

International Journal on

Advances in Internet Technology



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

Responsibility for the contents rests upon the authors and not upon IARIA, nor on IARIA volunteers, staff, or contractors.

IARIA is the owner of the publication and of editorial aspects. IARIA reserves the right to update the content for quality improvements.

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy or print, providing the reference is mentioned and that the resulting material is made available at no cost.

Reference should mention:

International Journal on Advances in Internet Technology, issn 1942-2652
vol. 14, no. 1 & 2, year 2021, http://www.iariajournals.org/internet_technology/

The copyright for each included paper belongs to the authors. Republishing of same material, by authors or persons or organizations, is not allowed. Reprint rights can be granted by IARIA or by the authors, and must include proper reference.

Reference to an article in the journal is as follows:

<Author list>, "<Article title>"
International Journal on Advances in Internet Technology, issn 1942-2652
vol. 14, no. 1 & 2, year 2021, <start page>:<end page> , http://www.iariajournals.org/internet_technology/

IARIA journals are made available for free, proving the appropriate references are made when their content is used.

Sponsored by IARIA

www.iaria.org

Copyright © 2021 IARIA

Editors-in-Chief

Mariusz Głąbowski, Poznan University of Technology, Poland

Editorial Advisory Board

Eugen Borcoci, University "Politehnica" of Bucharest, Romania
Lasse Berntzen, University College of Southeast, Norway
Michael D. Logothetis, University of Patras, Greece
Sébastien Salva, University of Auvergne, France
Sathiamoorthy Manoharan, University of Auckland, New Zealand

Editorial Board

Jemal Abawajy, Deakin University, Australia
Chang-Jun Ahn, School of Engineering, Chiba University, Japan
Sultan Aljahdali, Taif University, Saudi Arabia
Shadi Aljawarneh, Isra University, Jordan
Giner Alor Hernández, Instituto Tecnológico de Orizaba, Mexico
Onur Alparslan, Osaka University, Japan
Feda Alshahwan, The University of Surrey, UK
Ioannis Anagnostopoulos, University of Central Greece - Lamia, Greece
M.Ali Aydın, Istanbul University, Turkey
Gilbert Babin, HEC Montréal, Canada
Faouzi Bader, CTTC, Spain
Kambiz Badie, Research Institute for ICT & University of Tehran, Iran
Ataul Bari, University of Western Ontario, Canada
Javier Barria, Imperial College London, UK
Shlomo Berkovsky, NICTA, Australia
Lasse Berntzen, University College of Southeast, Norway
Marco Block-Berlitz, Freie Universität Berlin, Germany
Christophe Bobda, University of Arkansas, USA
Alessandro Bogliolo, DiSBef-STI University of Urbino, Italy
Thomas Michael Bohnert, Zurich University of Applied Sciences, Switzerland
Eugen Borcoci, University "Politehnica" of Bucharest, Romania
Luis Borges Gouveia, University Fernando Pessoa, Portugal
Fernando Boronat Seguí, Universidad Politécnica de Valencia, Spain
Mahmoud Boufaïda, Mentouri University - Constantine, Algeria
Christos Bouras, University of Patras, Greece
Agnieszka Brachman, Institute of Informatics, Silesian University of Technology, Gliwice, Poland
Thierry Brouard, Université François Rabelais de Tours, France
Carlos T. Calafate, Universitat Politècnica de València, Spain
Christian Callegari, University of Pisa, Italy
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Miriam A. M. Capretz, The University of Western Ontario, Canada
Ajay Chakravarthy, University of Southampton IT Innovation Centre, UK
Chin-Chen Chang, Feng Chia University, Taiwan

Ruay-Shiung Chang, National Dong Hwa University, Taiwan
Tzung-Shi Chen, National University of Tainan, Taiwan
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
Philippe Devienne, LIFL - Université Lille 1 - CNRS, France
Kamil Dimililer, Near East University, 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
Emiliano Garcia-Palacios, ECIT Institute at Queens University Belfast - Belfast, UK
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, Netherlands
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
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, Università di Pisa, 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-Tamulipas, 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
Orazio Tomarchio, Università di Catania, Italy
Dominique Vaufreydaz, INRIA and Pierre Mendès-France University, France
Krzysztof Walkowiak, Wrocław University of Technology, Poland
MingXue Wang, Ericsson Ireland Research Lab, Ireland
Wenjing Wang, Blue Coat Systems, Inc., USA
Zhi-Hui Wang, School of Software, Dalian University of Technology, China
Matthias Wieland, Universität Stuttgart, Institute of Architecture of Application Systems (IAAS), Germany
Bernd E. Wolfinger, University of Hamburg, Germany
Chai Kiat Yeo, Nanyang Technological University, Singapore
Abdulrahman Yarali, Murray State University, USA
Mehmet Erkan Yüksel, Istanbul University, Turkey

CONTENTS

pages: 1 - 13

Detecting Users from Website Sessions: A Simulation Study and Results on Multiple Simulation Scenarios

Corné de Ruijt, Vrije Universiteit Amsterdam, Netherlands

Sandjai Bhulai, Vrije Universiteit Amsterdam, Netherlands

pages: 14 - 21

Military REACH: A University-wide Collaboration

Fatemeh Jamshidi, Auburn University, USA

Abhishek Jariwala, Auburn University, USA

Bibhav Bhattara, Auburn University, USA

Katherine Abbate, Auburn University, USA

Daniela Marghitu, Auburn University, USA

Mallory Lucier-Greer, Auburn University, USA

pages: 22 - 35

FracBots: The Next IoT in Oil and Gas Reservoirs

Abdallah ALShehri, Saudi Aramco, Saudi Arabia

Klemens Katterbauer, Saudi Aramco, Saudi Arabia

pages: 36 - 45

A Framework of Web-Based Dark Patterns that can be Detected Manually or Automatically

Ioannis Stavrakakis, Technological University Dublin, Ireland

Andrea Curley, Technological University Dublin, Ireland

Dympna O'Sullivan, Technological University Dublin, Ireland

Damian Gordon, Technological University Dublin, Ireland

Brendan Tierney, Technological University Dublin, Ireland

pages: 46 - 59

Study for In-Vehicle-Network and New V2X Architecture by New IP

Lin Han, Futurewei Technologies, Inc, U.S.A

Lijun Dong, Futurewei Technologies, Inc, U.S.A

Richard Li, Futurewei Technologies, Inc, U.S.A

pages: 60 - 72

A Topic Modeling Framework to Identify Online Social Media Deviance Patterns

Thomas Marcoux, University of Arkansas at Little Rock, United States

Esther Mead, University of Arkansas at Little Rock, United States

Nitin Agarwal, University of Arkansas at Little Rock, United States

pages: 73 - 79

Twitter Search Interface for Looking Back at TV Dramas

Taketoshi Ushiyama, Kyushu University, Japan

Haruka Nagai, Kyushu University, Japan

Detecting Users from Website Sessions: A Simulation Study and Results on Multiple Simulation Scenarios

Corné de Ruijt

Faculty of Science
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
Email: c.a.m.de.ruijt@vu.nl

Sandjai Bhulai

Faculty of Science
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
Email: s.bhulai@vu.nl

Abstract—In this paper, we propose a click simulation model capable of simulating users' interactions with a search engine, in particular in the presence of user censoring. We illustrate the simulation model by applying it to the problem of detecting unique users from the session data of a search engine. In real click datasets, the user initiating the session may be censored, as unique users are often determined by their cookies. Therefore, analyzing this problem using a click simulation model, for which we have an uncensored ground truth, allows for studying the effect of cookie churn itself. Furthermore, it allows for studying how well clustering algorithms perform in detecting clusters of sessions that originate from a single user. To cluster sessions, we present and compare various constrained DBSCAN*-type clustering algorithms. From this comparison, we find that even though the clusters found by the best DBSCAN*-type algorithm did significantly outperform other benchmark clustering methods, it performed considerably worse compared to using the observed cookie clusters. This result remains under different simulation scenarios, though the results do improve when strengthening the user signal. While clustering algorithms may be useful to detect similar users for purposes such as user clustering, cookie tracking remains the preferred method for tracking individual users.

Keywords—Click models; Session clustering; HDBSCAN*

I. INTRODUCTION

This paper is an extension of our previous work on click model simulation and (Internet) session clustering, presented in [1]. In particular, we provide a more detailed description of the click simulation model and (H)DBSCAN* clustering algorithm with a maximum cluster size. Furthermore, we present the performance of the session clustering algorithm on multiple simulation scenarios. The latter was only briefly discussed in our earlier work, presented at the 2020 DATA ANALYTICS conference [1].

The current Internet environment heavily relies on cookies for the enhancement of our Internet browsing experience. These cookies are small pieces of data stored in the browser after being received from a server, along with a requested web page from that server. If the Internet user pushes subsequent requests to the server, the cookie is sent along, allowing the server to recognize the user and adjust its response accordingly. Hence, as cookies allow identifying users over multiple requests, they play a crucial role in session management, the personalization of websites and ads, and user tracking.

However, the usage of multiple devices, multiple browsers, and the focus on cookie management has made the problem of

identifying single users over multiple sessions more complex. One study reports that so much as 20% of all Internet users delete their cookies at least once a week, whereas this percentage increases to 30% when considering cookie churn on a monthly basis [2]. Not being able to track Internet users may lead to sub-optimal behavior of search engines and online ads, as these have less information about previous search and click behavior to infer the user's preference for certain items from. As cookie churn and the usage of multiple devices censor the underlying user who is generating web traffic, we call this user censoring.

Following the 2015 ICDM and 2016 CIKM machine learning challenges [3, 4], cross-device matching has in recent years received considerable scientific attention. Cross-device matching refers to the problem of identifying individual Internet users from a set of Internet logs, where Internet users may have been using multiple devices, and are therefore tracked as separate users. These studies, however, do have some limitations. Most approaches mentioned in the literature are limited to finding pairwise matches, i.e., pairs of sessions that are likely to originate from the same user. Such inference is insufficient if one is interested in identifying exclusive session clusters consisting of more than two sessions.

Furthermore, there seems to be ambiguity in what exactly is meant by cross-device matching, or by session clustering, and to what extent successful methods applied to one problem will also work well on other problems. The ICDM and CIKM competitions consider the problem from the perspective of an online advertiser, advertising on multiple websites. Other approaches (e.g., [2, 5, 6]) consider the problem from the perspective of a single website. At this point, it is unclear whether approaches that work well on a single website are likely to be successful in the online advertisement case, and vice versa. Apart from this multi versus single website perspective, most datasets studied seem to originate from large advertisers or search engines. This raises the question of how generalizable these approaches are for websites or advertisers with less traffic or less heterogeneous searches.

To allow for sensitivity analysis in session clustering, we consider the single website perspective, and propose a single query click simulation model that allows for cookie censoring. Simulation has two main advantages: 1) by adjusting the simulation parameters, we may study how session clustering algorithms perform on websites with different user browsing

characteristics. 2) It provides a ground truth, which, due to user censoring, is only partially observed in real world datasets. Apart from the ground truth being useful in evaluating clustering algorithms, it also allows for studying the effects of user censoring on typical website statistics, such as the number of unique visitors on a website. Although several models have been proposed in the literature that could be used to create a simulation model, they only capture a specific part of search behavior and/or censoring. To our knowledge, this paper is the first to combine these models into one click simulation model with censoring.

Besides introducing the simulation model, we compare several clustering approaches on multiple simulated datasets, where all clustering methods are based on the DBSCAN* and HDBSCAN* algorithms. To measure their effectiveness, we not only consider the error in terms of typical supervised clustering error measures, such the Adjusted Rand Index, but also in terms of the error in estimating overall web statistics. These include the number of unique users, distribution of the number of sessions per user, and the user conversion distribution.

This paper has the following structure. Section II discusses relevant literature related to session clustering. Section III discusses the simulation model, adaptations of (H)DBSCAN*, and experimental set-up. Section IV discusses the results, whereas Section V discusses the implications and ideas for further research.

II. RELATED WORK

A. Click simulation

Simulating click behavior is definitely not a new concept. Chuklin et al. [7, pp. 75-77] suggests using pre-fitted click models for this purpose, where the model is pre-fitted to public click datasets. One risk of using pre-fitted models is an availability bias: can the characteristics of public click datasets, commonly provided by large search engines, easily be generalized over all search engines? Also, these datasets do not always provide the type of information one is interested in, such as the device used to initiate a session.

Fleder and Hosanagar [8] provide a generative approach for modeling user preferences, which we will discuss in more depth in Section III-A. This model can be used as an alternative to model users' preferences for clicking on items. Using pre-fitted or generative models do have a trade-off in terms of accuracy vs interpretability. E.g., the former may have an accurate estimate of users' item preferences, but it provides little understanding of why this preference over different products has a certain shape, whereas for the latter, we expect this to be vice versa.

Several authors have studied how cookie censoring occurs. E.g., [2, 9, 10] consider cookie churn, whereas [11] considers specifically cross-device behavior. Results from these studies can be used to model cookie churn dynamics in a simulation model.

B. Identifying unique users from sessions

Identifying unique users from sessions can be seen as a specific case of the entity/identity resolution problem [6]. Though, what makes this problem special is the nature of the dataset. This often consists of a large number of sessions, of

which clicks and web page meta-data (such as the URL) are the main sources of information. Because of these characteristics, entity resolution algorithms that do not account for these characteristics are likely to fail in their objective.

Session matching can be applied from an online advertiser's perspective, as was the case during the 2015 ICDM and 2016 CIKM machine learning challenges [12, 13, 14, 15, 16, 17, 18, 19, 20, 21], or from the perspective of a single website [2, 5, 6]. What remains unclear is whether these two problems can be considered the same. Although in both cases the main motivation for cookie matching may be the same, e.g., increasing the click-through rate by means of personalization, the type of data is bound to be different. When advertising on multiple websites, the data seems to consist for a substantial part out of a large variety of visited URLs. Hence, proposed approaches from the advertisement perspective tend to rely heavily on natural language processing techniques [15, 16, 18, 19, 21, 22]. In case of a single website, the URLs or web pages' meta-data may be less diverse, and the "unique fingerprints" [23] users create while browsing a single website may therefore be less distinctive than on multiple websites.

Most often, both the single and multiple website perspectives are modeled as a binary classification problem. Here, a model is trained to identify whether two feature vectors describing sessions a and b originate from the same user. Striking is the success of tree-boosting methods for this task, which also in both the 2015 ICDM and 2016 CIKM machine learning competitions showed promising results. For a more in-depth discussion of the different methods applied in cross-device matching, modeled as a binary classification problem, we refer to Karakaya et al. [22]. Also worth mentioning is that many methods proposed to both the 2015 ICDM and 2016 CIKM competitions allow for overlapping user clusters. As the objective is to find pairs of sessions likely to originate from the same user, this may result in sessions a , b , c to be classified as $f(a, b) = 1$ and $f(a, c) = 1$, but $f(b, c) = 0$, f being the same user classifier. Such result may be undesirable in some practical applications.

A slight generalization of the cross-device matching problem is the cookie matching problem. Here we are given a set of sessions that are already partially clustered into users via cookies, but only partially due to some form of user censoring. I.e., cross-device matching and cookie matching only seem to differ on whether one assumes that user censoring only occurs because of cross-device usage, or also because of cookie churn. However, many approaches proposed in the literature can be applied to both problems. Hence, in these formulations, this distinction seems irrelevant. Various authors have considered the cookie matching problem, though under different names such as: 'user stitching' [6], 'visitor stitching' [5], or 'automatic identity linkage' [24]. Like in cross-device matching, these studies tend to allow for overlapping clusters.

One approach that does not allow for overlapping clusters is considered by [12], using classical bipartite matching algorithms such as the Hungarian algorithm. However, it is questionable to what extent these approaches are scalable, as the paper works with relatively small datasets. Furthermore, as users might have more than two cookies, bipartite matching will only solve a part of the problem.

Dasgupta et al. [2] also move beyond pairwise clustering. The authors consider a combination of several similarity measures to determine whether two cookies originate from the same user, and apply a greedy graph coloring algorithm to cluster a session graph into user clusters. However, since multi-device usage as we observe on websites now was not that much the case when the paper was published in 2012, the algorithm strongly relies on the assumption that only one device is used at a time. This allowed the authors to only consider non-overlapping cookies in terms of time as candidates for cookie matching, whereas in the multi-device case, such a constraint would not be able to identify unique users simultaneously using multiple devices.

In this paper, we will use the term session clustering to relate to the problem of identifying unique users from session data. We prefer this term, as our methods do not per se require having partial session clusters from cookies, something that would be the case in cookie matching. Furthermore, we seek non-overlapping clusters, whereas ‘matching’ relates to training a classifier to predict whether two sessions originate from the same user. However, still many of the methods discussed so far are applicable to this formulation of the problem.

We take a similar approach as [19] towards session clustering. This approach first trains a classifier that predicts whether sessions a and b originate from the same user (that is, share the same cookie in the data). Next, each session forms pairs with its K nearest neighbor (K -NN) sessions, after which each nearest neighbor is re-evaluated using the classifier on whether the session and neighbor indeed originate from the same user. All sessions included in the remaining pairs are subsequently clustered using a greedy clustering algorithm, from which all sessions in a cluster are also added to the set of session pairs.

This method shows some similarity with DBSCAN [25], where also K -NN is used to quickly identify similar data points. However, DBSCAN computes a (possibly approximate) minimum spanning tree (MST), from which a quick approximation can be made of the distances between points. Compared to the greedy clustering approach by [19], this leads to a considerable speed up. On the other hand, as DBSCAN misses a constraint on the maximum cluster size, we will turn to two of DBSCAN’s descendants: DBSCAN* and HDBSCAN* [26, 27], which can quite easily be adjusted to incorporate a maximum cluster constraint.

III. METHODS

A. Simulating click data with cookie-churn

We consider a simulation model that models how users behave when interacting with a search engine. We choose to simulate behavior on a search engine, and not behavior on other types of websites, as there is extensive literature on what type of parametric models are accurate for modeling user behavior on search engines [7]. Furthermore, apart from dedicated search engines, a search tool is also a common feature on websites having other purposes [28]. Hence, we believe it is likely that this behavior is also found elsewhere.

To avoid overcomplexifying the simulation model, we only consider the case in which users push one or multiple homogeneous queries to the search engine. I.e., the query itself is the same over all users, and one user may repeat this query

a number of times. Users do have different item preferences for the items the search engine may return. Furthermore, the item order may be different in each Search Engine Result Page (SERP). The simulation model consists of three parts. The first part models how users navigate through the SERP, the second part models how users’ utility function is determined, while the third part models how the session generating user is censored due to cookie churn or the usage of multiple devices. For reference, Table VI provides an overview of the most important variables in the simulation model.

1) *Simulating SERP interactions*: Two types of interactions between a user and the search engine are considered. First, users may push the (homogeneous) query to the server, and receive the SERP in response. Second, users may click on results in the SERP. At each interaction, the server checks whether the user has an active cookie. If not, a new cookie is sent along with the server’s response (that is, either the SERP, or the content page of a particular item in the SERP), and stored in the user’s browser. We will discuss how cookie churn is modeled in Section III-A2.

All interactions are stored by the server, which provides a label for the cookie, device and query-session. This query-session is defined in terms of a set of interactions with one SERP. Hence, where in practice a browser session is typically defined by some period of interaction, we deliberately choose to model a session as a set of interactions with one SERP, irrespective of the time between two interactions with this SERP.

To simulate clicks on a search engine, we employ the Simplified Chapelle-Zhang Model (SCZM) [29]. Although this model is known in the literature as the Simplified Dynamic Bayesian Network model (SDBN), we renamed the model as it is only a specific case of a Dynamic Bayesian Network. We choose to use SCZM for two reasons: 1) the model, though simple, seems to perform reasonably well in comparison with other parametric click models when predicting clicks [7]. 2) SCZM captures the ordering effect of items in the SERP. I.e., users may not always reflect their preferences correctly in their clicks, as their behavior is also determined by how items are ordered. Including this ‘cascade effect’ provides more realistic results.

To describe the simulation model, the following notation will be used. Let $i \in \{1, \dots, n\}$ be a query-session, which produces a SERP of unique items $S_i \subseteq \mathcal{V}$, with $\mathcal{V} = \{1, \dots, V\}$ the set of all items, indexed by v . We assume all SERPs $1, \dots, n$ to have the same number of items T . Let $u_i \in \mathcal{U}$ denote the user initiating query-session i , with $\mathcal{U} = \{1, \dots, U\}$ the set of all users. The user index u is used instead of u_i in case the precise query-session i is irrelevant. $r_i(t)$ denotes the item at position t in query-session i . Likewise, $r_i^{-1}(v)$ gives the position of item v in query-session i , and r_i^{\max} denotes the largest position of a clicked item in S_i , where $r_i^{\max} = 0$ if no items were clicked during query-session i .

SCZM considers three latent variables: $R_v^{(i)}$ denotes whether user u_i is attracted to item v during query-session i . This variable is also known as the relevance of item v for the user initiating session i . The probability of item v being relevant to user u in session i is given by $\phi_{u,v}^{(R)}$. $S_v^{(i)}$ denotes whether user u_i is satisfied with item v after having clicked the item, which happens with probability $\phi_{u,v}^{(S)}$, and $E_t^{(i)}$ denotes

whether user u_i will evaluate the item in position t during query-session i . Whether the item at position t in SERP i is clicked is denoted by the binary variable $y_t^{(i)}$.

