION

In the Intelligent Transport Systems (ITS), using cloud servers is inevitable. For providing a safe driving support service using cloud servers, masquerading vehicle information and spoofing a vehicle are threatening. In this research, we used V2X communication, obtained information from various objects, and described measures against data masquerade. We proposed a method that detects masqueraded data from information transmitted by vehicles to cloud servers. By using information of relay base stations in V2C communication and peripheral vehicle’s information in V2V communication,

and our measures are taken against masquerading vehicle information. By increasing our proposed method's threshold, the detection rates of masqueraded data were improved and vehicle information was made more reliable. Our proposed method can be adapted to depopulated regions by changing the amount of data of peripheral vehicle information required as the detection rates improve based on car densities. In overcrowded vehicle areas, we confirmed that our proposed method works most effectively because there were many peripheral vehicles satisfying the threshold. Further we confirmed that the proposed method works well without problems even under different propagation environment schemes (T109 or IEEE802.11p). False positive problems are our future tasks and we are also considering processing time improved. For future research I would like to conduct a demonstration experiment that also cooperated with LDM.

ACKNOWLEDGMENT

A part of this work was supported by KAKENHI (JP 16H02814) and MEXT program for the strategic research foundation at private universities.

REFERENCES

[1] Shuntaro Azuma, Manabu Tsuakda, and Kenya Sato,“ A Method of Detecting Camouflage Data with Mutual Vehicle Position Monitoring, ” VEHICULAR2017, July 2017.

[2] ETSI,“ Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Local Dynamic Map (LDM), ”EN 302 895 V1.1.1, September 2014.

[3] International Standardization Organization,“ Vehicle probe data for wide area communications, ”ISO 22837, 2009

[4] Ralf-Peter Schafer, Kai-Uwe Thiessenhusen, and Peter Wagner,“ A traffic information system by means of real-time floating-car data, ” DLR, January 2002.

[5] Yang, Yuchen Ou, Dongxiu Xue, Lixia Dong, and Decun,“ Infrastructure-based Detection Scheme of Malicious Vehicles for Urban Vehicular Network, ”Transportation Research Board 96th Annual Meeting, January 2017.

[6] Gongjun Yan, Stephan Olariu, and Michele C. Weigle,“ Providing VANET security through active position detection, ”Computer Communications, Vol.31, pp.2883-2897, July 2008.

[7] SPACE-TIME Engineering. Available from: https://www.spacetime-eng.com/en/ 2017.07.07

[8] Daniel Jiang, Luca Delgrossi,“ IEEE 802.11p: Towards an International Standard for Wireless Access in Vehicular Environments, ”Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE, pp.1550-2252, May 2008.

[9] Stephan Eichler,“ Performance Evaluation of the IEEE 802.11p WAVE Communication Standard, ”Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th, pp.1090-3038, October 2007.

[10] Xinzhou Wu, Sundar Subramanian, and Ratul Guha,“ Vehicular Communications Using DSRC Challenges, Enhancements, and Evolution, ” IEEE Journal on Selected Areas in Communications, Vol.31, No.9, pp.399-408, September 2013.

[11] Ryosuke Sugisaka, Asahi Aono, Ryota Ayaki, and Kenya Sato,“ Implementation and Evaluation of the Dynamic Map Based on Web, ” DICOMO2017, June 2017.

Internet of Things and Cloud Computing Enabling Circular Economy

A tool rental service

Johanna Kallio, Maria Antikainen, Outi Kettunen
VTT Technical Research Centre of Finland Ltd.
Espoo, Finland
e-mail: johanna.kallio@vtt.fi, maria.antikainen@vtt.fi, outi.kettunen@vtt.fi

Panu Korpipää
Finwe Ltd.
Oulu, Finland
e-mail: panu.korpipaa@finwe.fi

***Abstract* - Internet of Things, cloud computing and big data analytics are technological innovations that have the power to transform traditional businesses. These technologies can enable and accelerate a circular economy with closed material loops on a broader scale. We aim at providing information on how to disrupt prevailing linear business models by employing digital data, cloud computing and Internet of Things technologies. We give the reader an overview on circular economy and most prominent digital technologies affecting the incumbents of industry. We present the current deployment of a tool rental service and develop a scenario anticipating the possible future of the tool rental service. The envisioned tool rental scenario provides understanding on the effects of digital technologies and helps companies in identifying more sustainable and circular business models.**

Keywords - circular economy; Internet of Things; cloud computing; tool rental service.

I. INTRODUCTION

This paper explores how Internet of Things (IoT) and cloud computing can advance circular economy business model [1]. The concept of a circular economy (CE) describes an economy with closed material loops. CE focuses on reusing materials, and creating added value in products through services and technology-enabled smart solutions. This implies that the concept of CE is a continuous development cycle that aims to keep products, components and materials at their highest utility and value at all times, distinguishing between technical and biological cycles [2]. If EU manufacturing sector would adopt a CE business model, net material costs savings could be worth up to 570 billion euros per year and growth opportunities 320 billion euros by 2025 [3]. The circular concept fosters also wealth and employment generation against the backdrop of resource constraints [4], [5]. This transformation from linear "take-make-dispose" economy to circular one requires disruptive innovation in business models and technologies.

IoT is a general term relating to the various technologies for connecting, monitoring and controlling devices such as sensors, home appliances, vehicles and industrial devices over a data network. We define IoT as a computing concept where internet-enabled physical objects (e.g., sensors, actuators, tags, smart machines) can network and communicate with each other to achieve greater value and services by exchanging data and producing new information [6]. IoT relies on the three pillars related to the ability of smart objects: i) *to be identifiable*, ii) *to communicate* and iii) *to interact*. When objects can sense the environment and communicate, they become tools for understanding complexity and responding to it [6]. IoT is considered as being one of the key enablers for enhancing CE at large [3].

Advancement in digital technologies is making the current linear take-make-dispose economy more and more efficient, but it still fails to address resource and natural capital issues. However, this new connectivity between digital technologies and economy also offers the possibility to re-think the underlying system and support the development of CE. By combining the principles of CE with IoT and cloud technologies, greater opportunity may arise to scale new business models more effectively [7].

In this article, we present the current implementation of a CE tool rental service and outline a scenario suggesting the possible future of the tool rental service. The envisioned tool rental scenario offers understanding on the potential of digital technologies and helps stakeholders to identify increasingly sustainable and circular business models. The remainder of this article is organized as follows: Section II presents the main principles of CE and the benefits that digitalization can provide. This is followed by representative overview of potential IoT and cloud computing technologies to be capitalized on the CE business. In Section IV, we describe the current deployment of the tool rental service, as well as develop the scenario for the future developments. Finally, Section V concludes the paper and provides some directions for future work.