The model follows the cascade hypothesis, that is, it assumes a user always evaluates the first item ($E_1^{(i)} = 1$ for all $i = 1, \dots, n$), after which the user decides to evaluate subsequent items in the list according to the perceived attraction and satisfaction of the previous evaluated items in the list, according to

$$E_1^{(i)} = 1; \quad (1)$$

$$\mathbb{P}(R_v^{(i)} = 1) = \begin{cases} \phi_{u_i, v}^{(R)} & \text{if } v \in \mathcal{S}_i \\ 0 & \text{otherwise} \end{cases}; \quad (2)$$

$$\mathbb{P}(S_v^{(i)} = 1 | y_{r_i^{-1}(v)}^{(i)} = 1) = \begin{cases} \phi_{u_i, v}^{(S)} & \text{if } v \in \mathcal{S}_i \\ 0 & \text{otherwise} \end{cases}; \quad (3)$$

$$y_t^{(i)} = 0 \Rightarrow S_{r_i(t)}^{(i)} = 0; \quad (4)$$

$$E_{t-1}^{(i)} = 1, S_{r_i(t-1)}^{(i)} = 0 \iff E_t^{(i)} = 1, \quad t > 1; \quad (5)$$

$$y_t^{(i)} = 1 \iff E_t^{(i)} = 1, R_{r_i(t)}^{(i)} = 1. \quad (6)$$

To come up with reasonable values for $\phi_{u, v}^{(R)}$ and $\phi_{u, v}^{(S)}$, we used the same approach as in [8]. That is, users are represented by the vectors $\eta_u = (\eta_1^{(u)}, \eta_2^{(u)})$, $u \in \mathcal{U}$, where $\eta_1^{(u)}$ and $\eta_2^{(u)}$ are drawn from two independent standard normal distributions. Likewise, all items can be represented by the vectors $\psi_v = (\psi_1^{(v)}, \psi_2^{(v)})$, where again $\psi_1^{(v)}$ and $\psi_2^{(v)}$ are drawn from independent standard normal distributions. The probabilities $\phi_{u, v}^{(R)}$, and $\phi_{u, v}^{(S)}$ are then determined by the multinomial logits

$$\phi_{u, v}^{(R)} = \frac{e^{\omega_{u, v} + \nu^{(A)}}}{\sum_{v' \in \mathcal{V} \setminus \{v\}} e^{\omega_{u, v'} + \nu^{(A)}} + e^{\omega_{u, v} + \nu^{(A)}}}, \quad (7)$$

$$\phi_{u, v}^{(S)} = \frac{e^{\omega_{u, v} + \nu^{(S)}}}{\sum_{v' \in \mathcal{V} \setminus \{v\}} e^{\omega_{u, v'} + \nu^{(S)}} + e^{\omega_{u, v} + \nu^{(S)}}}, \quad (8)$$

with

$$\omega_{u, v} = -q \log \delta(\eta_u, \psi_v). \quad (9)$$

Here δ is some distance function, in our case Euclidean distance. $q \in \mathbb{R}^+$ is some constant value that models the users' preferences towards nearby items, and $\nu^{(A)}$, $\nu^{(S)}$ are salience parameters for attraction and satisfaction respectively.

The order in which items are presented is determined as follows. First, during a warm-up phase, we simulate clicks for $U_{\text{warm-up}}$ users, while randomly ordering the items such that all have equal probability of being positioned at positions $t = 1, \dots, T$. Next, we estimate the overall probability of each item being found attractive, and we use these probabilities as weights to determine the item order for subsequent query-sessions. More specifically, for each query-session i , we draw items \mathcal{S}_i from a multinomial distribution with parameters $\hat{\phi}_v / \sum_{v \in \mathcal{V}} \hat{\phi}_v$, $v = 1, \dots, V$; without replacement. The estimate of overall attraction is given by [7, p. 26],

$$\hat{\phi}_v = \frac{1}{|\mathcal{I}_v|} \sum_{i \in \mathcal{I}_v} y_{r_i^{-1}(v)}^{(i)}, \quad (10)$$

with

$$\mathcal{I}_u = \{\mathcal{S}_i : v \in \mathcal{S}_i, r_i^{-1}(v) \leq r_i^{\max}\}. \quad (11)$$

To avoid $\hat{\phi}_v$ to be (close to) zero, we impose a minimum probability of 10^{-5} for all $v \in \mathcal{V}$.

2) *Cookie censoring*: Cookie censoring is incorporated in the simulation model in two ways: by incorporating time and letting cookies churn after some random time \bar{T} , and by switching from device d to some other device d' . First, we consider the cookie lifetime $\mathcal{T}_{u, o, d}^{\text{cookie}}$ for the o -th cookie of user u on device d , and the user lifetime $\mathcal{T}_u^{\text{user}}$. Whenever the cookie lifetime of cookie o ends, but the current user lifetime is strictly smaller than $\mathcal{T}_u^{\text{user}}$, a new cookie o' is created, which lifetime is drawn from the cookie lifetime distribution F^{cookie} . For a period of $\mathcal{T}_{u, o', d}^{\text{cookie}}$, all click behavior of user u on device d will now be registered under cookie o' .

Second, after each query-session a user may switch from device d to d' , which happens according to transition matrix P . Whenever a user switches devices, we consider whether the user has used this device before. If not, a new cookie o' is created, and we draw a new cookie lifetime from F^{cookie} . However, the cookie lifetime $\mathcal{T}_{u, o, d}^{\text{cookie}}$ does not end prematurely when the user switches from device d to d' . If later on the user switches back to device d while the cookie lifetime $\mathcal{T}_{u, o, d}^{\text{cookie}}$ has not ended, the behavior of user u is again tracked via cookie o until another device switch occurs or cookie o churns.

Putting this censoring into practise requires us to provide five distributions: 1) a distribution F^{abs} for the time between query-sessions, which following [10] we will refer to as the *absence time*, 2) a distribution for the cookie lifetime (F^{cookie}), 3) a distribution for the user lifetime (F^{user}), 4) the device transition matrix P , and 5) the initial device probability F^{device} .

For the absence time distribution, we use some results from [10]. Although Dupret and Lalmas [10] fitted a Cox survival model to user absence data in order to estimate user lifetimes, we refitted the data mentioned in the paper with a different model for two reasons. First, there is ambiguity in the method used to model absence time. The authors fit a Cox survival model with one covariate. As the (log-)likelihood of a Cox survival model omits the estimation of the base hazard, the method for estimating this base hazard should be provided (e.g., the Breslow estimator). However, the paper does not report which method was used to fit the baseline hazard. Second, results from Dasgupta et al. [2] on cookie churn suggests that, when taking into account longer periods than 7 days, absence time has a fat-tailed distribution. We found that a Pareto-I with scale parameter $m = 1$ and shape $\alpha = 0.11$ seems to fit the data from [10] approximately well. This distribution was therefore used to model F^{abs} . To allow for absence times smaller than 1 (but still positive), we subtracted one from all drawn lifetimes.

To model the cookie lifetime, we used the results from [2], who find that a hyper-exponential distribution with one over the rate being equal to 50 seconds (with probability .06), 25 minutes (with probability .07), 14 hours (with probability .07), 15 days (with probability .18), and 337 days (with probability .62), fits reasonably well. Here, cookie lifetime is defined as the time difference between the first and last observed action from a single cookie. We consider time at a minute scale,

and therefore rounded up the first phase (50 seconds) to one minute.

The user lifetime is obtained by sampling from N cookie lifetime distributions, where N itself is drawn from a geometric distribution with parameter ρ . As the cookie lifetime distribution is modelled as a hyper-exponential, we will refer to this distribution as a repeated hyper-exponential distribution. Although we sample from the cookie life time distribution, the user lifetime is independent from the cookie lifetimes: they only share the underlying hyper-exponential distribution, not the realizations of that distribution.

To model device transition matrix P , we use the results from [11], who study device transitions between four devices: a PC, tablet, smartphone and game console. We adopted the transition probabilities found in this paper, where we dropped the game console as the found transition probabilities from and to this device were negligible. After dropping the game console, the probabilities were normalized to obtain transition matrix P . The initial device probability distribution F^{device} is also obtained using the results from [11], and is modeled as a multinomial distribution with parameter $\pi = (\pi_1, \pi_2, \pi_3)$; π_1, π_2, π_3 being the probability of the PC (Dev. 1), tablet (Dev. 2), and smartphone (Dev. 3) being the first device respectively. The normalized initial and transition probabilities from [11] are given by Table I.

TABLE I
INITIAL DEVICE AND DEVICE TRANSITION PROBABILITIES ADOPTED FROM [11]

	π	Dev. 1	Dev. 2	Dev. 3
Dev. 1	.64	.9874	.0042	.0084
Dev. 2	.11	.00256	.9697	.0046
Dev. 3	.25	.029	.0018	.9773

3) *Summary of the simulation procedure*: The entire simulation procedure is given in Algorithms 1 and 2 (see Appendix). The former describes how user preferences are obtained and how the overall popularity is determined, whereas the latter describes how clicks and cookie churn are simulated over a set of users.

For convenience, we have written the set of warm-up users as $\mathcal{U}_{\text{warm-up}}$, $\hat{\phi} = (\hat{\phi}_1, \dots, \hat{\phi}_V)$, and $\mathbf{y}_i = (y_1^{(i)}, \dots, y_T^{(i)})$. The location and scale parameter of the Pareto-I distribution are written as m and α , whereas the rate and rate probability of the hyper-exponential distribution are given by the vectors λ and \mathbf{p} . Last, let I_d be a 3×3 matrix where the d -th column contains all ones, whereas the rest of the matrix contains all zeros.

The simulation iterates over all users, where for each user new query-sessions are simulated until the user lifetime has elapsed. For each user, first the initial device is drawn, along with a cookie lifetime for that user on that device, and the total user lifetime. Next, query-sessions are simulated for each user in four steps. First, \mathcal{S}_i is (iteratively) drawn using the overall item popularity $\hat{\phi}$, and we simulate clicks using the SCZM model described in Section III-A1, which are stored in dataset \mathcal{D} . Second, we simulate the time until the next session. Third, the device of the next session is determined. Fourth, we check whether the last cookie on the new device has churned.

Algorithm 1: User simulation procedure

- 1 Draw $\eta_1^{(u)}, \eta_2^{(u)}, \psi_1^{(v)}, \psi_2^{(v)}$ i.i.d. from a standard normal distribution for all $v \in \mathcal{V}$ and $u \in \mathcal{U}$;
 - 2 Compute similarities $\omega_{u,v}$ according to (9);
 - 3 Compute the probability of attraction and satisfaction, using (7);
 - 4 Set $\hat{\phi}_v \leftarrow 1$ for all $v \in \mathcal{V}$;
 - 5 $\mathcal{D}_{\text{warm-up}} \leftarrow \text{SIMULATE_CLICKS}(\mathcal{U}_{\text{warm-up}})$;
 - 6 Recompute $\hat{\phi}$ according to (10);
 - 7 $\mathcal{D} \leftarrow \text{SIMULATE_CLICKS}(\mathcal{U} \setminus \mathcal{U}_{\text{warm-up}})$;
 - 8 **return** \mathcal{D} ;
-

If so, a new cookie is created with a corresponding new cookie lifetime.

Although Algorithm 2 assumes all users arrive at $t = 0$, we shift all times after the simulation to obtain click behavior spread out over time. Here we assume a Poisson arrival process with rate γ . I.e., the first query-session of user u starts some exponentially distributed time after the initial query-session of user $u - 1$. Note that these inter-first session times only depend on the time of the first session of the previous user, not on any other subsequent behavior of that user.

B. Session clustering

1) (H)DBSCAN*:

a) *Hierarchical clustering using Minimum Spanning Trees (MST)*: Before we discuss the adjustment made to the HDBSCAN* and DBSCAN* algorithms, we will first briefly describe the two algorithms. We first discuss the overlapping part in both algorithms, after which we discuss their differences. Following the terminology by [26, 27] and [30], let $X = \{X_1, \dots, X_n\}$ be a set of data points, let $\kappa_k(X_i)$ be the distance from point X_i to its k -th nearest neighbor (for some given value of $k \in \mathbb{N}$), and let $\delta(X_i, X_{i'})$ be some distance measure between points X_i and $X_{i'}$. Based on this original distance measure, DBSCAN* considers an alternative distance measure, which is named the *mutual reachability distance*, and is defined as follows:

$$\delta_k^{\text{mreach}}(X_i, X_{i'}) = \begin{cases} \max\{\kappa_k(X_i), \kappa_k(X_{i'}), \delta(X_i, X_{i'})\} & X_i \neq X_{i'} \\ 0 & X_i = X_{i'} \end{cases} \quad (12)$$

Although DBSCAN* does not specify the exact distance measure δ , we will (like in Section III-A1) assume this is Euclidean distance. The main motivation for introducing this mutual reachability distance is to better identify different clusters with high density of arbitrary shape, as the measure tends to push different high density clusters further apart.

Given the mutual reachability distance, (H)DBSCAN* represents each data point as a node in a complete weighted graph G , where the weights are simply the mutual reachability distances between data pairs. Using G , the algorithm first computes a minimum spanning tree (MST), which allows for fast identification of clusters. The MST is also used to approximate distances: the distance between two non-adjacent points X_i and $X_{i'}$ in the MST can be approximated by the path length $X_i \rightarrow X_{i'}$ in the MST. At the same time, this

distance is a lower bound on the actual distance: otherwise, $X_i \rightarrow X_{i'}$ would be adjacent in the MST.

From this MST, one can build a dendrogram of the data points in an agglomerative manner. First, (H)DBSCAN* assigns each data point X_1, \dots, X_n to separate clusters $\mathcal{B}_1^0, \dots, \mathcal{B}_n^0$. Here the superscript is used to indicate the hierarchy level of the cluster, which at this stage is zero. Second, it iterates through the edges in G , increasing in terms of their weights. For some edge (i, i') having the smallest edge weight, it finds the two clusters with the highest hierarchy levels h_i^{\max} and $h_{i'}^{\max}$, to which i and i' are assigned to respectively. Next, it and creates a new cluster $\mathcal{B}_j^{\max\{h_i^{\max}, h_{i'}^{\max}\}+1}$, which includes all data points included in the highest hierarchy clusters to which X_i and $X_{i'}$ were previously assigned to. If this process is repeated for all edges in G , the last edge will create a cluster containing all data, which occurs at level H .

b) DBSCAN:* The construction of the dendrogram occurs both in DBSCAN* and HDBSCAN* in the same manner. However, as both methods wish to find non-overlapping clusters, the two methods split ways from there. In DBSCAN*, one would take some value $\epsilon \in \mathbb{R}^+$, and remove all cluster merges in the dendrogram that were merged with a weight strictly greater than the chosen maximum distance ϵ . This would lead to a set of disconnected binary trees \mathcal{T} , and a set of singleton points \mathcal{N} . The singleton points are points for which their k -th nearest neighbor is already at a further distance than ϵ , and these points are consequently labeled as noise. All data points in one tree $\tau \in \mathcal{T}$ are labeled as one cluster.

c) HDBSCAN:* The underlying assumption of cutting the dendrogram at level ϵ , is that all clusters have (approximately) the same density. This density is in HDBSCAN* approximated by $\theta = 1/\epsilon$, i.e., close points imply high density. HDBSCAN* allows for different cut-off levels of ϵ , or similarly of θ , where the optimal cut-off level for some cluster is determined via the notion of *relative excess of mass*, which we will introduce in a moment.

More precisely, let M be some given minimum cluster size. To somewhat simplify notation, we let index j refer to any cluster, irrespectively of hierarchy h , such that h can be dropped. HDBSCAN* first creates a *condensed tree* from the dendrogram in the following way. It starts at the root of the dendrogram, having label j_0 , and considers its children. These were merged at some density $\theta_{j,j'}$, merging two clusters with labels j and j' . It then considers three options: 1) if both children have less than M points, all points in \mathcal{B}_j and $\mathcal{B}_{j'}$ “fall-out” of the cluster at density $\theta_{j,j'}$, implying that for densities greater than $\theta_{j,j'}$ all points in \mathcal{B}_j and $\mathcal{B}_{j'}$ are labeled as noise. 2) If only one cluster \mathcal{B}_j has less than M points, all points in \mathcal{B}_j fall-out at density $\theta_{j,j'}$, while the parent cluster label (j_0) is now continued for all observations in $\mathcal{B}_{j'}$. I.e., we replace label j' by j_0 , and as a result the exact cluster j_0 now refers to depends on whether we pick a density larger or smaller than $\theta_{j,j'}$. 3) If both children have more than M observations, clusters \mathcal{B}_j and $\mathcal{B}_{j'}$ keep their labels j and j' . I.e., label j_0 is not continued, and clusters \mathcal{B}_j and $\mathcal{B}_{j'}$ are considered separate clusters for densities larger than $\theta_{j,j'}$. After both children have been relabeled, this process is repeated using these new labels until all nodes have been relabeled.

The resulting condensed tree is essentially still the same

as the original dendrogram, but with different labels. I.e., by continuing the parent (option 2), some labels now may refer to different clusters, dependent on density θ . Let $\{1, \dots, m\}$ be the resulting set of labels from relabeling. For each label $j \in \{1, \dots, m\}$, let \mathcal{B}_j be the set of observations labeled j at the minimum density for which j exists. Furthermore, let $\theta_j^{\max}(X_i)$ and $\theta_j^{\min}(X_i)$ be the densities at which observation X_i falls off cluster j and the density at which X_i first occurs in cluster j respectively. Note that $\theta_j^{\min}(X_i)$ is either zero (when j is the label continued from the root node), or the density at which cluster j splits off from its parent, hence it has the same value for all $X_i \in \mathcal{B}_j$.

Clusters $\{\mathcal{B}_1, \dots, \mathcal{B}_m\}$ may still be overlapping. To find non-overlapping clusters, HDBSCAN* introduces the *relative excess of mass* of cluster j as $\sigma(j)$, which is defined by:

$$\sigma(j) = \sum_{X_i \in \mathcal{B}_j} [\theta_j^{\max}(X_i) - \theta_j^{\min}(X_i)]. \quad (13)$$

The relative excess of mass has an intuitive argument for clustering. Large values for $\sigma(j)$ imply that when increasing the density, the cluster remains more or less intact (apart from some noise points splitting off at higher densities). As a result $\theta_j^{\max}(X_i) - \theta_j^{\min}(X_i)$ becomes large. I.e., the relative excess of mass can be used as a measure of cluster quality. Hence, HDBSCAN* optimizes the sum of relative excess of mass over a subset of clusters $\{\mathcal{B}_1, \dots, \mathcal{B}_m\}$ such that this subset is non-overlapping.

2) Introducing maximum cluster sizes to HDBSCAN and DBSCAN*:* To return to the problem at hand: identifying small session clusters from the set of all sessions that may be originating from the same user, HDBSCAN* and DBSCAN* can obviously be used for this purpose. Apart from the earlier discussed benefit of speed by clustering via MST, incorporating noise points would also intuitively make sense in identifying potential users from sessions: we would expect that quite a large (though unknown) percentage of all sessions might still be from users only initiating a single session.

By tweaking parameters k , (the k -th nearest neighbor in nearest neighbor distance κ_k), ϵ (dendrogram cut-off point in case of DBSCAN*), and M (minimum number of points before a cluster is considered noise in HDBSCAN*) one can obtain session clusters that obey a maximum cluster size $\beta \in \mathbb{N}$. However, some early experiments with DBSCAN* and HDBSCAN* showed that the resulting clusters tended to either very large clusters, or labeled (almost) every point as noise. For that reason, we chose to adjust both algorithms, in order to obtain more small clusters having a size smaller than β .

To impose the clusters to be more fine grained, we impose a restriction on the maximum cluster size of the clusters found by (H)DBSCAN*. We do so in three different ways: max-size DBSCAN* (MS-DBSCAN*) imposes this restriction on DBSCAN*, whereas MS-HDBSCAN*⁻ and MS-HDBSCAN*⁺ are two ways to impose the restriction on HDBSCAN*.

First we consider MS-DBSCAN*. This algorithm is only a slight adaptation to the DBSCAN* algorithm described in Section III-B1. Given the binary trees \mathcal{T} , obtained by removing all nodes and edges in the dendrogram above distance ϵ , we further remove all cluster nodes j for which $|\mathcal{B}_j| > \beta$. Doing so results in two new sets: $\tilde{\mathcal{N}}$ and $\tilde{\mathcal{T}}$, again representing singleton

points that we assume to be noise, and all points in a tree $\tau \in \tilde{\mathcal{T}}$ receive the same cluster.

Second are the adaptations of HDBSCAN*. The first steps of these two adaptations are the same. First, all clusters $\mathcal{B}_j \in \{\mathcal{B}_1, \dots, \mathcal{B}_m\}$ having $|\mathcal{B}_j| > \beta$ are removed from the dendrogram. This, like in DBSCAN*, gives two sets: noise points \mathcal{N} and trees \mathcal{T} . Second, for each sub-tree $\tau \in \mathcal{T}$, we again optimize the total relative excess of mass subject to non-overlapping clusters. The difference between MS-HDBSCAN*⁻ and MS-HDBSCAN*⁺ arises when a leaf node of the condensed tree (that is, a label that does not split at some larger density into two new labels, though noise points may split off) of some condensed sub-tree τ is in the set of optimal non-overlapping clusters. In case of MS-HDBSCAN*⁻, all observations in \mathcal{B}_j are given the same label, whereas in case of MS-HDBSCAN*⁺, these are considered noise.

3) *Session cluster re-evaluation*: As one might have noticed, so far we have not used any information from the cookies. I.e., knowing which sessions have the same cookie could provide valuable information about the underlying user. In particular, we wish to train a model that can function as an alternative to standard distance measures δ , such as Euclidean or Manhattan distance, which we then again can plug into the adapted (H)DBSCAN* algorithms described in Section III-B2.

Obtaining session clusters with re-evaluation is done as follows. Assume we have a trained classifier $\hat{f}(X_i, X_{i'})$, which returns the probability of X_i and $X_{i'}$ originating from the same user. First, like in [19], we find for each point X_i the K nearest neighbors, which gives us a set \mathcal{X} of all nearest neighbor session pairs. Second, we compute $-\log(\hat{f}(X_i, X_{i'}))$ for all $(X_i, X_{i'}) \in \mathcal{X}$, and fill this into a (sparse) $n \times n$ distance matrix W . For all pairs $(X_i, X_{i'}) \notin \mathcal{X}$, we assume the distance is some large value δ_{\max} , which allows us to store W efficiently, and greatly speeds-up computations compared to evaluating all pairwise same user probabilities. Distance matrix W can subsequently be used as distance measure δ in the algorithms discussed in Section III-B to obtain new session clusters.

To train classifier \hat{f} , we first cluster a training set according to one of the models discussed in Section III-B, using Euclidean distance for δ . Second, for each cluster we add all unique session pairs into some training set $\mathcal{X}_{\text{clust}}$. Next, we start using the observed cookies: we treat each cookie as a cluster and determine all session pairs in these cookie clusters, where this set of pairs is denoted by $\mathcal{X}_{\text{cookie}}$. To determine for each session pair $(X_i, X_{i'})$ the correct label, we use the information from the observed cookie. If X_i and $X_{i'}$ have the same observed cookie, we set the target variable to one, whereas it equals zero otherwise. The final training set $\mathcal{X}_{\text{train}}$ is obtained by undersampling from $\mathcal{X}_{\text{clust}} \cup \mathcal{X}_{\text{cookie}}$.

Note that obtaining negative labeled training pairs from a point its K nearest neighbors follows the assumption that these are indeed more likely to be negatives than positives. If this assumption holds, sampling negatives from the nearest neighbors would intuitively help the classifier to learn more subtle patterns. I.e., the K nearest neighbors are close in terms of the common distance measure, but not according to the classifier.

4) *DBSCAN* with random clusters*: To benchmark the clustering approaches just discussed, we consider the following

benchmark. We first cluster the sessions using the ordinary DBSCAN* algorithm, in which way we obtain initial clusters $\mathcal{B}_1^0, \dots, \mathcal{B}_m^0$. Next, for each cluster \mathcal{B}_j^h ($h \in \mathbb{N}$, with initially $h = 0$), if $|\mathcal{B}_j^h| > \beta$, we iteratively select $\min\{s_{j,h}, |\mathcal{B}_j^h|, \beta\}$ points uniformly at random from \mathcal{B}_j^h to form a new cluster $\tilde{\mathcal{B}}$, and update $\mathcal{B}_j^{h+1} \leftarrow \mathcal{B}_j^h \setminus \tilde{\mathcal{B}}$. Here, $s_{j,h}$ is drawn from a geometric distribution with $p = 0.5$. This process continues until for all $j \in \{1, \dots, m\}$: $|\mathcal{B}_j^h| \leq \beta$ for some h , at which the remaining points in \mathcal{B}_j^h are labeled as one cluster.

Intuitively, we selected this benchmark as it captures the higher level hierarchy clustering of DBSCAN*, but not the low level clusters (as these clusters are picked at random). Therefore, comparing the previous methods with this random clustering approach allows us to assess whether the smaller size clusters reveal more information than the larger ones.

C. Experimental setup

1) *Simulation parameters*: Our experimental design consist of two steps. First, we consider a simulation base case on which we evaluate the clustering approaches discussed in Section III-B. In this base case, users' first query arrival follows a Poisson process with rate $\gamma = 0.2$ (minutes), after which subsequent behavior over time of a particular user is modeled according to F^{abs} , F^{cookie} , F^{user} , F^{device} , P , and π , of which the parameters were already given in Section III-A2. We used $U = 20,000$ users with $U_{\text{warm-up}} = 2,000$ (10%). Furthermore, we removed the first 250 sessions (not part of the first $U_{\text{warm-up}}$ users, who were only used to estimate the overall item popularity), as these would likely all be first sessions from new arriving users, and therefore including them may lead to a bias in the data. Likewise, we removed all observations after 43,200 minutes (30 days) to avoid the opposite bias: not having any new users. Users could pick from $V = 100$ items, and we choose as maximum list size $T = 10$.

For parameters that could not be adopted from the literature, we tried several parameter values and looked at three characteristics. First, we considered whether the click probability is decreasing in list position. Second, whether the attraction/satisfaction is centered around 0.5, with a standard deviation of approximately 0.1 to 0.2. Third, whether all sessions are somewhat spread out over time. This lead us to choosing users' preference for nearby items $q = 1$, user lifetime phases geometric parameter $\rho = 0.5$, and salience parameters $\nu^{(A)} = \nu^{(S)} = 5$. Figure 1 shows the three base case characteristics for the resulting simulated base case used in further inference. In the second step of the experimental design, we made adjustments to the latter parameters, that is, those not adapted from the literature. These adjustments will be discussed in Section IV-B.

2) *Features and MS-(H)DBSCAN* hyper-parameter settings*: The simulated dataset was split into a training and test set according to a 70/30 split over the users. I.e., users always are entirely in the training set, or entirely in the test set. For each session, we used the session's start time, observed session count (as observed by the cookie), number of clicks, and whether the session's SERP has at least one click as features. Furthermore, to obtain a vector representation of the items and interactions with the SERP, we first computed a bin-count table. This table contains per item the total number of clicks, skips (no click), and the log-odds ratio between clicks and



Figure 1. Summary of base case simulation.

skips over 30 percent of all sessions, which combined were used as item vector representations.

Next, for each session i , we concatenated all item vectors ψ_v , $v \in \mathcal{S}_i$, in order of their position, resulting in some vector \mathbf{a}_i with $3T$ elements. Additionally, we created four more session vectors. The first of these vectors is obtained by multiplying \mathbf{a}_i with a vector containing ones at those positions where a click occurred, whereas for the second vector, \mathbf{a}_i is multiplied with a vector containing ones at positions where the item was skipped (=not clicked). The third vector is obtained by multiplying \mathbf{a}_i with a vector containing ones at the last clicked position. To obtain the fourth vector, \mathbf{a}_i is multiplied with a vector of list positions for each item. In all cases, the vector multiplication is element-wise. Next, all five session vectors were concatenated to obtain one session vector representation.

The resulting concatenated session vector was further treated by computing all second order polynomial features, after which we normalized and applied the Yeo-Johnson [31] power scaler to make the distribution of each feature more Gaussian-like. We reduced the vector's dimension using a principle component analysis using seven principle components, the latter was chosen using the elbow method.

For each method, we experimented with $k \in \{1, 3, 5\}$ (here k as in κ_k , the distance to the k -th nearest neighbor). For DBSCAN*-like algorithms, we tried

$$\epsilon \in \left\{ \left(q_{\max} (q_{\min}/q_{\max})^{\ell/N} \right) \mid \ell \in \{1, \dots, N\} \right\}, \quad (14)$$

with $N = 9$ and q_{\min} , q_{\max} the minimum and maximum Euclidean distance, obtained by computing all pair-wise distances over 1,000 sampled session vectors. For HDBSCAN*-type algorithms, we set minimum cluster size $M = 2$.

For re-evaluation models, we took the approach already explained in Section III-B3. To train classifier $\hat{f}(\mathbf{a}_i, \mathbf{a}_{i'})$, we first run MS-DBSCAN* with the best found values for k and ϵ from earlier validation of MS-DBSCAN* on the training set to, together with the cookie clusters, obtain $\mathcal{X}_{\text{train}}$. Next, we computed the Manhattan, Euclidean, and infinity norm between \mathbf{a}_i and $\mathbf{a}_{i'}$, $(i, i') \in \mathcal{X}_{\text{train}}$ that were used as feature vector to train a logistic regression model. Although also other classifiers could be used, we considered that using a logistic regression model on a compressed input (the three

distance measures) would be a proper trade-off between model complexity and accuracy.

We selected for each point the $K = 1,000$ nearest neighbors to evaluate classifier \hat{f} on. All non-evaluated pairs received distance $\delta_{\max} = -\log(10^{-6})$. Next, the MS-(H)DBSCAN algorithms were evaluated using the new distance matrix W , where we experimented again with $k \in \{1, 3, 5\}$, and

$$\epsilon \in \left\{ q_{\min} + \frac{\ell(q_{\max} - q_{\min})}{N_{\text{re-eval}}} \mid \ell \in \{1, \dots, N_{\text{re-eval}}\} \right\}, \quad (15)$$

where $N_{\text{re-eval}} = 5$.

All algorithms excluding HDBSCAN* (i.e., including DBSCAN*) were trained using the `sklearn` package in Python [32] (version 0.22.1). `sklearn` was also used to compute error scores (see Section III-C3). We used the `hdbscan` package [33] (version 0.8.26) to obtain the dendrogram and condensed tree, based on which we could impose the maximum cluster size in the way described in Section III-B2. For both packages, the default parameters were used unless indicated otherwise.

3) Error metrics: We considered error metrics from two perspectives. First, we consider error measures with respect to overall website performance. More precisely, given some final clustering $\{\mathcal{B}_1^{\text{final}}, \dots, \mathcal{B}_m^{\text{final}}\}$, the following error measures are computed. 1) We compute the APE (absolute percentage error) between the real and estimated number of unique users (the latter being equal to m), 2) the Kullback-Leibler divergence (KL-divergence) between the real and estimated user session count distribution (the latter being equal to the cluster size distribution), and 3) the KL-divergence between the real and estimated user conversion distribution. Here, user conversion is defined as the fraction of items clicked per user over all shown (but not necessarily evaluated) items.

The second perspective is on the level of the clusters themselves, where we consider two error measures. To determine the quality of the clusters, we computed the adjusted Rand index (ARI) [34] between computed and real session clusters. Besides ARI, we also measure how well the model distinguishes whether each new session originates from an existing or already observed user, which is measured using the accuracy score.

Since ARI measures the overlap between the computed and real session clusters, we consider ARI to be our main error

measure, using the other error measures to study possible side-effects when optimizing for ARI.

IV. RESULTS

A. Results on base simulation case

Table II shows how the different models perform in terms of several error measures on both the training and test set. For each method, the shown results are the best results obtained under the different hyper parameters tried for that method under that dataset. I.e., in theory the hyper parameters might be slightly different between training and test, though in practice we found this was rarely the case.

The *OBS* model in the table are the scores one would obtain if the observed cookies would be used as clusters. Models using the classifier as distance measure are indicated using subscript p . What immediately becomes apparent is that compared to these observed cookie clusters, all methods perform considerably worse. Hence, in the scenario we consider: a single query where the true location $(\eta_1^{(u)}, \eta_2^{(u)})$ is only revealed by clicked and skipped item locations, our approaches do not come near what one would obtain if one would simply take the observed cookies.

However, the scores do reveal some interesting patterns. First, approaches using a probabilistic distance measure seem to overfit the data: they perform relatively well (compared to the other approaches) on various measures on the training set, but on the test set these results are mitigated. Here, MS-DBSCAN* seems to work best when considering multiple error measures. Looking at the results from different hyper-parameter settings for MS-DBSCAN* (Table III), we observe that selecting $k = 1$ performed best. Furthermore, due to our maximum size constraint the clusters did not alter for $\ell \geq 4$ ($\epsilon \geq 6.33$).

Furthermore, methods without a probabilistic distance measure do outperform the DBSCAN*-RAND method on most measures. I.e., they perform better at picking sessions originating from the same user from a given cluster B_j produced by DBSCAN*, than if we would pick session pairs at random. Although it is difficult to draw a firm conclusion, these findings might be an indication that the same user signal we try to infer from the click data is somewhat weak: if our methods would not pick up a signal at all, we would expect them to have the same result as the DBSCAN*-RAND method.

B. Results on multiple simulation scenarios

In order to judge the sensitivity of our findings on the parameter settings of the simulation model, we permuted the simulation settings to see if this would alter our results. In particular, we considered user distance sensitivity $q \in \{1, 2, 5, 10, 25, 50\}$ (denoted by USER_DIST_SENS_ $[q]$), number of items $V \in \{10, 100\}$ (denoted by ITEM_COUNT_ $[V]$), lifetime phases $\rho \in \{.15, .29, .43, .5, .57, .71, .85\}$ (denoted by LIFETIME_PHASES_ $[\rho]$), and salience $(\phi, \phi') \in \{1, 2, 5, 10\}^2$ (denoted by SALIENCE_ $[\phi][\phi']$). Whenever one parameter was permuted, the rest of the parameters was left at its value in the base case.

As re-running all models on all simulation settings would be computationally rather expensive, we only re-evaluated the best performing models on the simulation cases. Since in

our base case we found that the parameters $k = 1$, $\epsilon = (q_{\max} (q_{\min}/q_{\max})^{2/3})$ worked reasonably well, these parameters were used for MS-DBSCAN* and DBSCAN*-RAND. The maximum cluster size remained the same as in the base case.

Figure 2 shows how the models perform over the different simulation settings in terms of ARI, which is our main response variable of interest. The figure suggests that all cluster models do stochastically dominate DBSCAN*-RAND. Furthermore, MS-DBSCAN* seems to outperform the other clustering methods in terms of ARI. As assumptions like homogeneity of variance or normality do not hold in this case, we used a Kruskal-Wallis test, which rejects in this case that all median ARI scores over the different methods are the same (using significance level $\alpha = .01$, $p < 10^{-4}$). Pairwise (between MS-DBSCAN* and all other methods) one-sided pairwise Wilcoxon signed rank tests also indicate MS-DBSCAN* performed significantly better than the other methods (all p-values are smaller than 10^{-4}).

Table IV shows how MS-DBSCAN* performs on the various simulation cases. The rows in boldface have $\text{ARI} \geq 0.0025$. The results suggest that when strengthening the signal, that is increasing click probabilities, leads to some improvement in ARI. The most obvious way to do so is by decreasing the number of items (which, as we use bin counting, ensures each item has sufficient data for bin counting). However, these improvements remain small.

Table V shows how the different error measures correlate, using the error scores from all clustering algorithms on the various simulation cases. ARI seems to be weakly correlated with most other error measures, with the sign being in the desired direction (i.e., decrease in KL-divergence for both session count and conversion, but an increase in the new user accuracy). However, both ARI and the new user accuracy show a positive correlation with the percentage error in the number of unique users.

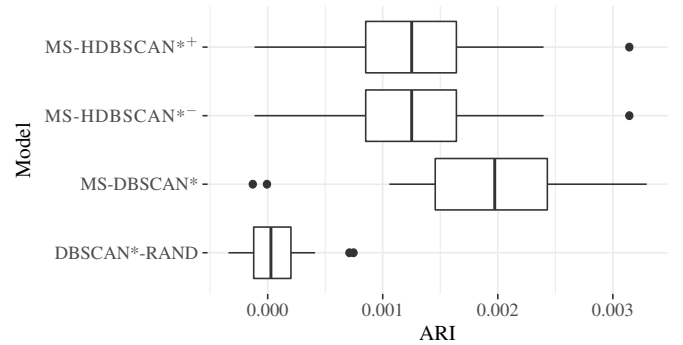


Figure 2. Scores over all simulations.

V. CONCLUSION AND DISCUSSION

In this paper, we presented a homogeneous query click simulation model, and illustrated its usage to the problem of uncovering users from their web sessions. The simulation model is composed of several models from which previous literature suggests that these models work well in explaining

TABLE II
RESULTS ON THE BASE CASE.

Model	Dataset	ARI	KL-div. session count	KL-div conversion	APE unique user	New user accuracy
MS-DBSCAN*	train	0.0012	0.55	0.13	15	0.56
MS-DBSCAN* _p	train	0.14	0.74	0.092	77	0.5
DBSCAN*-RAND	train	0.0002	1	0.096	0.011	0.42
MS-HDBSCAN* ⁺	train	0.0007	0.75	0.15	10	0.52
MS-HDBSCAN* ⁻	train	0.0007	0.75	0.15	10	0.52
MS-HDBSCAN* ⁺ _p	train	0.092	0.9	0.11	0.011	0.46
MS-HDBSCAN* ⁻ _p	train	0.1	0.9	0.11	0.011	0.46
<i>OBS</i>	<i>train</i>	<i>0.91</i>	<i>0.017</i>	<i>0.0032</i>	<i>15</i>	<i>0.95</i>
MS-DBSCAN*	test	0.0022	0.11	0.0026	60	0.56
MS-DBSCAN* _p	test	0.0015	1.4	0.13	6.8	0.4
DBSCAN*-RAND	test	0.0004	0.32	0.015	40	0.5
MS-HDBSCAN* ⁺	test	0.002	0.16	0.0042	53	0.55
MS-HDBSCAN* ⁻	test	0.002	0.16	0.0042	53	0.55
MS-HDBSCAN* ⁺ _p	test	0.0015	1.4	0.13	7.2	0.4
MS-HDBSCAN* ⁻ _p	test	0.0015	1.4	0.13	7.2	0.4
<i>OBS</i>	<i>test</i>	<i>0.91</i>	<i>0.1</i>	<i>0.0076</i>	<i>51</i>	<i>0.95</i>

TABLE III
ARI OF MS-DBSCAN* ON THE TRAINING SET OF THE BASE CASE.

ℓ	ϵ	k		
		1	3	5
1	0.013	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
2	3.44	0.0008	0.0004	0.0001
3	4.84	0.0011	0.0005	0.0004
4	6.33	0.0013	0.0006	0.0004
5	8.20	0.0013	0.0006	0.0004
6	10.76	0.0013	0.0006	0.0004
7	14.57	0.0013	0.0006	0.0004
8	20.94	0.0013	0.0006	0.0004
9	30.45	0.0013	0.0006	0.0004

typical patterns observed in click data, while remaining relatively simple. Such patterns include the position bias, cookie censoring, and users' utility over multiple products. Furthermore, we illustrated the simulation model on the problem of (partially observed) session clustering, that is, identifying unique users from their query-sessions. To solve the latter problem, we tested several mutations of (H)DBSCAN*, where these mutations differ from HDBSCAN*, or DBSCAN*, as they allow for incorporating a maximum cluster size. Furthermore, we consider both a Euclidean and probabilistic distance measure to determine whether a pair of sessions originated from the same user. The probabilistic distance measure was obtained using a pre-trained classification model.

Given a simulated dataset, we considered solving the problem of uncovering users from their web sessions by using (H)DBSCAN*-type clustering algorithms. Comparing (H)DBSCAN*-type algorithms with clusters one would obtain by using cookies, we found the accuracy of using cookies largely outperformed that of not using or partially using cookie data. This considerable difference seems to be due to two reasons. 1) The simulated censored cookies turned out to be rather accurate, implying that, assuming the parameters used for cookie censoring adapted from previous literature are accurate, censoring in cookie data does not impose that much of a problem in accurately measuring the metrics studied in this paper. These metrics being the number of unique users, user sessions count distribution, user conversion distribution,

the quality of session clusters (in terms of adjusted Rand index (ARI)), and estimating whether the next session originates from a new or existing user. 2) As we only consider a homogeneous query, the users' preferences are only revealed from the items users clicked, a signal the various (H)DBSCAN*-type algorithms find difficult to detect. Strengthening this signal, e.g., by increasing the number of clicks, leads to a small but significant improvement in ARI.

Other interesting observations include the difference between using Euclidean distance and a probability distance measure in the (H)DBSCAN*-type algorithms, the latter being obtained from training a classifier on detecting whether session pairs originate from the same user. The results show that the probabilistic classifier tends to overfit. Where some methods using probabilistic distance measures performed reasonable on the training set, they were outperformed by methods using Euclidean distance on the test set.

By studying the correlations between the various error metrics considered in this paper, we observe that some error measures show contradictory correlations. In particular, the positive correlation between cluster ARI and average percentage error in the number of unique users (.38), and between the accuracy in estimating whether the next session originates from a new user and the new user average percentage error (.95), indicate that optimizing for one of these error measures may lead to decreased performance in the other.

Although our findings suggest that the practicality of session clustering from single query click data is limited, the usage of the simulation model did allow for studying the sensitivity of the clustering algorithms on different click behavior, something that would not easily have been possible with real click data. It also allowed us to study the effects of user censoring caused by cookie churn or the usage of multiple devices. This showed that if we adopt models for cookie churn behavior found in the literature, this censoring only has a small effect on the accuracy of the website metrics discussed in this paper, with an exception for estimating the number of unique users.

Given our findings, a number of questions remain. First, it would be interesting to extend the simulation model to allow for multiple queries. As the solutions to the (multi-query)

TABLE IV
RESULTS MS-DBSCAN* ON OTHER SIMULATION CASES.

Simulation case	ARI	KL-div. conversion	KL-div. session count	APE unique user	New user accuracy
base_case	0.0021	0.0044	0.095	62	0.53
item_count_10	0.0025	0.0006	0.049	67	0.57
item_count_100	0.0015	0.0053	0.098	59	0.54
lifetime_phases_15	0.0014	0.014	0.14	56	0.56
lifetime_phases_29	0.0016	0.0076	0.1	59	0.55
lifetime_phases_43	0.0021	0.008	0.11	58	0.56
lifetime_phases_5	0.0021	0.0044	0.095	62	0.53
lifetime_phases_57	0.0019	0.0049	0.11	61	0.55
lifetime_phases_71	0.0019	0.0054	0.084	60	0.56
lifetime_phases_85	0.0028	0.005	0.092	61	0.53
salience_1_1	0.0012	0.0018	0.07	64	0.58
salience_1_2	0.0022	0.0019	0.059	65	0.57
salience_1_5	0.0014	0.0015	0.085	62	0.59
salience_1_10	0.0026	$< 10^{-4}$	0.084	62	0.58
salience_2_1	0.0011	0.0026	0.15	53	0.55
salience_2_2	0.002	0.0033	0.19	51	0.55
salience_2_5	0.0027	0.0005	0.12	56	0.57
salience_2_10	0.0018	$< 10^{-4}$	0.094	60	0.58
salience_5_1	$< 10^{-4}$	0.011	0.25	44	0.52
salience_5_2	$< 10^{-4}$	0.021	0.2	51	0.53
salience_5_5	0.0021	0.0044	0.095	62	0.53
salience_5_10	0.002	$< 10^{-4}$	0.079	62	0.57
salience_10_1	0.0016	0.0049	0.051	64	0.57
salience_10_2	0.0014	0.0034	0.065	66	0.57
salience_10_5	0.0018	0.0045	0.1	60	0.56
salience_10_10	0.0012	0.0001	0.042	71	0.63
user_dist_sense_1	0.0021	0.0044	0.095	62	0.53
user_dist_sens_2	0.0029	0.012	0.17	49	0.54
user_dist_sens_5	0.0033	0.0092	0.15	49	0.53
user_dist_sens_10	0.0026	0.002	0.12	57	0.56
user_dist_sens_25	0.0024	$< 10^{-4}$	0.13	59	0.57
user_dist_sens_50	0.0032	0.0002	0.09	64	0.56

TABLE V
CORRELATION MATRIX ERROR MEASURES.

	ARI	KL-div. conversion	KL-div. session count	APE unique user	New user accuracy
ARI	1.00				
KL-div. conversion	-0.15	1.00			
KL-div. session count	-0.16	0.60	1.00		
APE unique user	0.38	-0.52	-0.92	1.00	
New user accuracy	0.39	-0.59	-0.85	0.95	1.00

CIKM 2016 and ICDM 2015 cross-device matching competitions were quite successful, a logical hypothesis would be that incorporating multiple queries into the simulation model would improve the results obtained from (H)DBSCAN*-type algorithms. On the other hand, more diversity also causes clicks to be more spread across items that may lead to decreasing clustering performance.

Second, in this study, we only used a logistic regression model to approximate the probability of two sessions originating from the same user. Given the limited success of this approach so far, it would be interesting to consider other approaches. As the limited results seem to be due to overfitting, including regularization or using bagging could lead to better results.

Third, there is still limited knowledge on how cookie censoring occurs. Currently, multiple models exist in the literature, but most models only consider a specific type of censoring (e.g., only censoring by cross-device usage or cookie

churn), from which one cannot infer how these different types of censoring interact. Also, as discussed in Section III-A2, literature providing parametric models for cookie churn, user lifetime and absence time (the time between two sessions) seems to be contradictory in terms of tail probabilities. Hence, click simulation models that incorporate cookie censoring would benefit from studies taking a more holistic view on cookie censoring.