II. CIRCULAR ECONOMY AND DIGITALIZATION

A circular economy is commonly defined as an industrial system that is restorative or regenerative by intention and design [8], [9]. In CE, new business models are developed to reduce the need for virgin raw materials and to generate sustainable growth. The basic approach of CE is to eliminate waste by designing out of waste. Products are designed and optimized for a cycle of disassembly, reuse and refurbishment, or recycling, with the understanding that economic growth is based on reuse of material reclaimed from end-of-life products rather than extraction of resources. Circular design makes products easier to disassemble in

preparation for their next round trip. Reuse means the use of a product again for the same purpose in its original form or with little enhancement or change. Refurbishment means a process of returning a product to good working condition by replacing or repairing major components that are faulty or close to failure, and making 'cosmetic' changes to update the appearance of a product, such as cleaning, changing fabric, painting or refinishing. Any subsequent warranty is generally less than issued for a new or a remanufactured product, but the warranty is likely to cover the whole product (unlike repair). Accordingly, the performance may be less than as-new [2].

In CE, the concept of *user* replaces that of *consumer*. Unlike today, when a consumer buys, owns and disposes of a product, in CE, durable products are leased, rented or shared whenever possible [2]. If goods are sold, new models and incentives motivate consumers (users) to return or reuse the products or their components and materials at the end of their primary use. New performance-based business models are instrumental in translating products designed for reuse into attractive value proposals.

Digitalization's role as an important enabler of CE is widely accepted. One major benefit is that digitalization enables building visibility and intelligence into products and assets. Intelligence can be knowledge of the location, condition or availability of assets or materials [7]. Knowledge of the product location enables increased product accessibility and increases the product looping. Especially reverse logistics planning becomes easier. Knowledge of the product condition enables predictive and condition-based maintenance, advanced diagnostics and prognostics of the components and products. Predictive maintenance increases product reliability and availability and enables further remanufacturing with the historical knowledge of the product in order to ensure future lifetime and guarantee the quality of remanufactured products. Knowledge of the availability of the product allows, for instance shared use cases through digital platforms and market places, and improved recycling.

Digitalization also boosts the transformation towards service based business models [2], [3]. Thus, the shift towards product service systems (PSS) is suggested as being one of the key solutions in accelerating the transformation towards CE, and digitalization is a major enabler in this process [4]. RFID technology enables information collection on the usage history of products, which plays a central role in PSS-related business models, for example. This enables tracking the quality of returned products and facilitating the return flows into product life cycle management [10]. Also different digital technologies, such as utilization of artificial intelligence (AI) and data science bring novel ways to improve product life time, availability and looping. Moreover, digitalization can offer transparent access to data on products' resource consumption enabling optimization of product life cycles [11]

CE systems with interrelated cycles consist of large amounts of data. Digitalization provides new means to access this data. Decisions need to be made regarding the products' lifecycle stages, how waste materials should be reused, what type of logistical arrangements are needed and who are the actors involved in the value network. Understanding the value proposition in the value networks is essential [12]. Digitalization provides opportunities for the virtualization of distribution channels. Value can be delivered to the customers through digital channels, e.g., online shops and digital products. This can lead to reduce environmental impact and novel circular business models [12].

In CE, the coordination of material and information flows is crucial. Information about quantity and quality of products and their raw material contents need to be collected and retained. Digital technologies enable keeping the data together with materials in the cycle and make it possible to use waste as a resource [13].

Integration of digital intelligence provides opportunities to distribute knowledge, structure, ownership and different levels of customization. This allows more connected and durable relationships with the customers and end users. In addition, digital solutions enable circular business models through automated monitoring, control and optimization of resources and material flows [13]. CE also requires system knowledge and understanding for optimizing business models during the product and service life cycle. Overall, this is one of the advantages of digitalization: the capability to optimize big data and utilize artificial intelligence for solving circular challenges.

Yet, in order to gain all the above-mentioned benefits, there are still many challenges to be solved and gaps hindering digital technology-aided circular business models implementation. Overall, IoT technologies, cloud computing and digitalization in general have potential to disrupt current prevailing linear business models [14]. For instance, the incumbents of the media and music industries have experienced the enormous forces of start-ups' new business models based on digital data and technologies.

III. SENSORS AND COMMUNICATION

The concept of combining computers, sensors and networks to monitor and control devices has existed for decades [15]. Advancements in digital technologies are not only limited to embedded technologies, wireless communication protocols and small devices, but also huge amounts of data are being generated by these devices and can be utilized to improve businesses. In this section, we give a representative overview of available IoT technologies affecting CE.

A. Sensors

Recent improvements in wireless technologies and electronics have enabled the development of low-cost, low power and multifunctional sensors that are small in size and can communicate in short distances. Typically, these sensors consist of sensing, data processing and communication components. The deployment of sensors is mainly driven by three factors; decreasing price, improving computational power and smaller size, which enables their integration into smartphones and other small devices [16].

Sensors are often categorized based on their power sources, i.e., *active* or *passive*. Active sensors emit energy in environment, while passive sensors passively receive energy that is produced externally to the device. Passive sensors

require less energy, but active sensors can be used in harsh environmental conditions. Table I lists typical sensor types based on their functions [17].

The main challenges related to the use of sensors are power consumption, data security and interoperability. Many IoT applications need to run for several years over batteries, but charging and replacement may pose some issues especially in remote areas. However, recent advances in the fields of nano-scale accumulators and energy harvesting techniques have become promising choices [6]. IoT sensors are vulnerable for malicious attacks for several reasons; first, sensors are most of the time unattended that makes them easy to attack. Second, wireless communications make eavesdropping very easy and third, low energy and computing resources do not allow complex security implementations [18].

B. Networking

Data collected with sensors need to be communicated to other locations for integration and analytics. Internet Protocol (IP) is an open protocol that provides unique addresses to various internet-connected devices. IP networking represents a scalable and platform-independent technology having interoperability as the most essential objective. There are two IP versions, IP version 4 (IPv4) and IP version 6 (IPv6), which is the next generation protocol designed to provide several advantages over IPv4 [19].

TABLE I. DIFFERENT SENSOR TYPES.

Sensor types	Description
Temperature	Temperature sensor measures the amount of heat or cold. Temperature sensing is essential in quality control of an environment or internal factors.
Humidity	Humidity sensor detects the presence of water in the air or a mass. Humidity sensing in important for instance in industrial processes and human comfort.
Position	Position sensor measures the absolute or relative position of an object.
Occupancy	Occupancy sensor detects the presence of an object (e.g., people) even when they are stationary.
Motion	Motion sensor detects moving objects, for instance people or animals.
Velocity	Velocity sensor measures how fast the object moves or rotates.
Acceleration	Acceleration sensor measures changes in the velocity of objects that means how fast the object's speed changes.
Pressure	Pressure sensor measures the physical force applied by liquids or gases.
Flow	Flow sensor measures the rate of liquid or gas movement. Flow sensing is used especially in medical technology, industrial and building automation.
Sound	Sound sensor measures the level of noise in the environment.
Light	Light sensor measures the change or presence of light in the environment.
Chemical	Chemical sensor measures the concentration of chemicals, such as carbon dioxide.

TABLE II. COMMON NETWORK TECHNOLOGIES BY CONNECTION TYPES.

Connection type	Network type		
	PAN	***LAN***	***WAN***
Wired	USB	Ethernet	
Wireless	Bluetooth, RFID, NFC, Wi-Fi, ZigBee	Wi-Fi, WiMAX	WiMAX, LoRaWAN, Cellular technologies

Network technologies can be classified as wired or wireless. The main advantage of a wireless network is that users and devices can move around freely within the area of the network and get an internet connection, while wired connections are still useful for relatively more reliable, secured and high-volume network routes. The choice of technology depends mostly on the physical range to be covered [20]. When devices communicate with other nearby devices, they can use wireless personal area network (PAN**)** technologies. A local area network (LAN) connects networked devices over a limited area, such as an office building, school or home. When data have to be transferred over a large geographical distance, wide area network (WAN) technologies are used. The Internet is an example of the world's largest WAN.

The most common short-range wireless network technologies are Bluetooth [21], Near Field Communication (NFC) [22], Radio Frequency Identification (RFID) [23], Wi-Fi [24] and ZigBee [25]. Respectively, the most commonly employed wider range wireless network technologies are cellular technology, such as 3G or 4G, Low Power Wide Area Network (LoRaWAN) [26] and WiMAX [27]. Table II gives an overview of the mentioned networking technologies by their connection types [28].

The main challenges related to adoption of network technologies are associated with security, network interconnections and power consumption. Security will be a major concern wherever networks are deployed. Data transmission can be relied on several networks, such as Ethernet, cellular or other wireless networks, and different network technologies are often connected with gateways [29]. In addition to security, these interconnections pose challenges by adding complexity. In order to tackle the problem networked devices have with power consumption, energy-efficient networking is explored in research communities [30].

C. Data communication

Data communication focuses on how the data is streaming from the sensors towards the databases and application back-ends. The data communication includes a set of protocols that have been built for high volumes and large networks of devices. Constrained processing capabilities and limited battery resources restrict communication in sensor systems. Typical sensor application and data communication protocols considering processing capability and energy consumption are the following three [31]:

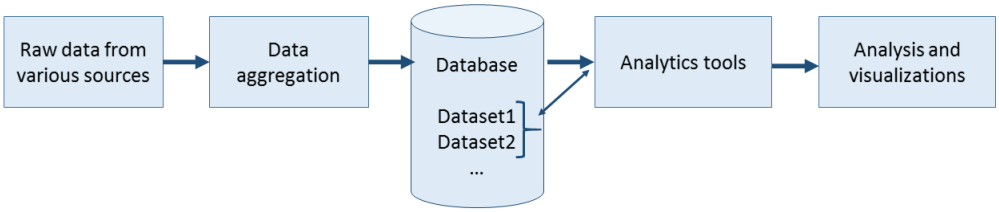


Figure 1. High-level overview of data processing.

- Message Queue Telemetry Transport (MQTT), which is a lightweight publish/subscribe based message protocol especially well-suited for running on limited computational power and lean network connectivity.
- Constrained Application Protocol (CoAP), which is designed specifically for machine-to-machine (M2M) applications, such as building and energy automation.
- Hypertext Transfer Protocol (HTTP), which is attractive option because of availability and compatibility of the legacy HTTP-stack on various platforms.

In many IoT solutions and sensor data platforms, the observational data from sensors is typically stored in simple and convenient data formats, which are easily accessible and already supported in many programming languages and common software development frameworks [32]. The most important examples of such data formats commonly applied in sensor systems are JSON, XML and CSV.

IV. CLOUD COMPUTING

The previous Section III presented the sensor, networking and data communication technologies that are relevant for IoT applications. Figure 1 depicts a simplified high-level data handling process for IoT applications receiving data from multiple data sources and in different data formats. The process of extracting and transforming the data into a suitable format and storing it in databases is often called data aggregation. The data aggregation prepares the data for analysis. This section focuses on cloud computing, which complements the previous IoT technologies in terms of data processing, storage and analysis.

A. Cloud platforms

We define an IoT platform as a middleware and infrastructure that enables interaction between the end-users and physical objects [32]. The IoT platform can be either local or provisioned from a cloud as illustrated in Figure 2. With cloud technology, platform's computation and storage resources can be made available on a need basis, without requiring major investment in new hardware or programming. Platform-as-a-Service (PaaS) is a category of cloud computing services for developing, managing and delivering applications. Software-as-a-Service (SaaS), on the other hand, provides an access to a cloud-based software or applications through which data can be stored and analyzed [33].

The most cloud platforms are implemented with the Representational State Transfer Application Programming Interface (REST API) [32]. Employing REST API also gives the advantage of easy integration with other web services, cloud environments and data processing toolchains. Thus, REST-based CoAP and HTTP are widely utilized data communication solutions to interconnect devices to the IoT platform.

Many vendors, such as Microsoft, HP, IBM and Oracle, provide commercial cloud-based IoT platforms for connecting sensors and actuators to the Internet. In addition, several open-source IoT platforms are available and often propose their own communication or middleware solutions. Mineraud et al. [32] provides an extensive evaluation of a number of available proprietary as well as open-source IoT platforms.

The main challenges of cloud platforms are associated with security, integration of multifold technologies, data ownership and data processing [32]. Efficient security mechanisms are essential for IoT platforms because interactions between different objects, as well as with humans, must be secured [18]. Furthermore, information produced from data is a threat for privacy and anonymity of users must be retained [34]. The interaction of IoT platforms and sensing devices is hindered by various technologies and missing de-facto communication standards. Currently, the interoperability issues are tackled by gateways, which support new types of devices (see Figure 2). The ownership of created

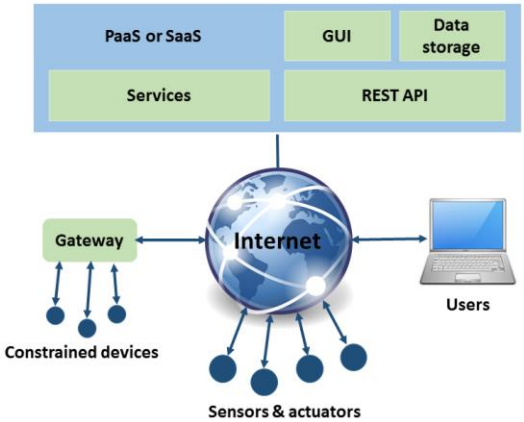


Figure 2. Cloud-based platform [27].