REFERENCES

- [1] C. de Ruijt and S. Bhulai, "Detecting users from website sessions: A simulation study," in *DATA ANALYTICS 2020, The Ninth International Conference on Data Analytics*, 2020, pp. 35–40.
- [2] A. Dasgupta, M. Gurevich, L. Zhang, B. Tseng, and A. O. Thomas, "Overcoming browser cookie churn with clustering," in *Proceedings of the fifth ACM international*

- conference on Web search and data mining. ACM, 2012, pp. 83–92.
- [3] ICDM, *ICDM 2015: Drawbridge Cross-Device Connections*, 2015, retrieved from: <https://www.kaggle.com/c/icdm-2015-drawbridge-cross-device-connections>, accessed November 17, 2021.
 - [4] CIKM, *CIKM Cup 2016 Track 1: Cross-Device Entity Linking Challenge*, 2016, retrieved from: <https://competitions.codalab.org/competitions/11171>, accessed November 17, 2021.
 - [5] S. Kim, N. Kini, J. Pujara, E. Koh, and L. Getoor, “Probabilistic visitor stitching on cross-device web logs,” in *Proceedings of the 26th International Conference on World Wide Web*. ACM, 2017, pp. 1581–1589.
 - [6] D. Jin, M. Heimann, R. Rossi, and D. Koutra, “node2bits: Compact time- and attribute-aware node representations,” in *ECML/PKDD European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2019.
 - [7] A. Chuklin, I. Markov, and M. d. Rijke, *Click models for web search*. Morgan & Claypool Publishers, 2015.
 - [8] D. Fleder and K. Hosanagar, “Blockbuster culture’s next rise or fall: The impact of recommender systems on sales diversity,” *Management science*, vol. 55, no. 5, pp. 697–712, 2009.
 - [9] D. Coey and M. Bailey, “People and cookies: Imperfect treatment assignment in online experiments,” in *Proceedings of the 25th International Conference on World Wide Web*. ACM, 2016, pp. 1103–1111.
 - [10] G. Dupret and M. Lalmas, “Absence time and user engagement: evaluating ranking functions,” in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 173–182.
 - [11] G. D. Montanez, R. W. White, and X. Huang, “Cross-device search,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 1669–1678.
 - [12] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, “Where you are is who you are: User identification by matching statistics,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 358–372, 2016.
 - [13] R. Saha Roy, R. Sinha, N. Chhaya, and S. Saini, “Probabilistic deduplication of anonymous web traffic,” in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 103–104.
 - [14] Q. Wang, “Recombining customer journeys with probabilistic cookie matching: A supervised learning approach,” in *Machine learning applications in operations management and digital marketing*, 2019, ch. 6, pp. 127–139.
 - [15] J. Lian and X. Xie, “Cross-device user matching based on massive browse logs: The runner-up solution for the 2016 CIKM cup,” *arXiv preprint arXiv:1610.03928*, 2016.
 - [16] N. K. Tran, “Classification and learning-to-rank approaches for cross-device matching at CIKM cup 2016,” *arXiv preprint arXiv:1612.07117*, 2016.
 - [17] R. Díaz-Morales, “Cross-device tracking: Matching devices and cookies,” in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 1699–1704.
 - [18] M. C. Phan, A. Sun, and Y. Tay, “Cross-device user linking: URL, session, visiting time, and device-log embedding,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 933–936.
 - [19] M. C. Phan, Y. Tay, and T.-A. N. Pham, “Cross device matching for online advertising with neural feature ensembles: First place solution at CIKM cup 2016,” 2016.
 - [20] R. Song, S. Chen, B. Deng, and L. Li, “eXtreme gradient boosting for identifying individual users across different digital devices,” in *International Conference on Web-Age Information Management*. Springer, 2016, pp. 43–54.
 - [21] U. Tanielian, A.-M. Tusch, and F. Vasile, “Siamese cookie embedding networks for cross-device user matching,” in *Companion Proceedings of the The Web Conference 2018*. ACM, 2018, pp. 85–86.
 - [22] C. Karakaya, H. Toğuş, R. S. Kuzu, and A. H. Büyüklü, “Survey of cross device matching approaches with a case study on a novel database,” in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 2018, pp. 139–144.
 - [23] L. Olejnik, C. Castelluccia, and A. Janc, “On the uniqueness of web browsing history patterns,” *annals of telecommunications-Annales des télécommunications*, vol. 69, no. 1-2, pp. 63–74, 2014.
 - [24] L. Jalali, M. Khan, and R. Biswas, “Learning and multi-objective optimization for automatic identity linkage,” in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 4926–4931.
 - [25] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.
 - [26] R. J. Campello, D. Moulavi, and J. Sander, “Density-based clustering based on hierarchical density estimates,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 160–172.
 - [27] R. J. Campello, D. Moulavi, A. Zimek, and J. Sander, “Hierarchical density estimates for data clustering, visualization, and outlier detection,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 10, no. 1, pp. 1–51, 2015.
 - [28] C. Luna-Nevarez and M. R. Hyman, “Common practices in destination website design,” *Journal of destination marketing & management*, vol. 1, no. 1-2, pp. 94–106, 2012.
 - [29] O. Chapelle and Y. Zhang, “A dynamic bayesian network click model for web search ranking,” in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 1–10.
 - [30] L. McInnes and J. Healy, “Accelerated hierarchical density clustering,” *arXiv preprint arXiv:1705.07321*, 2017.
 - [31] I.-K. Yeo and R. A. Johnson, “A new family of power transformations to improve normality or symmetry,” *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.
 - [32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
 - [33] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” *Journal of Open Source*

Software, vol. 2, no. 11, p. 205, 2017.

- [34] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

APPENDIX

TABLE VI
LIST OF NOTATION.

Variable	Description
$\{1, \dots, n\}$	Set of query-sessions, indexed by i
$S_i = \{1, \dots, T\}$	Set of items in SERP of session i , indexed by t
$\mathcal{V} = \{1, \dots, V\}$	Set of all items, indexed by v
$\mathcal{U} = \{1, \dots, U\}$	Set of all users, indexed by u
$r_i(t)$	Item $v \in \mathcal{V}$ at position t in the SERP of query-session i
$r_i^{-1}(v)$	Position of item v in the SERP of query-session i , zero if $v \notin S_i$
r_i^{\max}	Largest position of a clicked item $v \in S_i$, zero if no items were clicked
$R_v^{(i)}; \phi_{u,v}^{(R)}$	Attraction of user i for item v , with $\mathbb{P}(R_v^{(i)} = 1) = \phi_{u,v}^{(R)}$ given $v \in S_i$
$S_v^{(i)}; \phi_{u,v}^{(S)}$	Satisfaction of user i for item v , with $\mathbb{P}(S_v^{(i)} = 1) = \phi_{u,v}^{(S)}$ given $v \in S_i$
$E_t^{(i)}$	Whether item at position t in SERP i was evaluated
$y_t^{(i)}$	Whether item at position t in SERP i was clicked
η_u	Vector denoting the position of an user u in the user-item space
ψ_v	Vector denoting the position of an item v in the user-item space
$\nu^{(A)}, \nu^{(S)}$	Salience parameters for attraction and satisfaction
q	Users' preference for nearby items
$\omega_{u,v}$	Distance between user u and item v in the user-item space
$\hat{\phi}_v$	Overall estimated popularity of item $v \in \mathcal{V}$
F^{cookie}	Cookie lifetime distribution (hyper-exponential) with parameters λ and \mathbf{p}
$\mathcal{T}_{u,o,d}^{\text{cookie}} \sim F^{\text{cookie}}$	R.v. denoting the cookie lifetime for the o -th cookie of user u on device d
F^{abs}	User absence distribution (Pareto-I) with parameters α (shape) and m (scale)
$\mathcal{T}_{i,u}^{\text{abs}} \sim F^{\text{abs}}$	R.v. denoting the time between the i -th and $i+1$ -th session of user u
F^{user}	User lifetime distribution (sum of N_u hyper-exponentials) with parameters λ , \mathbf{p} and ρ (geometric parameter for N_u)
$\mathcal{T}_u^{\text{user}} \sim F^{\text{user}}$	R.v. denoting the user lifetime of user u
P, π	Device transition matrix and initial device probabilities

Algorithm 2: SIMULATE_CLICKS

```

1 Simulate_clicks ( $\mathcal{U}$ )
2   for  $u \in \mathcal{U}$  do
3     /* Draw initial device and cookie
4       lifetime, and draw the user's
5       lifetime */
6      $D \leftarrow \text{dic}(); i \leftarrow 1; o \leftarrow 1; t \leftarrow 0;$ 
7     Draw device  $d$  from  $\text{MULTINOM}(\pi); D[d] \leftarrow o;$ 
8     Draw  $\mathcal{T}_{u,o,d}^{\text{cookie}}$  from  $\text{HYPEREXP}(\lambda, \mathbf{p}); \mathcal{T}_u^{\text{user}}$  from
9        $\text{REPHYPEREXP}(\rho, \lambda, \mathbf{p});$ 
10
11    /* Simulate new query-sessions while the
12      user's lifetime has not elapsed */
13    while  $t \leq \mathcal{T}_u^{\text{user}}$  do
14      /* 1) Simulate clicks */
15      Draw  $S_i$  in its respective order by repetitively
16        drawing from
17         $\text{MULTINOM}(\hat{\phi}_v / \sum_{v' \in \mathcal{V}} \hat{\phi}_{v'}; v \in \mathcal{V} \setminus S_i);$ 
18      Draw  $R_v^{(i)}, S_v^{(i)}$  from  $\text{BERNOULLI}(\phi_{u,v}^{(R)})$  and
19         $\text{BERNOULLI}(\phi_{u,v}^{(S)})$  resp. for all  $v \in S_i;$ 
20      Compute  $E_t^{(i)}, y_t^{(i)}$ , and recompute  $S_v^{(i)}$  according
21        to Equations (1) to (6);
22      Append  $(i, u, o, S_i, y_i)$  to  $\mathcal{D};$ 
23      /* 2) Draw the time until the next
24        session and update  $t$  accordingly */
25      Draw  $\mathcal{T}_{i,u}^{\text{abs}}$  from  $\text{PARETO-I}(m, \alpha);$ 
26       $t \leftarrow t + \mathcal{T}_{i,u}^{\text{abs}}; i \leftarrow i + 1;$ 
27      /* 3) Update the device for the next
28        session */
29      Draw  $d'$  from  $\text{MULTINOM}(I_d P);$ 
30      if  $d' \neq d$  then
31        if not  $D.\text{exists}(d)$  then
32           $o \leftarrow o + 1;$ 
33          Draw  $\mathcal{T}_{u,o,d}^{\text{cookie}}$  from  $\text{HYPEREXP}(\lambda, \mathbf{p});$ 
34           $\mathcal{T}_{u,o,d}^{\text{cookie}} \leftarrow \mathcal{T}_{u,o,d}^{\text{cookie}} + t;$ 
35        else
36           $o \leftarrow D[d'];$ 
37         $d \leftarrow d';$ 
38
39      /* 4) Simulate cookie churn */
40      if  $t > \mathcal{T}_{u,o,d}^{\text{cookie}}$  then
41         $o \leftarrow o + 1;$ 
42        Draw  $\mathcal{T}_{u,o,d}^{\text{cookie}}$  from  $\text{HYPEREXP}(\lambda, \mathbf{p});$ 
43         $\mathcal{T}_{u,o,d}^{\text{cookie}} \leftarrow \mathcal{T}_{u,o,d}^{\text{cookie}} + t;$ 
44         $D[d] \leftarrow o;$ 
45
46  return  $\mathcal{D}$ 

```

Military REACH: A University-wide Collaboration

Fatemeh Jamshidi

Department of Computer Science and
Software Engineering
Auburn University
Auburn, Alabama, USA
Email: fzf0007@auburn.edu

Abhishek Jariwala

Department of Computer Science and
Software Engineering
Auburn University
Auburn, Alabama, USA
Email: avj0007@auburn.edu

Bibhav Bhattarai

Department of Computer Science and
Software Engineering
Auburn University
Auburn, Alabama, USA
Email: bzb0079@auburn.edu

Katherine Abbate

Project Manager, Military REACH
Auburn University
Auburn, Alabama, USA
Email: kma0057@auburn.edu

Daniela Marghitu

Department of Computer Science and
Software Engineering
Auburn University
Auburn, Alabama, USA
Email: marghda@auburn.edu

Mallory Lucier-Greer

College of Human Sciences
Human Development and Family Science
Auburn University
Auburn, Alabama, USA
Email: mluciergreer@auburn.edu

Abstract—At the federal level, a partnership composed of the Department of Defense (DoD), the Department of Agriculture (USDA), and colleges and universities throughout the United States of America work toward serving military families. Through this partnership, cooperative agreements are executed to support the needs of service members and their families. One such cooperative agreement between DoD, USDA, and Auburn University is Military REACH. This project aims to bridge the gap between military family research and practice by mobilizing peer-reviewed family science research into practical applications for military families and those who work on behalf of military families. At Auburn University, this project is an interdisciplinary collaboration between the Department of Human Development and Family Science, the Department of Computer Science, and the academic libraries. This paper aims to present the Military REACH website, the new searching functionalities added to the project to increase the number of active users, and a newly launched mobile application that is positioned to promote access to resources and assess the usefulness of the project's research summaries. In this paper, we present the functionality and qualitative data analysis of this additional aspect of the research.

Keywords—Military Families; Applications; Resources.

I. INTRODUCTION

For the past four years, the Auburn University Libraries and Computer Science Department have supported the University's research enterprise in a new way: by adopting a new collaborative model and serving as a high-level Information Technology (IT) and data-management consultants to faculty researchers who are pursuing external funding [1]. A practical example of this model in action is the Military REACH project at Auburn University funded by the Departments of Agriculture and Defense (USDA/NIFA Award No. 2017-48710-27339; PI, Dr. Mallory Lucier-Greer). The purpose of Military REACH is to make research accessible to policy makers, helping professionals, and military families in a manner that is inviting, easily understood, and meaningful for their everyday context [2]. Our

team works to critically evaluate empirical research related to military families and translate it into useful tools. These tools are actively disseminated to policy makers and military helping professionals to inform their decisions and practices as they work to support and enhance the lives of service members and their families. Specifically, the objective of this project is to provide high-quality resources to the Department of Defense (DoD) in the form of research and professional development tools across the spectrum of family support, resilience, and readiness. This work is primarily supported by the DoD's Office of Military Community and Family Policy. The purpose of this project is achieved through three primary deliverables, including:

- Provide timely, high-quality research reports at the request of DoD.
- Re-engineer, grow, and promote an online library of current research and its implications related to the well-being of military families.
- Design and market professional development opportunities, tools, and resources for youth development professionals.

The Military REACH Project is now in its fifth year and continuing at Auburn University for the foreseeable future; indeed, it has highlighted the library's value as an IT partner and led to research partnerships and collaborative funding proposals with other units on campus. This paper describes the related functions that are designed and implemented for each operator. The paper is organized as follows. In Section II, we provide pertinent background information about the project. Section III introduces our efforts to serve military families and covers the design and implementation of the website. Section IV demonstrates evaluation methods using Google Analytics. Section V provides evaluation results of the website. Section VI presents our mobile app as an important step forward.

Section VII presents the conclusion with suggestions for future directions.

II. RELATED WORK

Military REACH started by evaluating existing research in the context of Research Infrastructures (RI) and Digital Libraries (DL). Recent reviews of digital preservation [3] and projects that promote research and awareness in the areas of digital preservation include Curl Exemplars for Digital Archives (CEDARS) [4].

Two decades of research have worked to improve awareness of the digital preservation challenge and encouraged some organizations to improve the longevity of their digital resources. One of the most significant streams of research has been within cultural institutions, sometimes in collaboration with industry partners, to develop solutions to operational problems in these institutions [5]. National, regional, and University archives and libraries in Australia, Canada, Belgium, Denmark France, Germany, the Netherlands, New Zealand, Sweden, Switzerland, the U.K., the U.S., and elsewhere have investigated the implementation of institutional repositories, preservation, and strategies for Web archiving.

III. COLLABORATIVE EFFORTS TO SERVE MILITARY FAMILIES

Working closely with the Military REACH team in the Department of Human Development and Family Science, the library's IT department contributed to the original funding proposal and has guided network architecture, Web development, IT tools and solutions, sustainability, data management, accessibility, usage statistics, and automated methods for identifying recently published research.

A. Design and Implementation

The REACH Web application has an architecture that can be implemented in three layers, as shown in Figure 1.

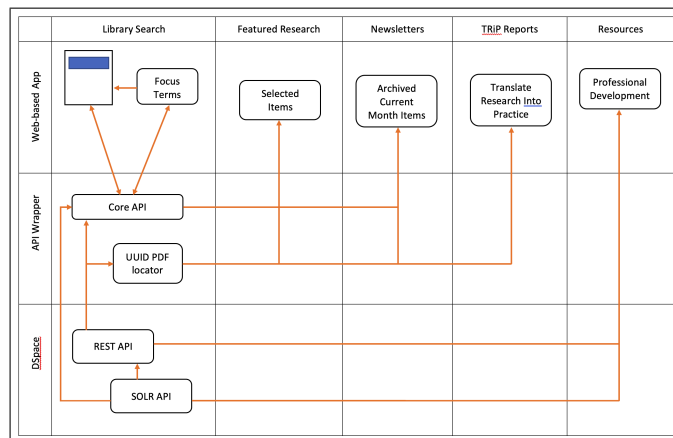


Figure 1. REACH System Architecture.

- **Web-based app:** This layer is the front-end of the application, where we mainly use Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript in Java Server Pages (JSP). Also, the Cascade Content Management System (CMS) used in this project, to manage the JSP, falls under this layer.

- **Application Programming Interface (API) Wrapper:** This layer is the back-end layer, where we use JAVA programs to write classes and methods that handle various functionalities of the website such as search, filter, sort, and many more functionalities.
- **DSpace:** DSpace is an open-source repository software package mostly used to create open access repositories for the scholarly and published digital content. DSpace is the central database of the application. All Military REACH related research articles are stored in this layer. DSpace uses Apache SOLR based search for metadata and full-text contents, all of which are stored in a relational database and supports the use of PostgreSQL. Also, DSpace is used to manage and preserve all the formats of digital content (PDF, Word, JPEG, MPEG, TIFF files). Likewise, it also allows a group-based access to control the setting of level-based permission to individual files.

1) Introduction to the Cascade Content Management Systems: To make the website easy to control and manage, Military REACH uses CMS. Cascade CMS is used in the Web application to manage site content, allowing multiple contributors to create, edit, and publish new entries. Content created in a Cascade CMS is stored in Cascade as an XML file and displayed in a presentation layer based on a set of templates. Programming languages such as Extensible Stylesheet Language Transformations (XSLT), and Velocity [6] are used to transform the Extensible Markup Language (XML) file into HTML/JSP pages.

Fundamental features of Cascade CMS are:

- Content creation (allows users to easily create and format content),
- Content storage (stores content in one place, in a consistent fashion),
- Workflow management (assigns privileges and responsibilities based on roles such as authors, editors, and administrators), and
- Publishing (organizes and pushes content live).

2) Cascade Content Management Systems Advantages: What makes Cascade particularly beneficial to a Web application, such as the Military REACH website, is the ease of updating resources and predefined pages. The “What You See Is What You Get” (WYSIWYG) editors included in the platform allow users to enter text and upload images with less basic knowledge of HTML or CSS (front end languages to make the website look appealing).

The other advantage of Cascade CMS is its collaborative nature. Multiple users can log on and contribute, schedule, or edit content to be published. Since the interface is browser-based; therefore, Cascade can be accessed from anywhere by multiple users. Similarly, Cascade CMS has an efficient, reliable way of sending frequent alerts to the users and site administrators of pages that have not been updated for a certain duration of time.

The use of in-built features such as the daily content report, task manager, and content review dates help collaborative teams stay updated on current tasks. Lastly, Cascade has a community of over 100,000 active users that are frequently

using the platform and are readily available to voice their experiences with using features and capabilities of Cascade.

3) Use of Cascade Content Management System in Military REACH:

- Two pages of the website, the Team members and Community Connections pages, are entirely made in the Cascade CMS. These pages can be easily updated by members of the team who may not necessarily have the technical knowledge of creating and updating web pages.
- Other pages, such as Home page, Family Focus page, and Contact Us page, are hybrid pages, where all of the texts displayed in the pages can be edited from Cascade. Other major functionalities within the hybrid pages are handled in the back-end JAVA classes.
- Therefore, having Cascade pages and hybrid pages simultaneously provides us with more flexibility for both the technical and non-technical team members to be involved in the organization.

IV. EVALUATION METHODS

Our evaluation methods are listed in this section.

A. Google Analytics

Military REACH has been using Google Analytics to access the user data since March 1, 2019 until present. Google Analytics data do not include any personally identifiable information. They are presented to stakeholders as aggregate data, making it a practical tool used in research settings without ethical concerns [7] [8]. The Web development team installed Google Analytics by adding a tracking tag for Military REACH to monitor the usability of the website. The tracking tags are a combination of JavaScript and computer programming language used to develop the website. The tracking tag code allows developers to receive data related to the users' behavior on the website. The data can proceed from diverse avenues. For example, the URL of the page and the device used to access the site. Tracking codes primarily collect data on the nature of the visit, such as the contents viewed, length of the session, average time on each page, location, and so on. This information is in a real-time, interactive dashboard format that can be viewed by logging in to Google Analytics.

B. User Engagement

This project focuses on several indicators from Google Analytics to evaluate the level of engagement. These indicators contain the number of returning users (n), bounce rate, number of pages accessed per session (n), mean session, and time spent on each page (minutes, seconds). The number of returning users reflects the number of sessions visited through the same client IP. A high number of returning users indicates a strong level of engagement with the Web-based platform [7][9]. The bounce rate is a percentage of single-page sessions in which there was no interaction with the page. A high bounce rate means minimal interaction with the page; however, it could also mean that users exit the page after finding what they were looking for right away. A low bounce rate can refer to a high overall engagement, especially for a multi-component platform like Military REACH. For example, there are not many available resources that would provide mental health support on

the platform's home page. Therefore, users will often need to interact with various searching tools and Web pages to access the required information. The number of pages per session indicates the number of Web pages that the user viewed in a single session. The mean session duration (minutes, seconds) provides information on the average duration of the time users spend on the website. There are different interpretations of measuring user engagement. For example, many pages per session could occur from a high level of engagement, while it could also cause a superficial exploration of several pages. Additionally, a long session duration can result from increased attention, but it could also be because the user keeps the Web page open while engaging in the other irrelevant activities.

C. Platform Improvement

Military REACH considers multiple indicators from Google Analytics to inform the improvement of the platform. These indicators include page views, mean duration of visit, and bounce rate when accessing resources provided on the website (e.g., Family Focus page, TRIP reports page). The most visited pages were observed in terms of their overall average time spent on the page to understand which tools or pages were most beneficial or viewed.

The entrance rate illustrates a proportion of sessions starting from a given page. In comparison, the exit rate results from a ratio of sessions ending from a given page. The information regarding the entrance rate may explain which Web page serves as the first impression for the users. The exit rate may indicate when users felt disengaged or had consumed adequate data needed for the session. Google Analytics provides information on the type of devices users are using to access the website. Such data can allow us to consider if implementing a mobile app for Military REACH would be practical or not. The three primary devices of interest to the current investigation are desktops, tablets, and mobile phones (counted here as mobile devices).

D. Marketing Strategy

Military REACH aims to reach as many users as possible. Therefore, we use a multi-pronged approach to inform our marketing strategy. The team connects with various military-connected organizations, especially around the United States. Twitter, Facebook, and LinkedIn accounts were also created to distribute awareness about the platform. To improve the marketing strategy, we also review Google Analytics to examine how the website is used and where the website it used. The methods include a direct link (i.e., typing the Web URL directly into a browser); organic search (i.e., entry through a search engine); and referrals via another website via social media via email. Understanding which ways are most accessible for users can help to improve the marketing strategy. Military REACH also examines the locations of users from different countries around the world and their proximity to military installations.