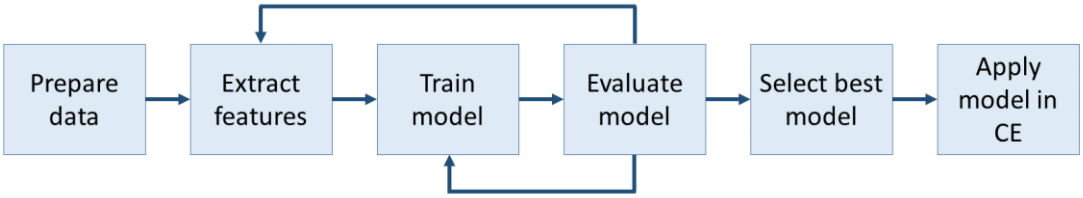


Figure 3. Machine Learning process in circular economy.

data is also a complex issue and regulation varies depending on whether the data is personal or machine-generated.

B. Data processing and storage

The key step for a successful analysis and use of any data is an organized and systematic approach to collecting, arranging, and presenting the incoming data [35]. Data processing methods can be divided in batch and stream processing. *Batch processing* starts with data acquisition and storing and continues with processing of the stored data [33]. For instance, Hadoop [36] is an open source batch processing framework targeted for distributed data storage and processing of unstructured data. Hadoop employs the MapReduce programming model for the processing of large-scale data clusters [37]. Hadoop Distributed File System is optimized to store large amounts of data. Another example of batch processing framework is Apache Spark [38], which is also capable of stream processing.

Especially large volumes of data and new IoT applications require real-time processing, where data items are processed as soon as they become available. This is called *stream processing* and it facilitates real-time action on the data, as well as filtering and aggregating it for efficient storage [33]. Open-source frameworks, such as the Apache tools Samza [39] and Storm [40] have been created for real-time or near real-time processing of streaming data. Flink [41], on the other hand, is an open-source hybrid framework for stream processing, which can also manage batch processing.

Compared to relational databases, requirements on performance and scalability typically presume non-relational databases. NoSQL (Not only SQL) databases are designed for storing unstructured data. Popular databases for storing streaming data are Apache Cassandra [42] and HBase [43]. True to form, all data processing frameworks and storing tools have different benefits and limitations, and the selection of tools depends on application.

C. Data analytics

Data analytics refers to examining of raw data with the purpose of refining this data as useful information that can be used as knowledge. Data analytics can be divided into three different categories: a) *descriptive* analytics describing what the data looks like, b) *predictive* analytics predicting what is going to happen, and c) *prescriptive* analytics describing what should happen to reach the goal [44].

Machine learning refers to system's ability to learn without being explicitly programmed. Machine learning explores the development of algorithms that can learn from and make predictions on data. Machine learning methods are often categorized as supervised or unsupervised. *Supervised learning* can apply what has been learned in the past to new data. *Unsupervised learning* can draw inferences from datasets [45].

Machine learning process is depicted in Figure 3. The process starts with data preparation. In this phase, the available raw data is visualized to get an overview of the data. The data may require some pre-processing actions, such as scaling, normalization and actions on missing data. The data is split in two parts, from which the first part is used for training and second part for testing. In the second phase, the most relevant features (i.e., variables) are selected. These features are used in construction of an accurate predictive model from various machine learning algorithm candidates.

Next, the chosen machine leaning algorithms are trained by using the training dataset. In the evaluation phase, the quality of the trained model is assessed with the test dataset. The target of the evaluation is to test the performance of the data model with testing dataset that has not been used for training the model. Approximately 20 - 30 % of the original data is used for the evaluation purposes. Another commonly used method is to leave a part of the whole dataset as test dataset, while the rest is used as training dataset, and repeat the process over the whole dataset. The method is called cross-validation. Before selecting the best solution, there might be several iteration cycles from feature extraction to model evaluation. Finally, the best model can be used for the target application in CE. The function performed by the model can relate to detecting a certain pattern in real time from sensor data or conducting a prediction based on the trend or pattern of the past data.

Different data analysis techniques are used based on the type of the addressed problem. Classification is a supervised machine learning technique, which produces a set of models or functions that distinguish dataset into different classes. The purpose is to predict the class for objects whose class is not known. There are plenty of methods that can be used for classifying the data, such as decision trees, frame- or rule-based expert systems, neural networks, Bayesian network and support vector machines. Clustering is typically an unsupervised learning technique, which divides dataset into meaningful groups having similar patterns. Clustering is a common technique for statistical data analysis and used for example by e-commerce, industry and health care sectors [46], [47].

V. TOOL RENTAL SERVICE

In this section, we describe how the previously presented digital technologies, namely sensors, cloud platforms and data analytics can be applied to a tool rental service to promote CE. We present a rapid experiment of the tool rental service and develop a scenario envisioning the possible future of the tool rental service. Our aim is to provide understanding on the possibilities and benefits of digital technologies for sustainable and circular business models. In general, digital technologies and better utilization of digital data can build visibility and intelligence into products and services.

A. Approach and data collection

Our AARRE project (Capitalising on Invisible Value - User-driven Business Models in the Emerging Circular Economy) explored user-driven circular business models and collaborated with multiple Finnish companies, Finnish organizations and Finnish decision makers in the CE field. The idea of a tool rental service is to offer an alternative for purchasing of tools, such as electric tools and cleaning equipment, which are used infrequently in urban economy. This kind of sharing economy can be an ecological option in certain conditions and on the other hand, facilitates the storage problem of goods in urban housing [48]. The goal of our empirical study is to provide input for the discussion on how to disrupt current prevailing linear business models by employing digital data and IoT technologies.

The planning and rapid experimenting of the tool rental scenario is based on several discussions with eight (8) AARRE project researchers, one start-up entrepreneur, and various companies. We also use interviews of nine (9) users and potential users of Liiteri, and Owela innovation tool results as a background material.

In addition, prior to the experimentation, a consumer panel was held in the project in order to identify current and future needs, ideas, enablers and barriers related to CE business models. There were 42 panelists divided into 5 groups. The discussion in each group was led by 1 - 2 scientists and based on a uniform list of questions. Several issues related to renting, leasing, borrowing and sharing were identified where IoT can be an enabler. Based on the discussions, the identified barriers related to renting, leasing, borrowing and sharing, are current lack of information on selection, availability, location and condition of the items. There is also a need for some additional information, such as instructions for use, as well as a need to monitor the usage in some cases. Tracking and tracing information in the logistics process would make renting easier and more cost-efficient. Technology could also support in providing a rating system concerning both the users and service providers.

B. Current deployment of the tool rental service

Our AARRE project participated in the design and implementation of a rapid experimental tool rental service called Liiteri [49] in collaboration with Finnish IT-startup CoReorient [50] and hardware store K-Rauta. The other co-operation partners were Helsinki Region Environmental Services Authority HSY, Technology Industries of Finland, SER-kierrätys, City of Espoo, Purjebägit Oy, Kierrätysverkko Oy, Metrosuutarit.fi, Pyörähuoltoovelle.fi and Kauppahalli24.fi.

Liiteri is an online platform where users can rent electric tools and house cleaning equipment. By registering to the Liiteri service, the user can choose the desired product and renting date via the online platform. The rental payment is made via the online service at the same time with product selection. When the payment has been processed, the user gets an access code to the 24/7 Liiteri self-service point, which is an intelligent container in the city centre of Helsinki. The user can pick up the rented gear any time from the Liiteri self-service point located at a central place of public transport connections. Alternatively, the user can choose a crowdsourced PiggyBaggy home delivery service [51]. An initial experiment to examine the utility of the Liiteri tool rental service was conducted and reported in another study [46].

C. Scenario for the tool rental service

In this section, we describe a future business scenario for the previously described Liiteri tool rental service. The scenario complements the available Liiteri tool rental service and aims to provide information for companies about the future possibilities of digital data collection, processing and analytics. Figure 4 depicts the envisioned tool rental service scenario. The data is collected by various means, such as sensors, mobile application, web services and data logs. Depending on the data and its application, different data processing and storing options can be used. Respectively, the choices of data analytics methods and visualizations are always dependent on the target application.

1) Personalized real-time services

Asset tracking refers to the method, which can be used to provide information on the location of physical equipment by using scanning a barcode or using location aware technologies, such as GPS, RFID or Bluetooth Low Energy (BLE). The asset tracking can be real-time or based on connected checkpoints. A node that is aware of its location and can send signals to smartphones or other mobile devices is called a beacon. BLE-based beacons are a worthy choice for improving shopping experience as they have high accuracy and support by the majority of mobile devices [52]. User's smartphone can interact with beacons placed in the Liiteri self-service point via a Liiteri mobile application (later referred to as the Liiteri app). BLE-based *beacons can be used for mobile door opening* when entering the 24/7 Liiteri self-service point [53].

Furthermore, the beacons can be utilized in *store navigation*. The beacons can enable users to be recognized in real-time when entering the 24/7 Liiteri self-service point and help them find easily the selected tool or other rented gear with nearby push notifications feature. A payment for the selected product can be discharged using a *beacon-based mobile payment* or a more conventional online payment service provided by the Liiteri app.

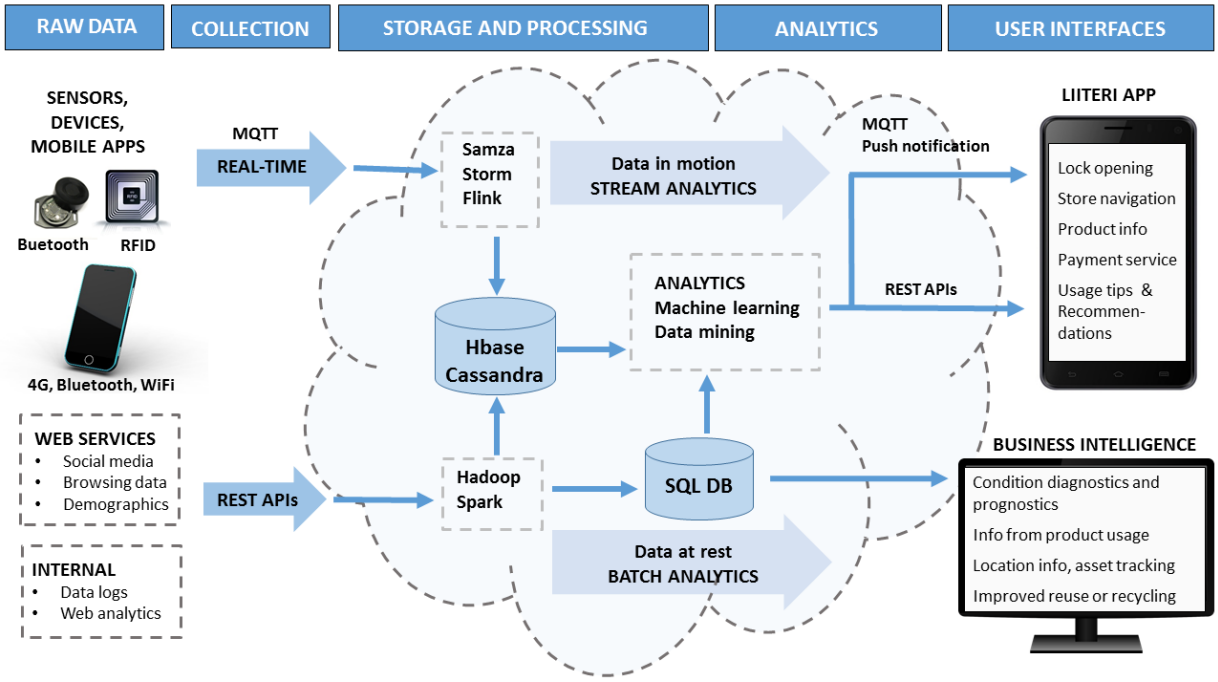


Figure 4. IoT-enabled tool rental service business scenario.

2) Other personalized services

At home, the users can be provided with a *guided usage service* by scanning the tool (equipped with barcode or RFID tag) with their phone, and even a guided replacement service in case of a broken part can be possible. The data from web services, such as social media, can be integrated with Liiteri data logs to provide other personalized services via REST APIs. For instance, the rental profile including demographic data and browsing history from the Liiteri app can be further analysed for *personalized recommendations*. The Liiteri app can also allow the users to *access product information and reviews* to help them make their decision. In these services, the data is analysed in batches and possible data processing frameworks can be Apache Hadoop or Spark. For data storage, Hadoop file storage or HBase are possible candidates.