V. EVALUATION RESULTS

The first version of the website was based on a single page application (March 2019 - November 2019). However, to better access our users' data, we switched to a multiple page application using Java Server Pages (JSP) and Servlets (November 2019 - present). The following are the results from

Google Analytics, which show the positive impact of this change in user engagement and platform functionalities.

A. User Engagement

We recorded a total of 1,806 users from on the initial iteration of the website platform between March 1, 2019 - November 2, 2019 (shown in Figure 2), then a total of 3,131 users between November 2, 2019 and June 11, 2020, after we switched to a multiple page application (shown in Figure 3). The last year of operation for the Military REACH platform served 9,059 users from June 11, 2020 - September 12, 2021; this is a meaningful boost compared to the total of 4,824 users from March 1, 2019 - June 10, 2020 (shown in Figure 4).

This improvement may be attributed to two fundamental functionalities focused on increasing user engagement. The first was implementing Android and iOS mobile apps to promote outreach (discussed in the next section). The second was adding an opportunity for researchers to share their own publications; researchers whose publications relate to military families have an opportunity to request their article be shared on the Military REACH website. This functionality has provided provide Military REACH's active users to be more involved in the project.

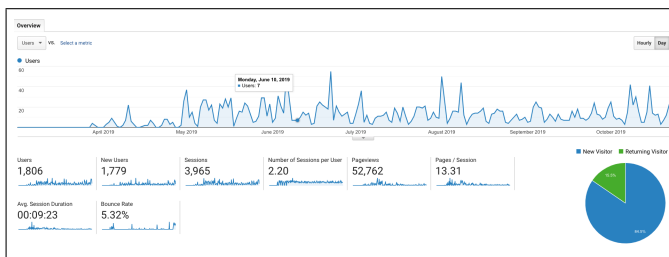


Figure 2. REACH overview presented in Google Analytics (March 1, 2019 - November 2, 2019).

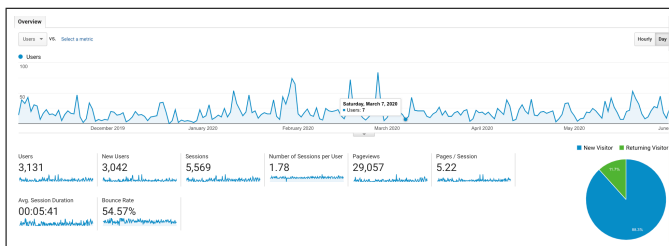


Figure 3. REACH overview presented in Google Analytics (November 2, 2019 - June 11, 2020).

The results show that user engagement is increasing because of social media marketing, conferences, and overall better efficiency and effectiveness of the website.

B. Platform Improvement

Table 1 presents details of the top ten most viewed pages. In March 2019 to November 2019; the Military REACH home page, which acts as the landing page, accounted for 51.41% (7,782/15,136) of all entries when the website was still a single page application using Angular and Typescript. However, after transforming to multiple page applications, users can access the resources they are looking for, using shared links on our

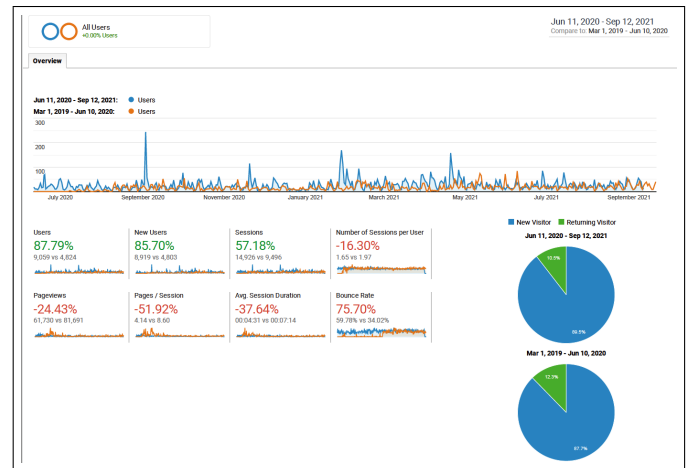


Figure 4. REACH overview presented in Google Analytics (June 11, 2020 - September 12, 2021 Compare to: March 1, 2019 - June 10, 2020).

social media or email. Table II is a representation of pages view compared in two period of times 2019-2020 and 2020-2021.

A list of devices used by Military REACH users to access the site is presented in Table III, indicating that the platform was accessed mostly via desktops (2,112/3,130, 67.43%) during July 2019 to June 2020. However, last year after implementing REACH mobile apps, users were more engaged using their cell phones. Table IV, represent the analysis of devices used by users two period of times 2019-2020 and 2020-2021. Further- more, sessions completed via desktops had a higher average session duration than those completed via other devices.

C. Marketing Strategy

Approximately 89.58% (2,804/3,129) of the users accessed the website from the United States. Table V shows that the users accessed the platform from around the world (Figure 5).

Google Analytics was a helpful tool to process the evaluation of the open-access, Web-based Military REACH platform.

The process evaluation provided information about the ways to keep users engaged, marketing strategies, and the aspects of the platform that required improvement.

VI. MILITARY REACH EFFICACY STUDY

To advance the work of the project and examine the usefulness of the research summaries created by Military REACH, our team created a mobile application that provides helping professionals (e.g., therapists, social workers) access updated research on military families. Our team has also tracked analytics for the app to better understand user engagement. Recently mobile applications have become more reliant on big data. Machine learning, big data, database, and deep learning concepts have been utilized not only in almost all the engineering fields, but also in other fields such as economics. It is a difficult task for Relational Database Management System (RDBMS) to manage the unstructured data. Firebase is a new technology to assist handling large amount of unstructured data [10]. Compared to RDBMS, Firebase is more efficient and faster. In this section we focus on the application of Firebase

TABLE I. REACH MOST VIEWED PAGES.

Page	Pageviews	Unique Pageviews	Avg. Time on Page
Change	79.54%	208.24%	82.39%
Total Nov 2, 2019 - Jun 11, 2020	27,111	18,800	0:01:20
Total Jul 15, 2019 - Nov 2, 2019	15,136	6,118	0:00:44
1 /homepage			
Nov 2, 2019 - Jun 11, 2020	4,244 (15.62%)	3,046 (16.15%)	0:01:09
Jul 15, 2019 - Nov 2, 2019	7,782 (51.41%)	2,393 (39.11%)	0:00:25
% Change	-45.46%	27.29%	171.80%
2 /Redirect			
Nov 2, 2019 - Jun 11, 2020	1,376 (5.06%)	634 (3.36%)	0:01:59
Jul 15, 2019 - Nov 2, 2019	135 (0.89%)	51 (0.83%)	0:00:46
% Change	919.26%	1143.14%	158.33%
3 /reachlibrary.jsp			
Nov 2, 2019 - Jun 11, 2020	1,127 (4.15%)	635 (3.37%)	0:00:30
Jul 15, 2019 - Nov 2, 2019	93 (0.61%)	49 (0.80%)	0:00:38
% Change	1111.83%	1195.92%	-22.78%
4 /Updates			
Nov 2, 2019 - Jun 11, 2020	862 (3.17%)	591 (3.13%)	0:02:13
Jul 15, 2019 - Nov 2, 2019	127 (0.84%)	58 (0.95%)	0:01:18
% Change	578.74%	918.97%	69.97%









TABLE II. REACH MOST VIEWED PAGES.

Page	Pageviews	Unique Pageviews	Avg. Time on Page
62,051 vs 81,370	42,585 vs 31,537	00:01:26 vs 00:00:57	14,969 vs 9,449
1 /			
Jun 10, 2020 - Sep 12, 2021	8,766 (14.13%)	5,743 (13.49%)	0:01:11
Mar 1, 2019 - Jun 9, 2020	37,197 (45.71%)	8,322 (26.39%)	0:00:41
% Change	-76.43%	-30.99%	72.17%
2 /Redirect			
Jun 10, 2020 - Sep 12, 2021	2,413 (3.89%)	1,158 (2.72%)	0:01:50
Mar 1, 2019 - Jun 9, 2020	1,495 (1.84%)	672 (2.13%)	0:01:52
% Change	61.40%	72.32%	-1.79%
3 /reachteam.jsp			
Jun 10, 2020 - Sep 12, 2021	2,216 (3.57%)	932 (2.19%)	0:01:36
Mar 1, 2019 - Jun 9, 2020	709 (0.87%)	338 (1.07%)	0:01:10
% Change	212.55%	175.74%	36.78%
4 /recruitment.jsp			
Jun 10, 2020 - Sep 12, 2021	1,759 (2.83%)	1,595 (3.75%)	0:04:24
Mar 1, 2019 - Jun 9, 2020	0 (0.00%)	0 (0.00%)	0:00:00
% Change	=%	=%	=%
5 /Families			
Jun 10, 2020 - Sep 12, 2021	1,615 (2.60%)	823 (1.93%)	0:01:01
Mar 1, 2019 - Jun 9, 2020	785 (0.96%)	393 (1.25%)	0:00:49
% Change	105.73%	109.41%	24.34%
6 /reachlibrary.jsp			
Jun 10, 2020 - Sep 12, 2021	1,471 (2.37%)	1,095 (2.57%)	0:00:37
Mar 1, 2019 - Jun 9, 2020	1,206 (1.48%)	674 (2.14%)	0:00:30
% Change	21.97%	62.46%	21.71%

TABLE III. DEVICES USED TO ACCESS MILITARY REACH

Device Category	Users	New Users
Change	176.01%	182.10%
Total Nov 2, 2019 - Jun 11, 2020	3,130	3,041
Total Jul 15, 2019 - Nov 2, 2019	1,134	1,078
desktop		
Nov 2, 2019 - Jun 11, 2020	2,112 (67.43%)	2,040 (67.08%)
Jul 15, 2019 - Nov 2, 2019	779 (68.57%)	737 (68.37%)
% Change	171.12%	176.80%
mobile		
Nov 2, 2019 - Jun 11, 2020	975 (31.13%)	956 (31.44%)
Jul 15, 2019 - Nov 2, 2019	332 (29.23%)	318 (29.50%)
% Change	193.67%	200.63%
tablet		
Nov 2, 2019 - Jun 11, 2020	45 (1.44%)	45 (1.48%)
Jul 15, 2019 - Nov 2, 2019	25 (2.20%)	23 (2.13%)
% Change	80.00%	95.65%

TABLE IV. DEVICES USED TO ACCESS MILITARY REACH

Device Category 	Acquisition		
	Users  	New Users 	Sessions 
	556.80%  9,077 vs 1,382	585.96%  8,938 vs 1,303	445.99%  14,960 vs 2,740
1. desktop			
Jun 11, 2020 - Sep 13, 2021	5,744 (63.73%)	5,695 (63.72%)	10,875 (72.69%)
Mar 1, 2020 - Jun 10, 2020	964 (69.65%)	905 (69.46%)	2,162 (78.91%)
% Change	495.85%	529.28%	403.01%
2. mobile			
Jun 11, 2020 - Sep 13, 2021	3,163 (35.09%)	3,139 (35.12%)	3,970 (26.54%)
Mar 1, 2020 - Jun 10, 2020	408 (29.48%)	386 (29.62%)	563 (20.55%)
% Change	675.25%	713.21%	605.15%
3. tablet			
Jun 11, 2020 - Sep 13, 2021	106 (1.18%)	104 (1.16%)	115 (0.77%)
Mar 1, 2020 - Jun 10, 2020	12 (0.87%)	12 (0.92%)	15 (0.55%)
% Change	783.33%	766.67%	666.67%

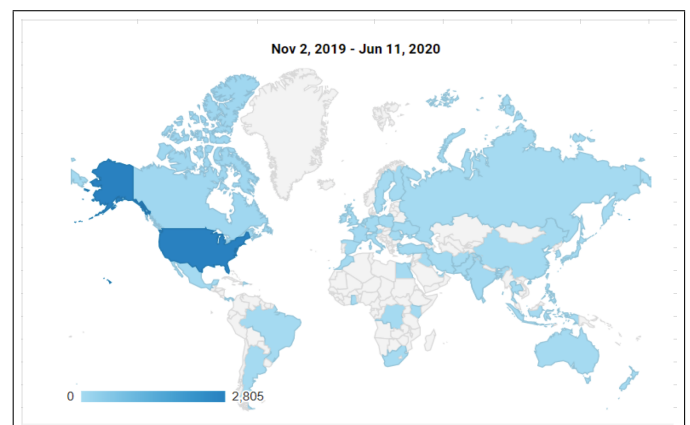


Figure 5. Map overlay about locations of users from Google Analytics.

TABLE V. LOCATIONS OF USERS FROM GOOGLE ANALYTICS 2019-2020.

	Acquisition		
	Users	New Users	Sessions
Change	175.93%	182.00%	169.62%
Total Nov 2, 2019 - Jun 11, 2020	3,129	3,040	5,565
Total Jul 15, 2019 - Nov 2, 2019	1,134	1,078	2,064
United States			
Nov 2, 2019 - Jun 11, 2020	2,804 (89.58%)	2,717 (89.38%)	5,193 (93.32%)
Jul 15, 2019 - Nov 2, 2019	1,050 (92.51%)	995 (92.30%)	1,969 (95.40%)
% Change	167.05%	173.07%	163.74%
Canada			
Nov 2, 2019 - Jun 11, 2020	88 (2.81%)	87 (2.86%)	106 (1.90%)
Jul 15, 2019 - Nov 2, 2019	4 (0.35%)	3 (0.28%)	8 (0.39%)
% Change	2100.00%	2800.00%	1225.00%
(not set)			
Nov 2, 2019 - Jun 11, 2020	29 (0.93%)	29 (0.95%)	29 (0.52%)
Jul 15, 2019 - Nov 2, 2019	48 (4.23%)	48 (4.45%)	48 (2.33%)
% Change	-39.58%	-39.58%	-39.58%
India			
Nov 2, 2019 - Jun 11, 2020	27 (0.86%)	26 (0.86%)	33 (0.59%)
Jul 15, 2019 - Nov 2, 2019	5 (0.44%)	5 (0.46%)	9 (0.44%)
% Change	440.00%	420.00%	266.67%
France			
Nov 2, 2019 - Jun 11, 2020	20 (0.64%)	20 (0.66%)	20 (0.36%)
Jul 15, 2019 - Nov 2, 2019	0 (0.00%)	0 (0.00%)	0 (0.00%)

TABLE VI. LOCATIONS OF USERS FROM GOOGLE ANALYTICS 2020-2021.

Country	Users	% Users
1. United States	8,155	90.11%
Jun 11, 2020 - Sep 12, 2021	4,389	90.94%
Mar 1, 2019 - Jun 10, 2020		
% Change	85.81%	-0.92%
2. (not set)	100	1.10%
Jun 11, 2020 - Sep 12, 2021	83	1.72%
Mar 1, 2019 - Jun 10, 2020		
% Change	20.48%	-35.79%
3. Canada	89	0.98%
Jun 11, 2020 - Sep 12, 2021	97	2.01%
Mar 1, 2019 - Jun 10, 2020		
% Change	-8.22%	-51.67%
4. China	89	0.98%
Jun 11, 2020 - Sep 12, 2021	18	0.37%
Mar 1, 2019 - Jun 10, 2020		
% Change	394.44%	163.67%
5. Philippines	60	0.66%
Jun 11, 2020 - Sep 12, 2021	12	0.25%
Mar 1, 2019 - Jun 10, 2020		
% Change	400.00%	166.63%
6. United Kingdom	58	0.64%
Jun 11, 2020 - Sep 12, 2021	19	0.39%
Mar 1, 2019 - Jun 10, 2020		

with Military REACH Android and iOS mobile apps. The paper also tries to demonstrate some of the features of Firebase for developing an Android app.

Firebase uses JavaScript Object Notation (JSON) files for storing data. The other servers use a table (rows and columns) format for storing data. There are a few cloud based servers, same as Firebase, such as AWS Mobile Hub. It is an integrated console that helps to create, build, test, and monitor the mobile apps that leverages AWS services. There is another framework called Cloud Kit- It, which is an Apple framework helping to

save data and store assets.

Military REACH uses Firebase to build and monitor data from the participants engaged with the app. In this study, our goal is to assess the usability of our articles.

A. Firebase

Firebase is a remarkable web application platform to help app developers build high-quality apps. It stores the data in JSON format which does not use query for inserting, updating, deleting, or adding data to it. It is the backend of a system that is used as a database for storing data [10].

Firebase available services are:

1) *Firebase Analytics*: It provides insight into app usage, similar to Google Analytics. It is a paid app measurement solution that helps in providing user engagement data. This main feature allows the application developer to understand how users are using the application. The Software Development Kit (SDK) has the feature of capturing events and properties on its own and also allows getting custom data.

Figure 6 represents Military REACH user engagement data including 205 active users and 34 minutes average engagement time. As presented in Figure 7, most of the participants were from United States.

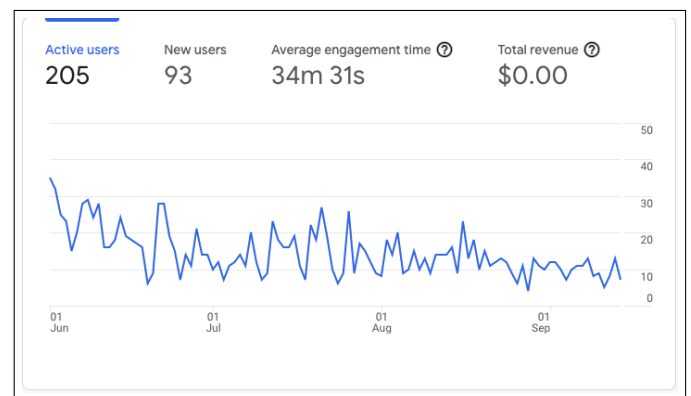


Figure 6. Acquisition overview, June 1 - September 13

Country	+ Active users	New users	Engaged sessions	Engagement rate	Engaged sessions per user	Average engagement time
Totals	205	93	1,525	76.98%	7.44	34m 31s
	100% of total	100% of total	100% of total	Avg 0%	Avg 0%	Avg 0%
1 United States	204	93	1,516	77.07%	7.43	34m 29s
2 Germany	1	0	3	75%	3.00	1m 29s
3 Israel	1	0	0	0%	0.00	0m 06s
4 Peru	1	0	1	100%	1.00	2m 20s
5 Puerto Rico	1	0	1	100%	1.00	7m 48s
6 Spain	1	0	4	57.14%	4.00	28m 56s
7 United Kingdom	0	0	0	0%	0.00	0m 00s

Figure 7. Location overview, June 1 - September 13

2) *Firebase Cloud Messaging (FCM)*: FCM is a paid service which is a cross-platform solution for messages and notifications for Android, Web Applications, and IOS. Military REACH uses FCM to notify users whenever a new article is available to them to review.

3) *Firestore Authentication*: Firestore Authentication supports social login provider like Facebook, Google GitHub, and Twitter. It is a service that can authenticate users using only client-side code and it is a paid service. It also includes a user management system whereby developers can enable user authentication with email and password login stored with Firestore [10].

4) *Real-time Database*: Firestore provides services like a real-time database and backend. An API is provided to the application developer allowing application data to be synchronized across clients and stored on Firestore's cloud. The client libraries are provided by the company which enables integration with Android, IOS, and JavaScript applications.

5) *Firestore Storage*: It facilitates a secure file transfer regardless of network quality for the Firestore apps. It is integrated with Google Cloud Storage which is cost-effective object storage service. The developer can use it to store a variety of data types such as images, PDFs, and videos.

6) *Firestore Notifications*: It enables targeted user notifications for mobile app developers and the services are freely available.

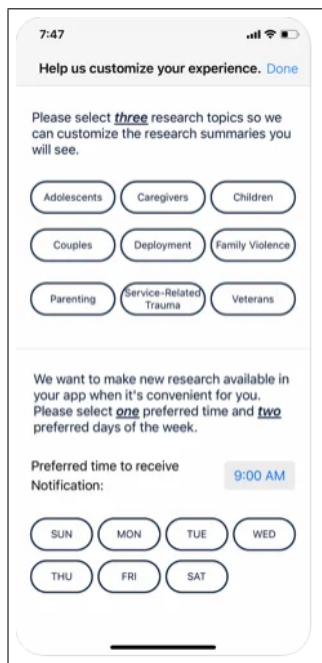


Figure 8. Military REACH App Category Selection

VII. CONCLUSION AND FUTURE WORK

The Google Analytics results helped Military REACH to analyze their website's usage to better serve military families. It shows that after adding more features to the search functions, users are interacting with the website in practical ways and spending more time on the website. Compared to the first two years, website usage almost tripled last year.

According to the Google Analytics results, 31% of users have access to the website through their phone. In response, to facilitate the accessibility of Military REACH resources, the team created a mobile application (app).

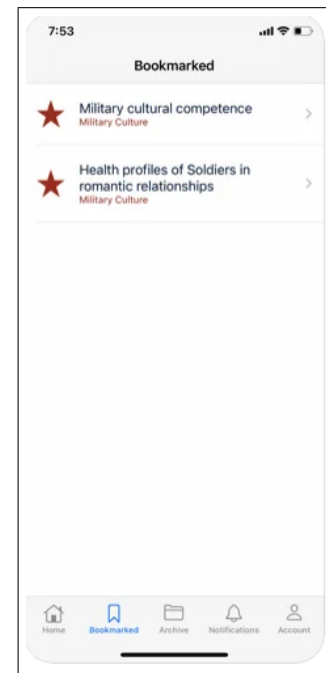


Figure 9. Military REACH App Home page

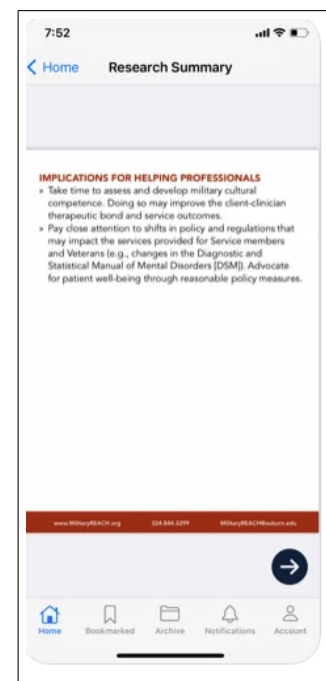


Figure 10. Military REACH Articles Format

Figure 11. Military REACH App Surveys

In the future, Military REACH plans to conduct a pilot test of a newly developed mobile app that will be used for the dissemination of REACH reports, mainly Translating Research Into Practice (TRIP) reports. The team will conduct an efficacy study to examine the impact of our mobile app and TRIP reports specifically for helping professionals who directly serve military families. Survey data will be collected from participants (i.e., primary data collection) using Qualtrics (a survey software used at Auburn University), a secure online data collection tool. This data will help us understand the users' military family knowledge better, their confidence in serving military families, their satisfaction and reaction to the app, and make the military family research accessible to everyone.