The tool availability and location information can be utilized to *provide a connection with next user*. This promotes collaboration among users and facilitates *crowdsourced delivery of tools* from user to user. During the delivery process, communication can be handled via the Liiteri app. *The deliverer can earn points*, which she or he can spend for the tool rental or redeem for cash. Home delivery service of the tools, in turn, can utilize the real-time navigation information and the customer's location information for route planning and *optimization of delivery routes* to reduce driving time, fuel consumption and exhaust gases.

3) Condition-based maintenance

Sensors can measure for example acceleration, temperature, vibration and humidity of tools or parts of the tools. The sensor data can be exploited in business intelligence in many ways, such as in condition-based maintenance (CBM). The condition-based maintenance is a maintenance procedure based on the information collected through condition monitoring and can be used for *diagnostics* and *prognostics* [54]. Prognostics based maintenance is often called as predictive maintenance. The goal of condition-based maintenance is to increase the lifetime of tools, improve the maintenance efficiency and safety. The condition-based maintenance of the tools can employ streaming data from multiple sensors, both from component or system level. This scenario is composed of various tools in different locations and cloud computing for (near) real-time monitoring of these tools.

The tools can be equipped with affordable sensors connected to the Liiteri cloud-based platform using cellular networks. The data streams require real-time processing for assessing current failures (diagnostics). On the other hand, batch processing is needed for predicting possible future failures (prognostics). A possible solution for data processing can be Apache Spark or Flink, which are hybrid platforms and capable of stream and batch processing of data. NoSQL database, for instance HBase or Cassandra can be used for data storage. Different machine learning or data mining techniques, such as classification or regression analysis can be used for diagnostics and prognostics purposes. Automatic failure diagnostics can turn refurbishment into a potential and accessible option. A trivial consequence from refurbishment and increasing the lifetime of tools is decreasing the use of natural resources and waste.

TABLE III. SERVICES BASED ON THEIR INTELLIGENCE TYPES.

Intelligence type	Service	Description
Knowledge on location	Mobile door opening	Beacon-based service for mobile door opening when entering the 24/7 Liiteri self-service point.
	Mobile payment	Beacon-based mobile payment service for the selected product.
	Store navigation	Beacon-based real-time navigation service.
	Guided usage	Basecon-based guided usage service for the rented tool.
	Optimization of delivery route	Optimization of delivery routes based on real-time navigation information and the next user's location information.
Knowledge on condition	Diagnostics	Real-time service for assessing current failures.
	Predictive maintenance	Maintenance service based on predicting future failures.
	Decision support for future loops	Prognostics service for predicting RUL to support on decision between the reuse, remanufacture or recycling.
	Improved product design	Design service based on tool usage and user feedback data.
Knowledge on availability	Crowdsourced delivery of tools	Service based on tool availability and location information to provide a connection with next user.

4) Improved reuse, remanufacture and recycling

A prognostics model employing machine learning can be developed for predicting a Remaining Useful Lifetime (RUL) and considering sustainability aspects in decision-making, i.e., deciding when maintenance is economically viable, environmentally bearable and equitable compared to reuse, remanufacture or recycle [55]. In this scenario, batch processing is used, and for instance Apache Hadoop or Spark are possible data processing options. For data storage, HBase would be a good choice.

The main idea of predicting remaining useful lifetime is to build a machine learning model for normal and failure events based on one or more features, which are in this case failure predicting variables, such as rising temperature or unusual vibration. The analysis results help to decide between the *reuse, remanufacture or recycling* activities in order to maintain both economic and environmental value of the tools or their parts as high as possible. In addition, the tool usage data collected with sensors and the user feedback can be analysed *to improve future product design and performance*.

IoT and cloud computing creates visibility and intelligence into assets. As described in Section II, this intelligence can be knowledge of the location, condition or availability of the tools. Table III lists the envisioned services according to their intelligence categories as introduced in [7].

VI. CONCLUSION AND FUTURE WORK

IoT technologies and digitalization in general enable novel business models based on CE, offering a significant potential to disrupt current prevailing linear business models [56]. As a case study to evaluate the hypothesis, we presented an instance of a CE tool rental service and outlined a scenario suggesting the possible future of the concept.

A. Discussion

Tool sharing itself is not a totally novel idea and for instance city of Toronto has a tool library [57]. The Toronto tool library is a centralized warehouse where people can store and share their tools to the local neighborhood with an annual membership payment. The novelty of our tool rental service lies on more advanced use of technology. The envisioned tool rental service scenario uses sensors, networking, cloud computing, and data analytics for selling services instead of goods, for designing products for regeneration and for creating added value through services. Generally, offering services instead of selling goods reduces the environmental footprint of product manufacturing and the private ownership of goods. The target of this study was to awake discussion among companies on how to create CE business by employing digital data, IoT and cloud computing.

Organizing and managing the growing amount of data from increasing number of sources is crucial for a successful novel CE service. A notable data analytics challenge is related to heterogeneous data sources and extracting data from different data storage locations in large scale. The chosen methods need to overcome variety, heterogeneity and noise of the data. Data storage is required to be highly robust and resilient to failures. Using the available data to a maximum benefit requires a comprehensive understanding of the meaning, behavior, and relations of the data. Data mining is typically a human-operated process, where the data is explored by supervised or unsupervised methods, aiming to gain understanding and to find patterns or models that can lead to a business advantage. Acquired models can then be deployed to automatically process the dedicated part of the data stream or storage.

Our current database implementation is centralized but multidisciplinary colloboration towards more circular businesses could beneftit from the use of decentralized databases. Blockchain technology is a decentralized database that allows for the chronological recording and secure storage of transaction data. Blockchains can be used to register tool rentals and transfers between tool renters making system more transparent and increasing trust. Even though blockchains provide pseudonymity meaning that transactions are transparent but not explicitly connected to individuals, sensitive information should not be shared [58].

When applying IoT and cloud computing in real-world implementations information security must always be considered. Information security and privacy procedures, in terms of data integrity, access control, and system availability, need to be in place to protect the information and service provisioning of relevant actors [18]. Moreover, the scenario raises questions about who owns the created data, on which terms the data can be shared with others, and what kind of legislation should be in place to prevent the sharing of sensitive data. General Data Protection Regulation (GDPR) is an EU wide regulation for data privacy that aims to protect all EU citizens' personal data and regulate the international business related to the personal data [59]. Before the

commercial implementation of the envisioned services, considerations regarding data protection and privacy must be done according to the GDPR. All personal data must be protected, stored and processed appropriately.

In order to create future services that are increasingly capable of adapting to the user and the surrounding external conditions, development in technologies enabling learning from data have become pivotal. The users are individuals, whose needs and preferences vary. The service provider operates on a large population, but gains an advantage from being able to adapt to the user and threat each one individually. There are several use cases in CE, where machine learning could be applied in general. Related to the service addressed in this study, the most potential usages include detecting the user preferences in current context in order to guide suitable actions for the user, and modeling different types of users having varying types of purchase behavior. As an additional benefit, adapting to the user enables targeting marketing or commercials for the user based on their past behavior. Predicting product demand highly accurately becomes feasible, which facilitates further cost savings. Moreover, AI and machine learning methods present a considerable prospect in materials and products process flow optimization, which becomes especially crucial when the system is scaled up.

B. Limitations and future work

As in every research, there are some noteworthy limitations with this study. The scenario planning has been qualitative and it includes researchers' and interviewees' subjective interpretation. The previous study [48] examined the utility of the Liiteri tool rental service and the results indicated that renting could be an attractive choice to consumers, if crucial consumer expectations are identified and met. This attitude change is partly attributable to technological improvements, such as widespread digital cloud platforms, which make sharing of goods easier. On the other hand, especially young people prefer services to ownership of goods.

In order to proceed towards commercial utilization of the presented concept, further work on the IoT-enabled tool rental service scenario includes addressing multiple technical, usability, and profitability aspects, in addition to the environmental viewpoint. The main technological challenges include the overall cost-efficient logistics on large scale, data security, interoperability of IoT sub-systems, and interface with the consumers. The usability of the service must meet or exceed the level of current modern competing e-commerce platforms. Overall, the business challenges are the same as for e-commerce and retail services in general. To be commercially viable, the service must maximize the user satisfaction and minimize the costs through intelligent use of digital technologies.

Future work needs to evaluate which parts of the system are efficient as automated, and which parts need to be managed by human operators for an optimal profitability, usability and security. The challenges and opportunities are similar as, for example, in Amazon's automated grocery store recently deployed within their office building. The store relies on cameras and sensors to detect what shoppers take from the shelves, and what they put back. Cash registers become largely obsolete, as customers are billed after leaving the store using credit cards, based on the virtual shopping cart collected while physically shopping.

Full examination is needed on the types of products that provide most benefit from the CE service concept in terms of environmental benefit. Moreover, next steps of future research endeavours could focus on implementation of a next generation IoT-enabled CE service for a larger group of people, with the systematic collection of consumer experiences for further service improvement.

Despite the various remaining challenges, there is a growing amount of consumers that see the environmental friendliness as an added value factor. This could facilitate the accumulation of consumers in the early phases of development, towards ultimately better collaboration between the human use of technology and the environment.

ACKNOWLEDGMENT

This research has been conducted as a part of the AARRE (Capitalising on Invisible Value – User-Driven Business Models in the Emerging Circular Economy) project. The authors would like to express their gratitude to the Green Growth Programme of Funding Agency Business Finland (formerly known as Tekes), the Technical Research Centre of Finland (VTT), the case companies and other parties involved in the AARRE project. In addition, the authors want to acknowledge the valuable contribution of the people who took part of this study.

REFERENCES

[1] J. Kallio, M. Antikainen and O. Kettunen, "A tool rental service scenario - IoT technologies enabling a circular economy business model," Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOM), 2017 The Eleventh International Conference on, IARIA, Nov. 2017, pp. 60-65.

[2] Ellen Mac Arthur foundation, "Towards the Circular Economy vol 1: Economic and business rationale for an accelerated transition," 2012.

[3] J. Manyika, M. Chui, J. Bughin, R. Dobbs, P. Bisson and A. Marrs, "Disruptive technologies: Advances will transform life, business, and the global economy," McKinsey Global Institute, 2016.

[4] P. Ghisellini, C. Cialani and S. Ulgiati, "A review on circular economy: the expected transition to a balanced interplay of environmental and economic systems," Journal of Cleaner Production 114, pp. 11-32, 2016, doi: 10.1016/j.jclepro.2015.09.007.

[5] A. Wijkman and S. Kristian, "The circular economy and benefits for society: Jobs and Climate Clear Winners in an Economy Based on Renewable Energy and resource Efficiency," The Club of Rome Report, November 2015.

[6] D. Miorandi, S. Sicari, F. De Pellegrini and I. Chlamtac, "Internet of things: Vision, applications and research challenges," Ad Hoc Networks 10(7), pp. 1497-1516, 2012, doi:10.1016/j.adhoc.2012.02.016.

[7] A. Morlet et al. "Intelligent assets: Unlocking the circular economy," Ellen Mac Arthur foundation, 2016.

[8] J. Butterworth et al. “Towards the Circular Economy vol 3: accelerating the scale-up across global supply chains,” Ellen MacArthur Foundation, 2014.

[9] D. A. R. George, B. C. Lin and Y. Chen, "A circular economy model of economic growth," Environmental Modelling & Software 73, pp. 60-63, 2015, doi: 10.1016/j.envsoft.2015.06.014.

[10] A. Pagoropoulos, D. C. A. Pigosso and T. C. McAloone, ”The Emergent Role of Digital Technologies in the Circular Economy: A Review,” Procedia CIRP, 2017, pp. 19–24, doi: 10.1016/j.procir.2017.02.047.

[11] H. Kagermann, “Change Through Digitization—Value Creation in the Age of Industry 4.0,” Manag. Perm. Chang., Wiesbaden: Springer Fachmedien Wiesbaden, pp. 23–45, 2015, doi:10.1007/978-3-658-05014-6_2.

[12] V. Salminen, H. Ruohomaa and J. Kantola, ”Digitalization and big data supporting responsible business co-evolution,” Adv. Intell. Syst. Comput., vol. 498, Springer, Cham, pp. 1055–67, 2017.

[13] H. Wilts and H. Berg, “The digital circular economy: can the digital transformation pave the way for resource-efficient materials cycles?,” International Journal of Environmental Science & Natural Resources, 7(5), pp. 1-4, 2017, doi:10.19080/IJESNR.2017.07.555725.

[14] M. E. Porter and J. E. Heppelmann, “How smart, connected products are transforming competition,” Harvard Business Review, 92(11), pp. 64-88, 2014.

[15] M. Weiser, “The computer for the 21st century,” IEEE pervasive computing, 1(1), pp. 19-25, 2002.

[16] J. Greenough. “The Internet of everything 2016,” BI Intelligence, 2016.

[17] J. Fraden, Handbook of modern sensors: physics, designs, and applications, Springer Science & Business Media, 2015.

[18] R. Roman, J. Zhou, and J. Lopez, “On the features and challenges of security and privacy in distributed internet of things”, Computer Networks, 57(10), pp. 2266-2279, 2013, doi: 10.1016/j.comnet.2012.12.018.