REFERENCES

- [1] F. Jamshidi, A. Jariwala, B. Bhattarai, K. Abbate, D. Marghitu, and M. Lucier-Greer, "Building a web-based environment to support sponsored research and university-wide collaborations," WEB 2020 : The Eighth International Conference on Building and Exploring Web Based Environments, Sep. 2020.
- [2] L. Nichols, K. Abbate, C. W. O'Neal, and M. Lucier-Greer, "Mobilizing family research: Evaluating current research and disseminating practical implications to families, helping professionals, and policy makers," Southeastern Council on Family Relations Conference, Jul. 2019.
- [3] H. R. Tibbo, "On the nature and importance of archiving in the digital age," *Adv. Comput.*, vol. 57, Jan. 2003, pp. 1–67.
- [4] K. Russell, "Digital preservation and the cedars project experience," *New review of academic librarianship*, vol. 6, no. 1, Apr. 2000, pp. 139–154.
- [5] S. Ross and M. Hedstrom, "Preservation research and sustainable digital libraries," *International journal on digital libraries*, vol. 5, no. 4, Apr. 2005, pp. 317–324.
- [6] A. Deshpande, A. Göllü, and L. Semenzato, "The shift programming language and run-time system for dynamic networks of hybrid automata," in *Verification of Digital and Hybrid Systems*. Springer, Jun. 2000, pp. 355–371.

- [7] E. A. Song, "A process evaluation of a web-based mental health portal (walkalong) using google analytics," *JMIR mental health*, vol. 5, no. 3, Jul. 2018, p. e50.
- [8] D. J. Clark, D. Nicholas, and H. R. Jamali, "Evaluating information seeking and use in the changing virtual world: the emerging role of google analytics," *Learned publishing*, vol. 27, no. 3, 2014, pp. 185–194.
- [9] E. A. Vona, "A web-based platform to support an evidence-based mental health intervention: lessons from the cbits web site," *Psychiatric Services*, vol. 65, no. 11, Jan. 2014, pp. 1381–1384.
- [10] C. Khawas and P. Shah, "Application of firebase in android app development-a study," *International Journal of Computer Applications*, vol. 179, no. 46, 2018, pp. 49–53.

FracBots: The Next IoT in Oil and Gas Reservoirs

Abdallah A. Alshehri, Klemens Katterbauer

EXPEC Advanced Research Center

Saudi Aramco

Dhahran, Saudi Arabia

abdullah.shehri.8@aramco.com, klemens.katterbauer@aramco.com

Abstract— Fracture Robots (FracBots) technology is a game-changing technology that has been developed to revolutionize upstream operations. FracBots are magnetic induction (MI)-based wireless sensor nodes that have the inter-node wireless communication, sensing and localization estimation capabilities. FracBots are miniature devices that can operate as wireless underground sensor networks (WUSNs) inside hydraulic fractures to collect and communicate important data and generate real-time mapping. A large number of FracBots is deployed to establish FracBot-to-FracBot connectivity, making the technology the first IoT (Internet of Things) to generate and exchange data inside the reservoir without human intervention. In addition, a novel artificial intelligence (AI) framework is designed for the real-time sensor selection for subsurface pressure and temperature monitoring, as well as reservoir evaluation. The framework encompasses a deep learning technique for sensor data uncertainty estimation, which is then integrated into an integer-programming framework for the optimal selection of sensors to monitor the reservoir formation. The results are rather promising, showing that a relatively small numbers of sensors can be utilized to properly monitor the fractured reservoir structure.

Keywords- *Wireless underground sensor network; magnetic induction communication; FracBot network; 4IR; artificial intelligence; formation evaluation; robotics; reservoir mapping.*

I. INTRODUCTION

Sensing deep in the reservoir has always been a major objective to enhance reservoir formation understanding and optimize the recovery from the reservoir. In the early days of the oil and gas industry, determination of reservoir formation properties was based on assumed geological formations and structures encountered on the surface [1]. Furthermore, retrieved rock cuttings assisted in getting a better understanding of the reservoir formation, however, this information is limited to a small area and may not be representative of the reservoir formation as a whole or taking into account the heterogeneity in the reservoir. Another challenge for mature reservoirs is to determine the sweep efficiency in the reservoir, where besides production information and some surface reservoir monitoring, such as seismic or electromagnetics, there is no overall in-situ reservoir monitoring system available [2, 3]. As the reservoirs are dynamic, permanent monitoring of the reservoir is crucial to determine the saturation flow and the fracture channels. Hence, an in-situ monitoring of the reservoir becomes quintessential in order to overcome the

existing challenges of limited information away from the wellbores.

The 4th industrial revolution (4IR) has become a major transformer of the upstream petroleum industry. Major advances were already achieved in enhancing production, performing real-time monitoring of wells and reservoirs and also forecast potential reservoir risks and workover requirements [4, 5, 6]. Several advances were also achieved in performing maintenance and installation operations remotely via the help of 4IR technology [7]. The main objective is to improve productivity and cost-effectiveness of the operations, as well as enhance safety. This allows to conduct maintenance in a much shorter time period and also allows to conduct the operations around the clock.

Enhancing production from and monitoring reservoirs are critical components for ensuring the effectiveness of oil and gas operations and maintain its sustainability. For this, sensing is an essential area that allows to monitor the reservoir in real-time and investigate its evolution. Continuous sensing further allows monitor the behavior of a reservoir over time and forecast its future production potential. Conventional surface sensing covers an extensive area of the reservoir. However, the resolution and challenge connected to the multiple solutions of the inverse problem represent a significant problem. The challenge arises primarily from the lack of direct measurements and observations in the reservoir. Furthermore, challenges arising from placing large measurement equipment downhole for an extensive period of time may render this approach. While surface sensing enables to cover an extensive area and deduce easier the correlations between different measurements, as well as the causes and effects, subsurface sensing operations are significantly more challenging. This is due to the lack of direct measurements and observations of the reservoir structure and formation, as well as challenge to place measurement equipment downhole [8,9]. In order to overcome this challenge related to the lack of direct measurements, a more direct approach to sensing in the form of subsurface reservoir sensors is essential.

Miniaturized downhole sensors have been developed in recent years, allowing to achieve permanent downhole sensing that is both robust and efficient [9, 10]. Reference [11] presented a temperature insensitive pressure sensor based on fiber-optics that has a size of only 125 micrometers. The authors demonstrated the ability to measure pressure levels over a significant range with minimal temperature

effects, which may make these sensors applicable for downhole sensing. Similarly, reference [12] presented a fiber-optic FabryPerot gas refractive index sensor for high temperature applications. The miniaturized sensor allows to measure up to 800 degrees Celsius, outlining the feasibility of high temperature permanent downhole monitoring with low power consumption.

In general, microseismic and tiltmeter surveys are ones of many technologies available to characterize reservoir hydraulic fractures but they are expensive, approximate, and time consuming. Moreover, they are conceptual approaches that do not unfortunately provide useful information about the inner workings of hydraulic fractures. However, reference [10] presented innovative wireless sensors for the mapping of hydraulic fractures in subsurface reservoirs. The results outline the ability to accurately map fractures with a hybrid solution of electromagnetic and magnetic conduction wireless communication in order to overcome excessive path losses within the reservoir environment. Communication losses between the sensors represent a major challenge in addition to the power requirements of the sensors, requiring that there is sufficient proximity between the wireless sensors such that the data is adequately transmitted. These advancements lead to the feasibility of downhole sensing in the reservoir with data transmission being conducted wirelessly [6]. Powering these downhole sensors for long period to maximize the sensing duration in the downhole environment is a major challenge. All sensors do not require to operate at the same time due to the connectedness of the reservoir and partial redundancy of the downhole sensors. This operational feature helps to achieve the objective of maximizing data acquisition while minimizing overall power consumption. However, this objective leads to the problem of selecting the minimal number of sensors while achieving the target objective of the most accurate downhole sensing. These selection schemes can typically be classified in coverage schemes, target tracking and localization schemes, single mission assignment schemes and multiple missions assignment schemes [7]. Coverage schemes are selection schemes that ensure the sensing coverage of the location or the targets of interest, while target tracking and localization schemes focus on the selection of sensors for target tracking and localization purposes. The mission assignment schemes focus on the selection of sensors for a single or multiple mission that have to be accomplished.

In this work, we review the FacBot technology and demonstrate a novel intelligent sensor selection framework for the optimization of sensor selection in real-time for flow and fracture monitoring. We generated a platform for FracBot development including software and hardware elements. To this end, we have contributed in five areas as follows: first, we developed a novel cross-layer communication framework for MI-based FracBot networks in dynamically changing underground environments, and thoroughly modeled the efficiency and performance of the network. Second, we developed a novel magnetic induction (MI)-based

localization framework that exploits the unique properties of the MI field to determine the locations of the randomly deployed FracBot nodes in hydraulic fractures. Third, we developed an accurate energy model framework of a linear FracBot network topology that gives feasible FracBot transmission rates while respecting the constraints of a realistic energy harvesting paradigm. All together, these elements demonstrate that important new capabilities including 3D mapping of a hydraulic fracture and on-going measurement of reservoir parameters in-situ are possible using wireless underground sensor networks (WUSNs). Fourth, we designed, developed, and fabricated MI-based FracBot nodes. To validate the performance of our solutions in our produced prototype of FracBot nodes, we developed a physical MI-based WUSN testbed. Finally, we develop a novel intelligent sensor selection framework for the optimization of sensor selection in real-time for flow and fracture monitoring. The objective of the framework is to maximize longevity of the operations while maintaining measurement accuracy and flow detection ability.

II. FRACBOTS SYSTEM

A typical oil reservoir environment with a hydraulic fractures has been described in Figure 1 displaying the tentative placement of the FracBots. The research challenges of current wireless sensor networks (WSNs) are addressed to position wireless underground sensor nodes (FracBots) in cracks during the hydraulic fracturing procedure in order to be capable to work efficiently in underground settings. A short system lifetime, trouble in launching wireless signals, and high path loss are included in these challenges [13]

The structure design of the MI-based FracBot network has been illustrated in Figure 1, which has two layers:

- FracBot (sensor nodes): They are small nodes placed into the fracture throughout the hydraulic cracking process. The nodes positions are roughly uniform and linear inside the fracture because the fracture is extremely narrow. The FracBots are wireless nodes that have powerless source, but they are charged from EM radiation transferred wirelessly from the base station located at the wellbore.
- The base station: It is made up of a big dipole antenna at the wellbore, is linked to an above-ground connection.

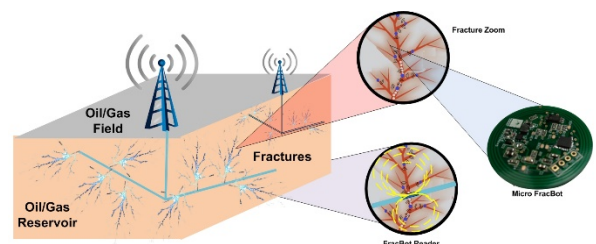


Figure 1. The structure of the FracBots network.

A. FracBot Architecture

FracBots are active micro-wireless sensors injected inside the hydraulic fracture during the hydraulic fracturing process. The FracBot node is furnished with a processor, a transceiver, an antenna, a sensing unit and a harvesting unit. It harvests energy transmitted from the base station, which permits it to execute sensing tasks and to wirelessly communicate collected data back to the base station using MI-based communication.

B. Network Architecture

Fractures dimensions are nominally millimeters wide and some meters high, can reach up to 100 m long. The FracBots are assumed for current purposes to be almost static and uniform in the fractures. Therefore, a static network scheme for the FracBot system in the fracture is envisioned as described in Figure 2. This indicates that energy is transmitted and collected in a single-hop energy method while sensed data is communicated in a multi-hop mode. We suggest a three-stage operational arrangement based on the structure design described earlier.

1) *A single-hop emitted energy phase:* The base station releases energy through a crack and communicates with the FracBot sensors. The base station is situated at the wellbore and provided with high power communication antenna which permits the use of low frequency RF to emit EM waves and transmit the energy via the fracture environment to the MI-based FracBots spread out in the hydraulic fracture.

2) *A multi-hop MI-based transmission phase:* The FracBots gather essential energy through harvesting, sense related reservoir parameters, and use the MI communication technique to communicate quantities to the nearby neighbor sensor, and by successive repeating, the uplink with the multi-hop communication path is utilized to communicate the information to the base station.

3) *A backbone communications phase:* In this phase, the base station collects the sensing information from the FracBots in the fracture and then sends the information via an aboveground gateway.

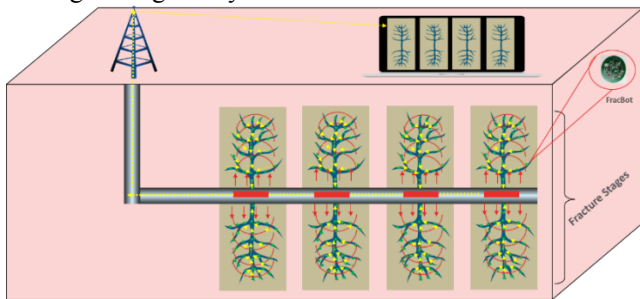


Figure 2. The FracBots network.

III. WIRELESS FRACBOT NETWORKS ENERGY

A wireless channel model in hydraulic fracture is described for both MI communications and energy transmission. The suggested FracBot network comprises of two types of channels described as follows:

A. Downlink Wireless Channel Model

To radiate energy and communicate information to the FracBots in the fracture, the base station antenna emits EM waves at low MHz frequency. The EM waves are affected by harsh environment, and numerous fluids including oil/gas and water in the fracture. The key ingredients surrounding the fracture are reservoir rocks, as displayed in Figure 1. Thus, the fluids and substances influence the downlink path loss as in Eq. (1) [14].

$$L_{DL} = 10 \log \left(\frac{N^2 \omega^2 \mu_2^2 l^2 r^4 k_2^2 \sin^2 \theta}{64 R_c R_i d_{DL}^2} \left[\frac{1}{k_1^2 d_{DL}^2} + 1 \right] e^{-2d_{DL}/\delta} \right) \quad (1)$$

Where θ is the angle of the coil positions, N is the coil number turns, R_c is the resistance of the coil antenna, and r is the radius of the coil. k_1 , k_2 are the wavenumbers inside and outside the fracture, l is the length of the base station antenna, δ is the skin depth inside the fracture, R_i is the input resistance of the base station antenna, μ_2 is the reservoir and rocks effective permeability, w is the angular frequency, and d is the distance between the base station and the FracBot. We use the following values throughout this paper. The reservoir rock has similar to that of air (i.e., $\mu_2 = \mu_0 = 4 \times 10^7$ [H/m]). As explained later using magnetic permeability, the permeability μ_1 inside the fracture, if occupied with magnetic proppants, is assessed in Eq. (3). We used the following parameters to calculate the permeability. The ratio of p_{para} and p_{ferro} are 30% and 10%, respectively, the proportionality constant \hat{c} is 0.993, and the magnetic susceptibilities χ_{ferro} is $\chi_{Fe3O4} \approx 5 \times 10^{-4}$ for temperatures under 853[K]. The material employed to yield the high- μ proppants can regulate this effective permeability. The effective permittivity inside the fracture is set to be $\epsilon_1 = 3.5\epsilon_0$ (crude oil) while the permittivity of the matrix / reservoir and rock is set to be $\epsilon_2 = 2\epsilon_0$ (sand and clay mixture). If we primarily suppose absolute oil production, the conductivity outside the fracture is set to be $\sigma_2 = 0.001$ S/m, while the effective conductivity in the fracture is low, on the order of $\sigma_1 = 10^{-4}$ S/m. A base station transmitting power of 50 watts with 20 m dipole antenna are used. $R_i = 75\Omega$ is the input resistance. The operating frequency is 10 MHz for the antennas (the dipole and the coils), 5 mm radius and 10 as the number of turns of the coils. The coil resistance is $R_c = 0.2 \Omega$. The minimum received power is $P_r = -100$ dBm and the converting rate of the energy at the FracBot sensor is $\eta = 80\%$.

Figure 3 illustrates the power received at FracBots as a function of the distance between the base station and the FracBots in the hydraulic fracture. The energy transfer framework displays the received power by the FracBots. It indicates that the energy model can overcome the hydraulic fracture environment restrictions. For instance, at a distance of 30 m from the base station, the received power is about -50 dBm, it is adequate to power the very low power wireless FracBots.

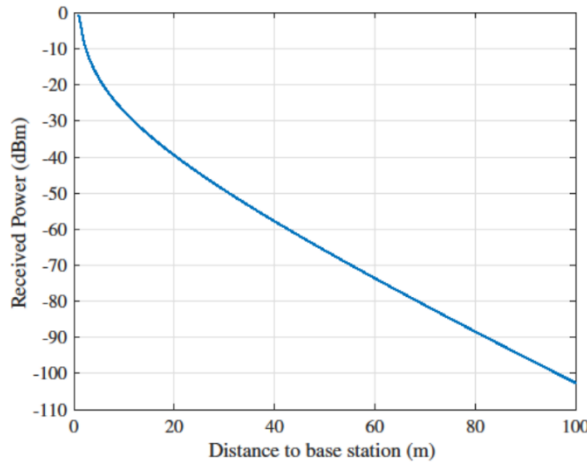


Figure 3. Power received by FracBots from the base station.

B. Uplink MI Channel Model

To send and transmit collected data by the FracBot sensors to the base station in the multi-hop mode, the uplink channel between two adjacent FracBots is employed as presented in Figure 2. The MI technique, to propagate signals and accomplish constant channel settings through the small size of the coils, utilize the near magnetic field of coils. MI communication is extremely appropriate for underground environments. The distinctive MI-based channel formed in the fracture medium is covered by the uplink channel capacity. Reference [15] attains this capacity:

$$C_{UL} = f_U \log_2 \left[1 + \frac{2\pi^2 P_t M^2 f_U^2}{R_c^2 N_{noise}} \right] - f_L \log_2 \left[1 + \frac{2\pi^2 P_t M^2 f_L^2}{R_c^2 N_{noise}} \right] \quad (2)$$

Where f_L is the lower frequency of the channel bandwidth, f_U is the upper frequency of the channel bandwidth, N_{noise} is the noise power, and P_t is the transmission power. This uplink channel capacity demonstrates the impacts of the hydraulic fracture environment to calculate a feasible data rate via the MI-communication link among the FracBot nodes.

Through the intermediate FracBot nodes, a multi-hop route forms between the FracBot nodes transmitter and the base station. A magnetic field is created between the transmitter and receiver coils, as proposed in [16]. The quality of the MI communication is impacted by the magnetic permeability of the medium which is the key environmental element. The resistance of copper coil will alter with respect to the variable temperatures in hydraulic fracture, particularly, while the permeability of matrix and water is similar to that of air (i.e., $\mu_0 = 4 \times 10^{-7}$ [H/m]) at room temperature. Depending on the composites of the underground magnetic content, the medium permeability also behaves differently. The effects of medium permeability and temperature are governed as [16]:

$$\mu = \mu_0(1 + x) = \mu_0 \left(1 + p_{para} \frac{\hat{c}}{T} \right) + p_{ferro} x_{ferro} \quad (3)$$

$$R = 2\pi r N R_0 [\alpha_{Cu}(T - T_0)] \quad (4)$$

Where, μ_0 is the air permeability, R is the coil resistance, χ and χ_{ferro} are the magnetic susceptibilities of the medium and ferromagnetic contents, respectively. \hat{c} is a constant, p_{ferro} and p_{para} are the ratio of ferromagnetic and paramagnetic composites, respectively, T [K] is the actual hydraulic fracture temperature, T_0 [°K] is the room temperature, $\alpha_{Cu} = 3.9 \times 10^{-3}$ [K] is the copper coil's temperature coefficient and R_0 [Ω/m] is the resistance of a unit length of coil at room temperature. Stokes theorem is used to obtain the self and mutual inductance is analytically.

$$M(T, \sigma) = \frac{\mu \pi N^2 r^4 \delta(d, \sigma) \cos \theta}{4d_{BL}^3} \quad (5)$$

Where, $\delta(\cdot, \cdot)$ is attenuation caused by the skin depth effect and σ [S/m] is the medium conductivity. Between the two MI transceivers, the path loss of MI communication can be described as

$$L_{UL}(d, f_0, \theta, T, \sigma) = \frac{2(2R^2 + \omega_0^2 M^2)}{\omega_0^2 M^2} \quad (6)$$

Thus, the estimated uplink channel bandwidth is achieved by

$$B_{UL}(T, \sigma) = \frac{R(\sqrt{2} - 1)}{\mu \pi^2 r N^2} \quad (7)$$

The lowest transmitting power amount needed to facilitate inter-communication among FracBots over the MI-based channel in hydraulic fracture is displayed in Figure 4. The required transmission power rises dramatically as the distance between the two FracBot nodes rises, as a result of the complex transmission medium. To assure the MI-link quality, this distance must be optimized. The path loss and the frequency response of MI channels at different temperatures in the hydraulic fracture environment is exhibited in Figure 5. The path loss rises, when the operating temperature and the transmission range rise, resulting in degradation of the quality of the communication link.

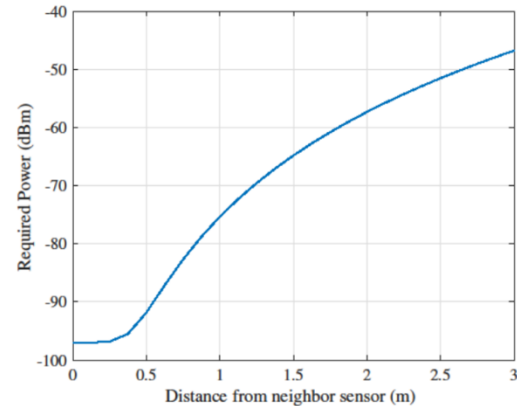


Figure 4. Required power to transmit data from FracBot to neighbor FracBot.

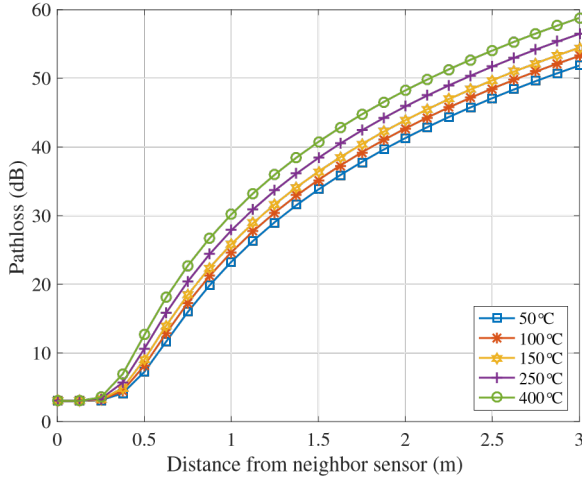


Figure 5. Path loss of magnetic induction at different hydraulic fracture temperatures.

C. Energy Consumption and Energy Harvesting Model

To charge the whole FracBot network, the downlink energy charging functions in one-hop fashion. The size of FracBot nodes is very minor which restrict the battery capacity due to the very narrow fracture. Hence, the very low size battery is not able to keep sufficient power for the FracBots to operate the communication and conduct sensing tasks. Due to this limitation, to store the harvested energy for the FracBot operations, the battery is replaced by ultra-capacitor. Accordingly, as the size of sensed information transmitted by FracBots is determined by the collected energy, it is essential to acquire precise energy model for charging and consumption process. To model the energy harvesting from the base station installed in the oil well, the recent results for an energy transfer model were implemented [16]. As a function of the distance from the Base station to a particular FracBot, the equivalent path loss can be calculated for the downlink channel by [14]:

$$E_i^h = T_i^c \eta_i P_{TX} L_{DL} \left(\sqrt{l^2 + \left(\sum_{j=1}^n d_j \right)^2} \right) \quad (8)$$

Figure 6 shows the collected energy over the distance between the base stations and the FracBots in the hydraulic fracture in a one-hour charging time. The power received by the FracBot nodes overcoming the hydraulic fracture conductivity constraints is revealed by the wireless energy charging model. For example, the harvested energy is around -10 dBmJ at a 25 m distance from the base station that is sufficient to charge the very low power MI-FracBots.

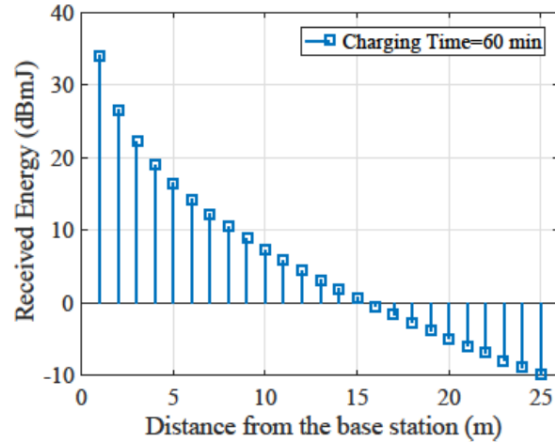


Figure 6. Harvested energy in FracBots network.

IV. FRACBOT FUNCTIONALITIES

The basic functions have been developed. First, we have developed an innovative cross layer communication model for Magnetic Induction networks in altering underground environments, coupled with selections of coding, modulation and power control and a geographic forwarding structure. Second, we have developed an innovative MI-based localization framework to capture the locations of the randomly deployed FracBot nodes by exploiting the exceptional properties of the MI-field.

A. Environment-Aware Cross-layer Communication Protocol

We present a distributed cross-layer framework for MI-based WUSNs [18]. A cross-layer framework is recommended for WUSNs in oil reservoirs as an alternative of taking the classical layered protocol method which is the 7-layer Open Systems Interconnection model (OSI Model). To improve MI communication in WUSNs, it is executed in a distributed manner to jointly enhance the communication functionalities of different layers. Our solution attains optimal energy consumption and high throughput efficiency with low computational complication, and also fulfills the quality of service (QoS) requirements of diverse applications. These properties qualify our solution as a valuable for practical applications. The cross-layer solution framework includes the following:

- 1) Evaluation for the major environment facts of underground reservoir affecting the transmission qualities of MI-based communication.
- 2) Three-layer protocol stack for WUSNs in oil reservoir.
- 3) Cross-layer framework to conjointly enhance communication functionalities of various layers.
- 4) Distributed Environment-Aware Protocol (DEAP) proposal to realize the projected cross-layer framework.

Figure 7 demonstrates the protocol stack for environment-aware cross-layer protocol design and its key contributions. Firstly, the distributed cross-layer framework accounts for

environment information of oil reservoirs that influences the MI-based communications qualities. MI channel models are established to consider the effects of the physical layer functionalities. The effects of temperature, electrical conductivity, magnetic permeability, and coil resistance are studied. This is to capture their effects on the MI-communication parameters, such as the path loss, the bandwidth, and the interference. Second, the protocol stack consists of three-layer stacks: a data link layer, a network layer and a physical layer. The communication functionalities for each layer of a protocol stack are recognized, for example, medium access control (MAC), routing algorithms, modulation and forward error coding, and the statistical quality of service (QoS) comprising of transmission reliability and packet delay. These parameters are analyzed to find out their effects on MI-based communications. Third, the proposed cross-layer framework addresses all functionalities of each protocol layer. To optimize MI communication in WUSNs, it is executed in a distributed manner to jointly optimize the communication functionalities of various layers. Finally, DEAP is recommended to comprehend the cross-layer framework and solve its optimization problem in a disseminated manner. The DEAP process comprises a distributed power control, an evaluation of a multiple access scheme for a data link layer and a two-phase decision process for executing a routing algorithm for the network layer.

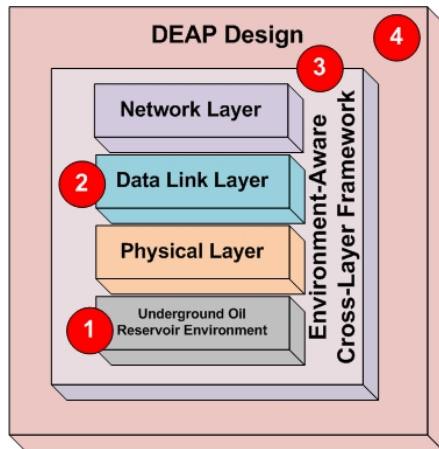


Figure 7. Protocol stack of environment-aware cross-layer protocol design.

Thus, the DEAP achieves both optimal energy savings and throughput gain concurrently for practical application and provides statistical QoS guarantee. Evaluation findings indicate that cross-layer framework outclasses the layered protocol solutions with 6 dB throughput gain and 50% energy savings. Furthermore, the distributed framework comprises of two-rounds per node decisions that involves single-hop neighbor data and has uncomplicated computation process. As a result, consistent and effective communication is recognized by the distributed cross-layer design for MI communication in the challenging underground environments.

B. FracBots Localization Framework

We introduce a MI-based localization for FracBots in the hydraulic fracture [19]. We suggest an innovative MI-based localization solution, which uses the spinoff of magnetic induction communication (received magnetic field strength (RMFS)) and the promising features of MI channel. By using RMFS, it guarantees the accuracy, simplicity, and ease of the localization scheme. MI-based communication is very appropriate for oil reservoirs due to its distinctive multi-path and fading-free propagation features. Unknown sensor locations are provided by the MI-based localization in randomly-deployed wireless sensor systems in underground environments. By capitalizing on the unique features of the magnetic induction communication including fading-free and multi-path propagation features, it generates approximate distances, between two neighboring nodes and between nodes and base stations, with very accurate RMFS measurements. Our solution develops an MI-based localization framework to integrate Weighted Maximum Likelihood Estimation (WMLE) and Semidefinite programming (SDP) relaxation techniques to generate very accurate localization in underground environments. It mutually applies both fast initial positioning and fine-grained positioning to attain high positioning precision in WUSNs to provide a rapid and precise positioning in different noise systems (low and high) while sustaining high computational efficiency under various underground environment situations. Our localization framework is summarized as follows:

- 1) RMFS measurements for designing localization in hydraulic fracture.
- 2) Localization framework for WSNs in hydraulic fracture.
- 3) Quick early positioning by varying Direction Augmented Lagrangian Method (ADM).
- 4) High resolution positioning from Conjugate Gradient Algorithm (CGA).

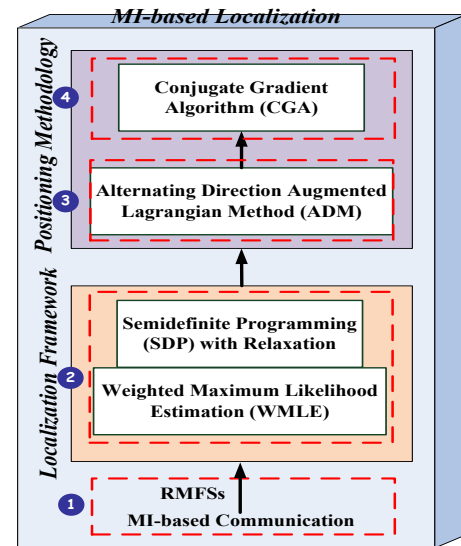


Figure 8. MI-based localization system.

Figure 8 displays the structure of MI-based localization system. The first step is to attain the approximate distance from received magnetic field strengths (RMFSs) via the developed channel models. Next, the localization framework is formulated as the problem creation of combined WMLE and SDP reduction for precise FracBot positioning from noisy distance estimations. Third, an efficient initial positioning is gained from a fast algorithm, called ADM, to provide approximate but useful location results. According to the initial results, a fine-grained positioning obtained from the powerful Algorithm (CGA) is finally fed to improve localization accurateness in a time-efficient way.

V. FRACBOT NODES AND TESTBED

The key component of WUSN is the sensor node; mainly in reservoirs monitoring and hydraulic fracture mapping. Thus, we develop a miniaturized FracBot node to validate the feasibility and capability of using MI-based communication in underground environments. Particularly, we design and realize a FracBot node that can be used to gather useful data about hydraulic fracture such as temperature, pressure, chemistry composition and other variables. The FracBot is designed based on major electronic components including Microcontroller (MCU) and RFID/NFC chip. This chip launches the communications among the FracBots using Near Field/MI-based technique. The key design concepts are:

- 1) Low energy requirements (feasibility and implementation in aggressive environments).
- 2) MI communication (RFID/NFC technology with passive/active sensors).
- 3) Multi-purpose FracBots (support several sensing applications).
- 4) Hardware miniaturization (hardware is designed in small footprint).

To implement these key design concepts, we create a design roadmap to proficiently develop the FracBot node in terms of hardware and software as described in Figure 9.

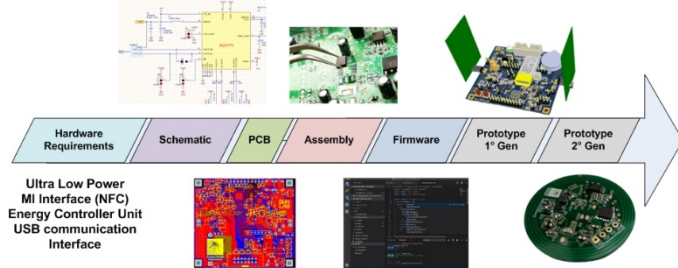


Figure 9. Roadmap of the FracBot design.

The roadmap skeletons the steps of design after determining the idea and requirements. Component selection is a broad process, requiring picks from a wide-range of available products, and it directives how the remaining phases proceed. Prototyping and software development is extremely constrained, encompassing the development of a model sensor

node and associated software. Prototype design is first achieved in a schematic diagram and then as a printed circuit board. Then, the firmware and software are executed. After this stage, a completed circuit has been prepared to the final step which is testing and verification.

Restricted characteristics are essential for designing an effective node that withstands operations in severe environments with high temperature, and pressure, high path loss and limited energy. Moreover, to improve every component based on their requirements, the very small size is needed as it can protect development time, board space, and cost. The key features of our proposal are a long operating time, ultra-low power, an efficient communication layer, a processing function, and sensing capabilities and energy-harvesting. The concurrent employment of all five characteristics allows the node to operate in a perpetual powered status. The FracBot node will encompass mainly a microcontroller, a temperature sensor, an energy harvesting unit and a transceiver. The feasibility of energy harvesting will be exhibited using this FracBot node.

A. FracBot node design and development

The design and development of FracBot node are based on near field communication (NFC) for a physical layer coupled with an energy collecting feature and very low power requirements [20]. Two types of FracBots are created: a FracBot active node and a FracBot passive node. Figure 10 demonstrates the active FracBot prototype, which entails of a microcontroller, an energy management unit (EMU), USB communication, a temperature sensor, a NFC transceiver (passive and active), and a super-capacitor.

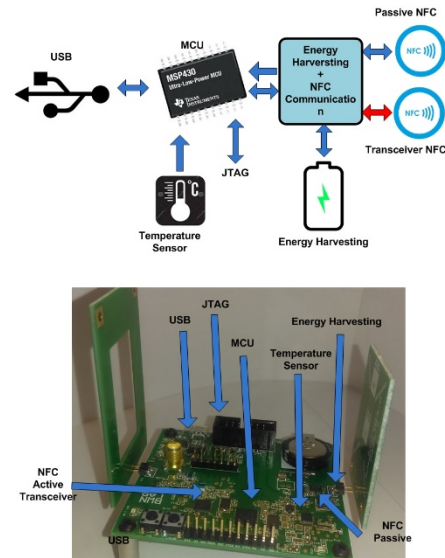


Figure 10. Block diagram and prototype of the FracBot active node.

The FracBot active node has sophisticated functions and consumes minimal energy since the FRAM technology has been exploited. The microcontroller features used in the active node are very low energy, a high processing speed, and several

interfaces. In addition, Figure 10 shows a block diagram of the active node featuring the interconnection block of the node, which comprises of the microcontroller, the JTAG interface, the energy harvesting circuit, USB communication, the temperature and the MI transceiver. The JTAG interface permits us to program and access all variables of the code and stop the code from running at a pre-defined point (breakpoints).

The FracBot passive node is a passive node that does not have a transceiver but a transmitter only relaying the data to the active node. Its prototype and diagram are shown in Figure 11. It comprises of the microcontroller, the temperature sensor, the USB interface, and the NFC active tag. The NFC transceiver of the active node can access, through the established link, the NFC tag memory, change the configurations of the nodes and generate energy by harvesting energy output. As shown in the block diagram, the node is capable to launch a bidirectional communication with RFID/NFC transceiver.

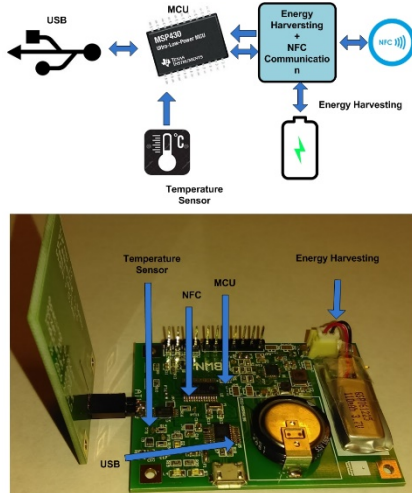


Figure 11. Block diagram and prototype of the FracBot passive node.

B. FracBot Node Software/ Firmware

Firmware is a special type of computer software used to control components hardware of electronic devices at low-level. Low power firmware is categorized by the capability to switch between active and low-power modes with the guarantee of the functionalities and operation continuation. This feature contributes in significant energy reduction on microcontroller unit (MCU). The software is optimized according to the advanced control of MCU and all peripherals. The most advanced microcontroller considers efficient power control, instant wakeup, intelligent autonomous peripherals and interrupts in its operation. Inefficient firmware codes are not preferred since they slow the function and require a lot of energy. There are many examples of inefficient firmware properties such as software delay loop, uninitialized ports and data format conversions. Other example is math operation set as division and floating-point operations which could cause critical operation issues. To avoid such issues, the MCU pins

requisite to be configured with correct function to moderate the energy waste. To avoid software delay loop, a timer is required in interval mode configuration to enable the MCU to enter the sleep mode during the interval time. This help the MCU to not run at maximum power during the interval time. Division and floating-point operations require large computational efforts which consume a lot of the processing time and big part of the memory. To avoid that, the math operations can be configured at fixed point [21]. The design of the FracBot nodes incorporates advanced energy strategies to optimize the energy consumption based on the energy availability. It also employs very low energy profile to balance between the hardware and the software/firmware in all components operation. Furthermore, using ULP tools and energy tracer permit the development of efficient codes [21].

C. FracBots Performance Evaluation

After thorough studies have been theoretically conducted, little work has been devoted to evaluate a sensor node (FracBot) in underground-like environments to validate the theoretical results. Toward this end, we design and implement an experimental testbed simulating a reservoir environment that comprises of numerous media such as air, sand, water, and stone with few FracBot nodes as demonstrated in Figure 12. One of the crucial outcomes is that the performance of the FracBot is influenced by sand and stone media. They reduce the energy transfer, and eventually harm MI signal propagation. Hence, the evaluation of hardware enables the designers to apprehend the challenges, enhance the electronic design and minimize essential assets to reduce the hardware size.

1) FracBot Propagation Evaluation:

The FracBot MI propagation is evaluated at the operating frequency of 13.56 MHz. The investigations are done according to the received power measured using a signal analyzer. We also examine the MI field produced by the transceiver with and without modulation. In addition, we examine magnetic induction signal propagations in the air. We measure and study the effect of the antenna alignment on the received power. Figure 13 shows the schematic of the experimental arrangement and the real setup in the laboratory. In this scenario, MI interaction is measured at distances between 0 and 25 cm and angles of 0, 30°, 60° and 90°, respectively.

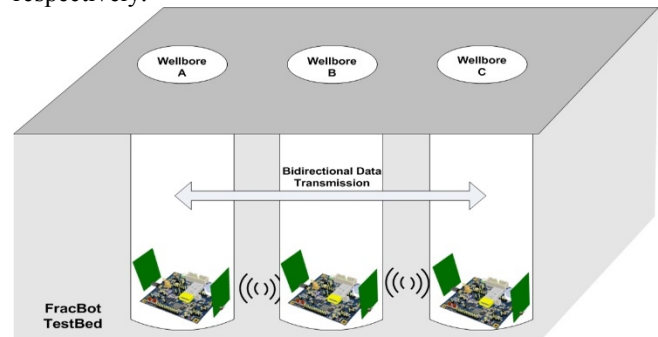


Figure 12. A model of physical testbed in hydraulic fracture.

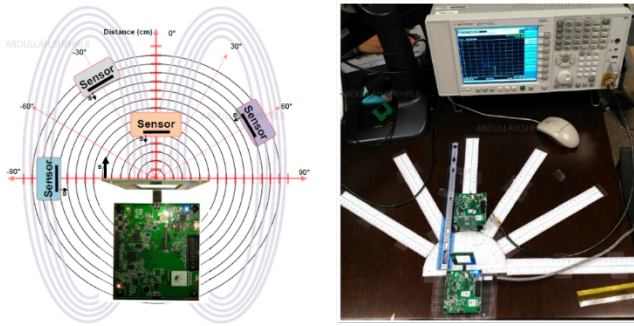


Figure 13. FracBot experimental setup.

2) *Angular Analysis*: The direction and the alignment of the transceiver and the receiver of the FracBots is one of the complications in MI-based communication. In the angular study, we perform measurements at 0, 30°, 60° and 90° angles. The results of distances between 6 and 25 cm, compared to those under 6 cm reveal minor variants. Figure 14 displays the power analyses of the angular variations. At distances of 6 cm and beyond, the angle between antennas (the transceiver and the receiver) affects the received power slightly, less than -2 dBm.

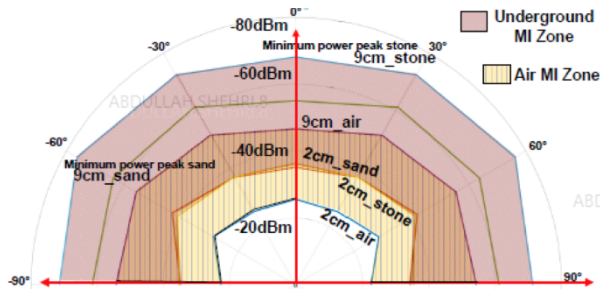


Figure 14. Angular plots of received power (air, sand/stone).

The angular study shows that the MI field radiated at 13.56 MHz is omni-directional. It enables the Base station to assess the location of each sensor and produce a fracture map, when this characteristic is incorporated with the received signal strength indicator (RSSI) measurement. The FracBot MCU needs 50 ms to complete all reading tasks and then stock them in the NFC transponder. This task consumes 33μW of the energy available in the storage system. Based on the angular analysis, the node can function constantly by harvesting energy of the MI field if the receiver is positioned at 23 cm or nearer to the succeeding FracBot node. As an outcome, the received power in the area of 6-25 cm is approximately -50 dBm that delivers adequate energy to the node each hour and allows it to transmit information in a 50 ms time frame. After 25 cm, the received power is less than -50 dBm, that is not enough to power the node each hour. As a result, the node require to collect the necessitated energy and transmit information within a time frame of 50 ms every 2 hours at minimum. It is worth to mention that the FracBot can operate in an intermittent status if the MI signal strength is lower than -50 dBm.

3) *FracBot Underground Testbed*: To measure the FracBot nodes performance, we design and develop a testbed

similar to underground environment comprising of a plastic container containing water, sand, and stone, demonstrated in Figure 15. The system involves several underground settings, comprising dry soil, wet soil, stone and dry soil with stone. The testbed setting permits to position the FracBots at different depths until 14 cm with a adjustable distance between the nodes. This flexibility enable changes of the experimental setup to easily evaluate the FracBot nodes performance. Using the spectrum analyzer, we measure the MI circuits characteristics such as MI propagation and antenna tuning.

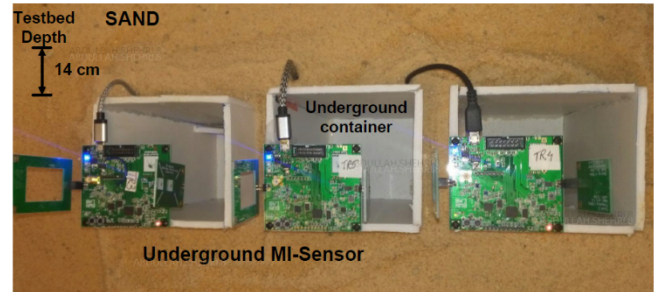


Figure 15. Underground testbed of the FracBot.

To assess the transmission link, we wirelessly link the NFC tag of first FracBot to the transceiver of second FracBot. The FracBots conduct one communication task every 3 minutes and one temperature reading per minute in the laboratory. For experimental purposes, the data transmission of long interval can be simulated by the adjustable interval time in a short time. The nodes utilize NFC technique, but as they are intended to operate in air, a consistent reference test and data analysis in air is essential. The node is examined to transmit in air and with a sand obstacle.

Table 1 displays the experimental performance for OOK and ASK modulations with data rates of 26 and 1.6 kbit/s. In an underground environment, the modulation OOK at data rate of 1.6kbit/s, compared with that at 26 kbit/s, lowers the transmission error. However, in stone, ASK modulation does not work for both rates due to high attention. On the other hand, OOK modulation works but at a higher transmission error than that in sand for both rates. Former study in underground field claims 10 MHz as an optimum frequency with data rate of 1 kbit/s [15]. To estimate the transmission link among FracBots, the nodes are located at 5 cm distant from each other, as shown in Figure 15 because of the restriction posed by the sensitivity of the off-the-shelf transponder chip limited to -50 dBm. At 5 cm, the signal strength is -50 dBm. Beyond 5 cm, the signal quality will degrade as well as the communication becomes impossible.