[19] The IPv6 Forum [Online]. Available from: http://www.ipv6forum.com/ 2018.05.29.

[20] N. Kaur and S. Monga, “Comparisons of wired and wireless networks: A review,” International Journal of Advanced Engineering Technology, V(II), pp. 34-35, 2014.

[21] Bluetooth, “A look at the basics of Bluetooth technology” [Online]. Available from: http://www.bluetooth.com/Pages/Basics.aspx 2018.05.31.

[22] NearFieldCommunication.org [Online]. Available from: http://nearfieldcommunication.org/ 2018.05.29.

[23] R. Want, “An introduction to RFID technology,” IEEE Pervasive Computing, 5(1), pp. 25-33, 2006, doi: 10.1109/MPRV.2006.2.

[24] Wi-Fi Alliance [Online]. Available from: http://www.wi-fi.org/ 2018.05.29.

[25] ZigBee [Online]. Available from: htpp://www.zigbee.org/ 2018.05.31.

[26] LoRa Alliance [Online]. Available from: https://www.lora-alliance.org/ 2018.05.29.

[27] WiMAX Forum [Online]. Available from: http://wimaxforum.org/ 2018.05.29.

[28] D. E. Comer, Computer networks and internets, Pearson, 2015.

[29] J. Latvakoski, A. Iivari, P. Vitic, B. Jubeh, M. B. Alaya, T. Monteil and M. Kellil, ”A Survey on M2M Service Networks. Computers,” 3(4), pp. 130-173, 2014, doi: 10.3390/computers3040130.

[30] D. Feng, C. Jiang, G. Lim, L. J.Cimini, G. Feng and G. Y. Li, “A survey of energy-efficient wireless communications,” IEEE Communications Surveys & Tutorials, 15(1), pp. 167-178, 2012, doi: 10.1109/SURV.2012.020212.00049.

[31] V. Karagiannis, P. Chatzimisios, F. Vazquez-Gallego and J. Alonso-Zarate, “A survey on application layer protocols for the internet of things,” Transaction on IoT and Cloud Computing, 3(1), pp. 11-17, 2015.

[32] J. Mineraud, O. Mazhelis, X. Su and S. Tarkoma, “A gap analysis of Internet-of-Things platforms,” Computer Communications, 89, pp. 5-16, 2016, doi: 10.1016/j.comcom.2016.03.015.

[33] M. Díaz, C. Martín and B. Rubio, "State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing,” Journal of Network and Computer Applications, 67, pp. 99-117, May 2016, doi: 10.1016/j.jnca.2016.01.010.

[34] S. Satyadevan, B. S. Kalarickal and M. K. Jinesh, “Security, trust and implementation limitations of prominent IoT platforms,” proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), Springer, Cham, 2015, pp. 85-95, doi: 10.1007/978-3-319-12012-6_10.

[35] D. Pyle, Data preparation for data mining (Vol. 1), Morgan Kaufmann, 1999.

[36] Apache Hadoop [Online]. Available from: http://hadoop.apache.org/ 2018.05.29.

[37] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters”, Communications of the ACM, 51(1), pp. 107-113, 2008, doi: 10.1145/1327452.1327492.

[38] Apache Spark [Online]. Available from: https://spark.apache.org/ 2018.05.29.

[39] Apache Samza [Online]. Available from: http://samza.apache.org/ 2018.05.29.

[40] Apache Storm [Online]. Available from: https://storm.apache.org/ 2018.05.29.

[41] Apache Flink [Online]. Available from: https://flink.apache.org/ 2018.05.29.

[42] Apache Cassandra [Online]. Available from: http://cassandra.apache.org/ 2018.05.29.

[43] Apache HBase [Online]. Available from: http://hbase.apache.org/ 2018.05.29.

[44] T. H. Davenport, “Competing on Analytics,” Harward Business review, 84(1), p. 98, 2016.

[45] R. S. Michalski, J. G. Carbonell and T. M. Mitchell, eds. Machine learning: An artificial intelligence approach, Springer Science & Business Media, 2013.

[46] S. Theodoridis and K. Koutroumbas, Pattern recognition, Fourth Edition. Elsevier, 2009.

[47] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible interference, Elsevier, 2014.

[48] M. Antikainen, M. Lammi and H. Paloheimo, “Creating value for consumers in CE - Tools as a service,” The XXVIII ISPIM Innovation Conference (ISPIM), June 2017, p. 1-12.

[49] Liiteri [Online]. Available from: http://www.liiteri.net/ 2018.05.29..

[50] CoReorient website [Online]. Available from: http://www.coreorient.com/ 2018.05.29.

[51] Piggy Paggy Beta [Online]. Available from: http://www.piggybaggy.com/ 2018.05.29.

[52] P. Kriz, F. Maly and T. Kozel, “Improving indoor localization using bluetooth low energy beacons,” Mobile Information Systems, pp. 1-11, April 2016, doi: 10.1155/2016/2083094.

[53] J. Potts and S. Sukittanon, “Exploiting Bluetooth on Android mobile devices for home security application,” Southeastcon 2012, IEEE, 2012. pp. 1-4, doi: 10.1109/SECon.2012.6197001.

[54] A. K. Jardine, D. Lin and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," Mechanical systems and signal processing, 20(7), pp. 1483-1510, 2006, doi: 10.1016/j.ymssp.2005.09.012.

[55] B. Iung and E. Levrat, "Advanced maintenance services for promoting sustainability," Procedia CIRP, 22, pp. 15-22, 2014, doi: 10.1016/j.procir.2014.07.018.

[56] M. Antikainen, M. Lammi and T. Hakanen, "Consumer service innovation in a circular economy – the customer value perspective," The 4th International Conference on Serviceology (ICServ2016), September 2016.

[57] Toronto Tool Library [Online]. Available from: https://torontotoollibrary.com/ 2018.05.29.

[58] P. Boucher, S. Nascimenco and M. Kritikos, "How Blockchain Technology Could Change Our Lives: In-depth Analysis," European Parliament, 2017.

[59] The EU General Data Protection Regulation (GDPR) Portal [Online]. Available from: https://www.eugdpr.org/ 2018.05.29.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
issn: 1942-2679

International Journal On Advances in Internet Technology
issn: 1942-2652

International Journal On Advances in Life Sciences
issn: 1942-2660

International Journal On Advances in Networks and Services
issn: 1942-2644

International Journal On Advances in Security
issn: 1942-2636

International Journal On Advances in Software
issn: 1942-2628

International Journal On Advances in Systems and Measurements
issn: 1942-261x

International Journal On Advances in Telecommunications
issn: 1942-2601