Table 1. Experimental performance of the ASK and OOK modulation.

Environment	Modulation	Date rate (kbit/s)	Error (%)
Air	ASK	26	2
Air	OOK	26	1

Sand	ASK	26	70
Sand	OOK	26	78
Sand	ASK	1.6	40
Sand	OOK	1.6	32
Stone	OOK	26	87
Stone	OOK	1.6	58

VI. REAL-TIME INTELLIGENT SENSOR SELECTION

In order to efficiently and long-term deploy subsurface sensors, it is crucial to optimize the sensor capability to sense as well as extend the lifetime of each sensor as long as possible. An essential part of optimizing the sensor capability to sense in the reservoir formation is to optimally select the best number of sensors. There are several trade-offs that have to be taken into account such as the battery utilization of sensors as well as need to have multiple close sensors being in operation during the same time. Specifically, one aims to reduce the number of sensors being in operation at the same time, while maintaining sufficient sensing reach. The resulting problem can then be transformed into a sensor selection problem. The sensor selection problem is mathematically defined as given a set of sensors $S = \{S_1, \dots, S_n\}$, then we need to select the best subset with k sensors that satisfy one or multiple missions. The challenge that arises from this problem is in most instances NP-complete, which implies that there is no polynomial-time algorithm for solving the problem. This represents a major challenge for real-time data interpretation and the optimization of the sensors as in order to be able to have a recommendation available within an acceptable timeframe, an approximate solution is only feasible [22]. We will demonstrate a novel intelligent sensor selection framework for the optimization of sensor selection in real-time for flow and fracture monitoring. The objective of the framework is to maximize longevity of the operations while maintaining measurement accuracy and flow detection ability.

A. Method

We have developed an innovative real-time sensor utilization optimization framework that incorporates a deep learning driven optimization framework connected to a subsurface fracture network model. This forms then a crucial part of the sensor selection optimization problem that aims to optimize in real-time to minimize the number of sensors required in order to maintain sufficient data quality. This challenge is equivalent to maximize the longevity of the sensors deployed while maintaining sufficient reservoir coverage in order to limit the uncertainty in the multi-data interpretation.

The framework incorporates a deep learning approach for the sensor measurements combined with a fast iterative solver for real-time optimization of the sensor selection. The framework is outlined in Figure 16.

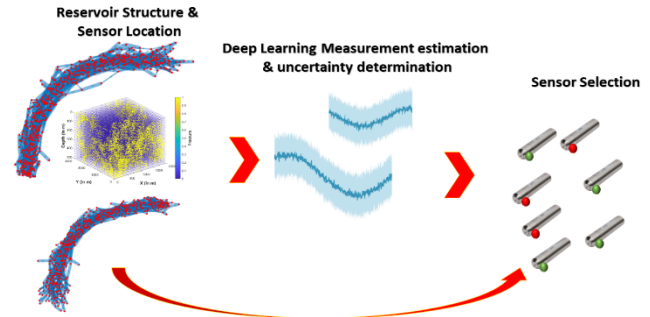


Figure 16. Framework representation with the fracture network structure and the uncertainty estimates.

First, a fracture-flow reservoir model is established using a connectivity and sensing data quality determination approach. The assumption is that the flow between injecting and producing wells is primarily within the fractures with only limited flow in the matrix structures. This is in line with conventional assumptions when utilizing discrete fracture network models, as well as observations on fractured carbonate reservoir rocks, where the flow is primarily in the fractures. The network flow model is then integrated into a deep learning framework for the sensor data estimation and the uncertainty in the estimates. The deep learning framework utilizes a feedforward network structure for determining from the sensor derived flow measurement data based on multiple potential scenarios in terms of the reservoir formation condition. The estimations relate to whether the sensors are close to the matrix or in the fracture, and what the water saturation in the vicinity of the sensor is. The main objective of the deep learning framework is to have a data-driven approach to the estimation of the fracture and water saturation in the vicinity of the sensor based on pressure and temperature measurements. The sensor selection problem is then posed as an integer optimization problem as outlined below:

$$\begin{aligned} \min f^T z \\ \text{s.t. } C_z > 0 \\ U_z \leq b_u, \forall_i \in N \end{aligned} \quad (9)$$

The integer optimization problem is solved in real-time where the vector f is the cost function dependent power consumption over time of the sensors. For each update time step, the cost function is updated from the previous, implying that if the sensor i is operational, then f_i is gradually increasing, while for the inactive sensors, f_i may remain constant or is reduced in case the sensors can be recharged. The constraint $C_z > 0$ ensures that there is for each reservoir area at least one sensor that covers this area. The matrix C is the connectivity matrix between the sensors and the area, implying that $C_{ij} = 1$ if the j -th sensor covers the i -th area. This ensures that each area is covered, and that the sensor can connect and transfer data between each other. Data transmission is a crucial

The constraint $Uz \leq b_u$ implies that the data sensing reliability for each node is maintained, implying that the sensing uncertainty must be below a threshold value. The matrix or vector U is the sensing reliability matrix, and b_u is the reliability threshold. The last constraint is a binary constraint, indicating whether the sensor is active ($z_i = 1$) or inactive ($z_i = 0$). For solving the integer optimization problem, we utilized a fast and efficient branch and bound method, via utilizing a feedback approach incorporating the solutions of previous optimizations. The framework is easily scalable to larger flow network models, allowing in near real-time to optimize the selection of sensors and maintain longevity of the sensor deployments.

B. Results

We examined the framework on a complex fracture network structure in 2D in order to outline the performance of the framework. The 2D model is a graph-based model consisting of 500 nodes and 1000 different network structure realizations. We have displayed in Figure 17 two examples of the different network realizations and connection between the fracture network nodes. The realizations illustrate the considerable difference between the connectedness of the fracture network which reflects the general challenge of monitoring and determining the fracture network structure and connectedness between the fractures. We then utilized a deep learning approach to estimate the uncertainty of the data based on the network structure. The data set was divided 75/15/15 into a training, validation and test dataset, and a fully connected feedforward neural network structure was used.

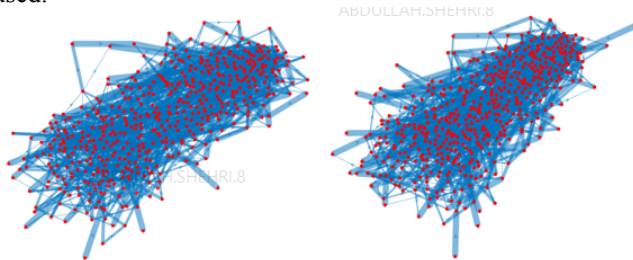


Figure 17 Different realizations of the fracture network structure.

For the optimization, we used a scaled conjugate gradient approach given the substantial size of the problem. The sensors record pressure and temperature data at each location, and for each of the sensors an interpreted uncertainty parameter is computed. The uncertainty parameter varies from 0 to 1.5, where a higher uncertainty parameter indicates stronger uncertainty in the measured data. The uncertainty measurement parameters are derived from multiple repeat measurements of the sensors that are then classified in terms of their accuracy and variation. The training, validation and testing results of the deep neural network are displayed in Figure 18. The estimation results are rather strong, outlining overall accurate estimation of the sensor data uncertainty, with the larger number of data points

for lower uncertainties only marginally affecting the estimation quality for higher uncertainties.

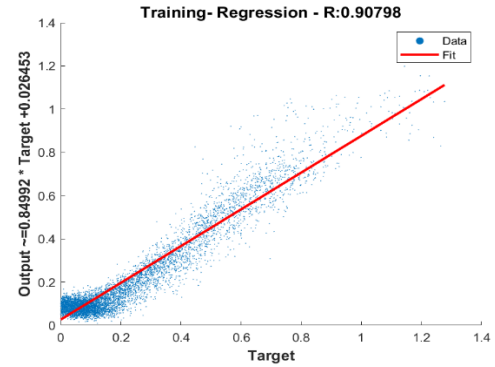


Figure 18 a. Comparison of the neural network estimation of the data uncertainty.

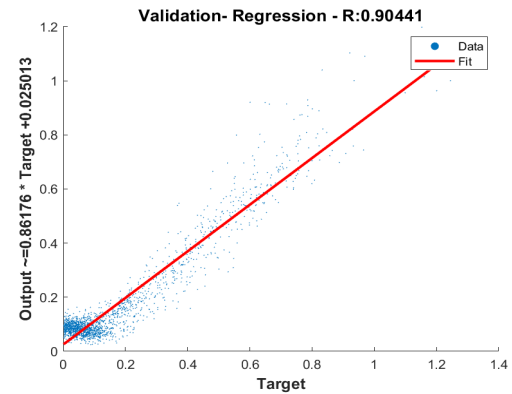


Figure 18 b. Comparison of the neural network estimation of the data uncertainty.

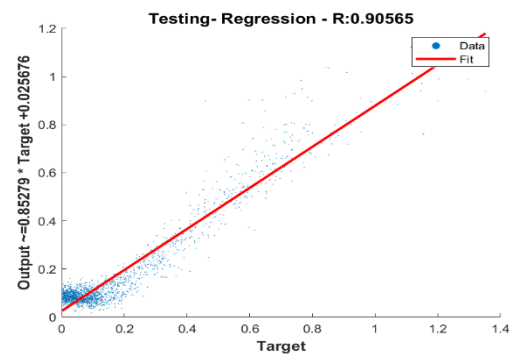
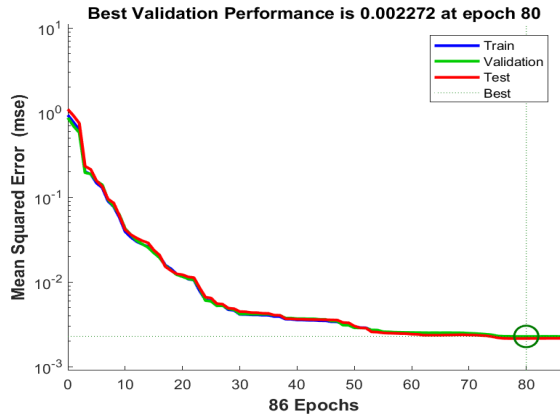


Figure 18 c. Comparison of the neural network estimation of the data uncertainty.



Utilizing the deep learning network model, we then solved the sensor selection problem in real-time under uncertainty. The uncertainty matrix U is updated in each simulation step to reflect the changing reservoir conditions as well as sensing parameters. The cost vector f for the sensors is increased in each step for the active sensor components, reflecting the power utilization of sensor and to penalize excessive usage of an individual sensor. In case the sensor is not anymore operational f_i (e.g., lack of power), then f_i was set to positive infinity. The timeframe for the sensor optimization was from April 1st, 2019 until January 11th, 2020, where the sensors were optimized every 15 days. The optimization results are displayed in Figure 19 outlining the active sensors in green and the inactive in black.

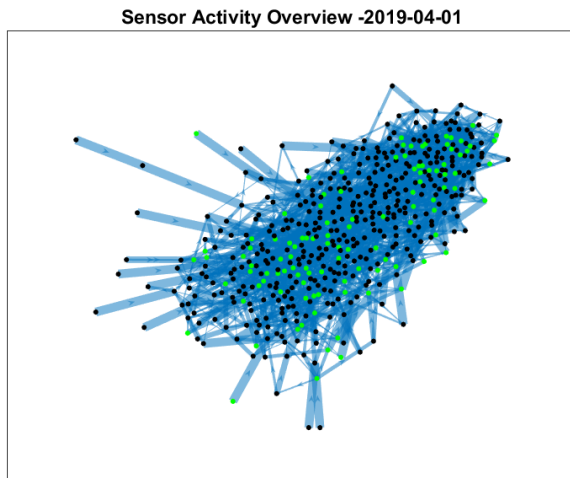


Figure 19 a. Overview of the selected sensors for different time steps.

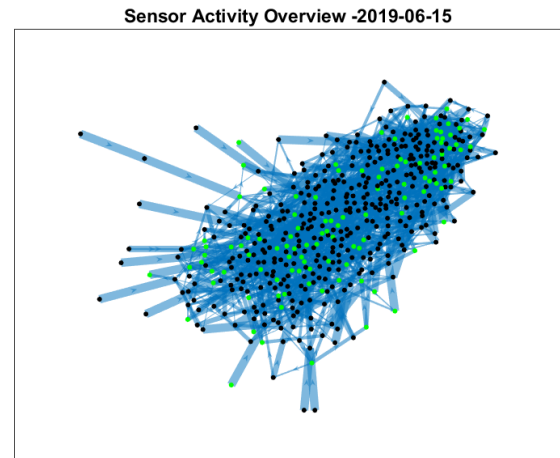


Figure 19 b. Overview of the selected sensors for different time steps.

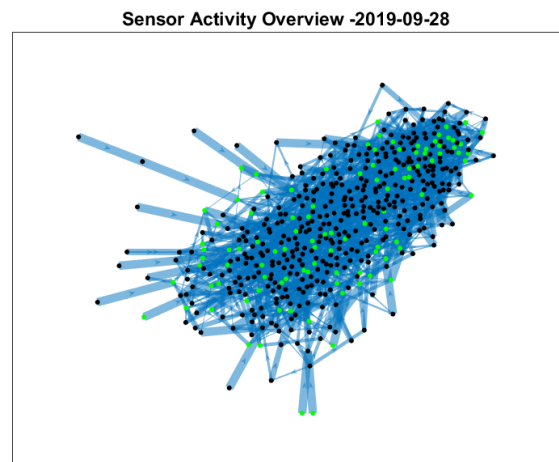


Figure 19 c. Overview of the selected sensors for different time steps.

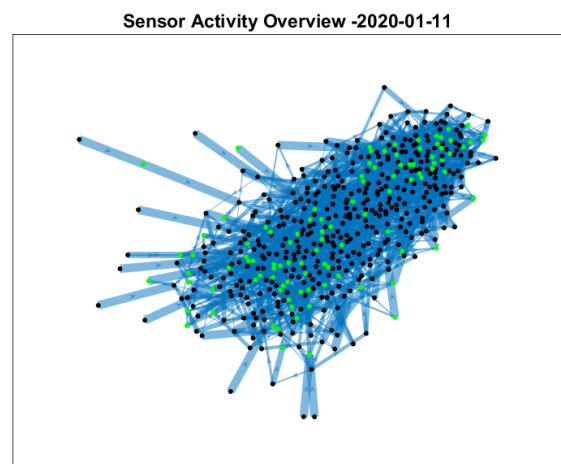


Figure 19 d. Overview of the selected sensors for different time steps.

As observed there are certain sensor clusters that are active for longer durations indicating that these sensors are

placed in crucial fracture intersection points as well as exhibit a low degree of measurement uncertainty. This is confirmed via a sensor utilization analysis for the 500 sensors in Figure 20 and Figure 21. The indication is that most sensor are rarely active, or solely active for a short period of time, while there are a few sensors that are heavily utilized and operational for more than 250 days out of 285 days .

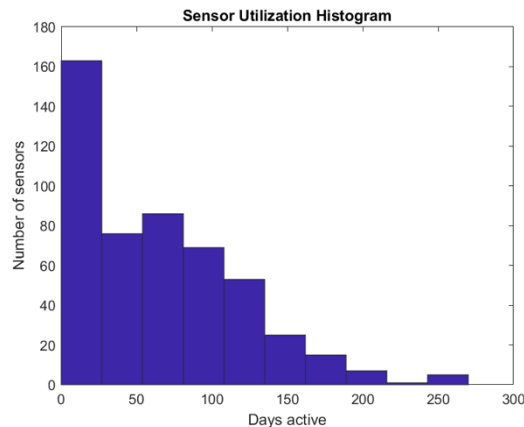


Figure 20. Sensor utilization histogram.

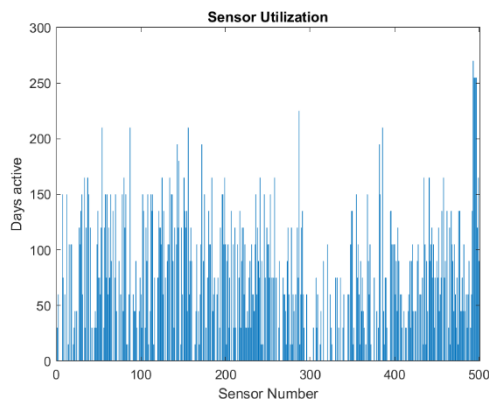


Figure 21. Sensor utilization in days.

VII. CONCLUSION

This paper proposed FracBots systems for monitoring oil and gas reservoirs, mapping hydraulic fractures and collect other wellbore parameters. We established a platform of the FractBots comprising of software and hardware solutions. We formulated and developed three key functions. We developed cross layer communication model for magnetic induction networks in altering underground environments to enable the communication in dynamically changing underground environments. We developed an innovative MI-based localization framework to capture the locations of the randomly deployed FracBot nodes by exploiting the exceptional properties of the MI-field. We developed an energy model framework for a linear FracBot network topology to estimates FracBot data transmission rates while respecting harvested energy constraints. We designed and

developed novel prototypes of wireless FracBots for potential use as a platform for a new generation of WUSNs for monitoring hydraulic fractures and unconventional reservoirs, and measuring other wellbore parameters. We developed the hardware of the MI-based wireless FracBots for short-range communication using near-field communication (NFC) as a physical layer combined with an energy-harvesting capability and ultra-low power requirements. Finally, to examine the functionalities of FracBot nodes in air, sand, and stone media, a physical MI-based WUSN test bed was implemented. Experiments indicated that the constructed FracBots can form a transmission link and transfer data over ASK modulation using a data rate of 1.6 Kbit/s and a minimum receiver sensitivity of -70 dBm. The hardware development and the testbed analyses allow us to better understand the environment challenges, improve the electronic sensitivity and optimize the minimum resources that are necessary to miniaturize the FracBot hardware.

In addition, we presented a novel AI driven sensor selection framework for the optimal selection of subsurface pressure and temperature sensors in a fractured reservoir. The framework presents the ability to optimize the selection of sensors for subsurface sensing in real-time, thereby maximizing the overall coverage of the sensors for efficient waterfront tracking. The results outline the ability to efficiently and long term perform reservoir sensing if the sensors are optimally selected and utilized.

References

- [1] A. Alshehri, "FracBots: The Next Real Reservoir IoT," The Fifteenth International Conference on Systems and Networks Communications (ICSNC 2020), Porto, Portugal Oct. 18- 22, 2020
- [2] K. Katterbauer, I. Hoteit, and S. Sun, "EMSE: Synergizing EM and seismic data attributes for enhanced forecasts of reservoirs," *Journal of Petroleum Science and Engineering*, 2014, 122, pp. 396- 410.
- [3] K. Katterbauer, I. Hoteit and S. Sun, "History Matching of Electromagnetically Heated Reservoirs Incorporating Full-Wavefield Seismic and Electromagnetic Imaging," *SPE Journal*, 2015, 20(5), pp. 932- 94.
- [4] T. Ertekin and Q. Sun, "Artificial intelligence applications in reservoir engineering: a status check," *Energies*, 2019. 12(15), P. 2897.
- [5] R. Miftakhov, A. Al-Qasim, and I. Efremov, "Deep Reinforcement Learning: Reservoir Optimization from Pixels," *International Petroleum Technology Conference*, Dhahran, 2020.
- [6] P. Panja, R. Velasco, M. Pathak, and M. Deo, "Application of artificial intelligence to forecast hydrocarbon production from shales," *Petroleum*, pp. 75- 89, 2018.
- [7] S. Fumagali, "Robotic Technologies for Predictive Maintenance of Assets and Infrastructure," *IEEE Robotics & Automation Magazine*, 2018. 25(4), pp. 9-10.
- [8] A. Davarpanah, B. Mirshekari, T. Jafari, and M. Hemmati, "Integrated production logging tools approach for convenient experimental individual layer permeability measurements in a multi-layered fractured reservoir," *Journal of Petroleum Exploration and Production Technology*, 2018, 8(3), pp. 743- 751.

- [9] F. Sana, K. Katterbauer, T. Al-Naffouri, and I. Hoteit, "Orthogonal matching pursuit for enhanced recovery of sparse geological structures with the ensemble Kalman filter," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2016, 9(4), 1710- 1724.
- [10] Offshore Magazine Business Briefing, "Miniature downhole sensors offer improved shock resistance," *Offshore Magazine*, 2013.
- [11] A. AlShehri and A. Shewoil, "Connectivity Analysis of Wireless FracBots Network in Hydraulic Fractures Environment," *Offshore Technology Conference Asia*, Kuala Lumpur, 2020.
- [12] J. Xu, X. Wang, K. Cooper, G. Pickrell, and A. Wang, "Miniature Temperature-Insensitive Fabry-Perot Fiber Optic Pressure Sensor," *IEEE Photonics Technology Letters*, 2006, 18(10), pp. 1134- 1136.
- [13] M. Akkas, I. Akyildiz, and R. Sokullu, "Terahertz Channel Modeling of Underground Sensor Networks in Oil Reservoirs," *IEEE Global Communications Conference*, 2012.
- [14] A. Alshehri, S. Lin, and I. Akyildiz, "Optimal Energy Planning for Wireless Self-Contained Sensor Networks in Oil Reservoirs," *IEEE International Conference on Communications*, 2017.
- [15] H. Guo and Z. Sun, "Channel and Energy Modeling for Self-Contained Wireless Sensor Networks in Oil Reservoirs," *IEEE Transactions on Wireless Communications*, 2014, 13(4), pp. 2258- 2269.
- [16] Z. Sun and I. Akyildiz, "Magnetic Induction Communications for Wireless Underground Sensor Networks," *IEEE Transactions on Antennas and Propagation*, 2010, 58(7), pp. 2426- 2435.
- [17] S. Lin, I. Akyildiz, et al. "Distributed Cross-Layer Protocol Design for Magnetic Induction Communication in Wireless Underground Sensor Networks," *IEEE Transactions on Wireless Communications*, 2015, 14(7), pp. 4006- 4019.
- [18] I. Akyildiz, H. Schmidt, S. Lin, and A. Alshehri, "Environment-Aware Cross-layer Communication Protocol Design in Underground Oil Reservoirs," *U.S. Patent No. 10,117,042*, 2018.
- [19] S. Lin, A. Alshehri, Wang, P. et al. "Magnetic Induction-Based Localization in Randomly-Deployed Wireless Underground Sensor Networks," *IEEE Internet of Things Journal*, 2017, 4(5), pp. 1454- 1465.
- [20] C. Martins, A. Alshehri, and I. Akyildiz, "Novel MI-based (FracBot) sensor hardware design for monitoring hydraulic fractures and oil reservoirs," *Th 8th IEEE Annual Ubiquitous Computing, Electronic Mobile Comm. Conference*, 2017.
- [21] B. Finch and W. Goh, "MSP430™ Advanced Power Optimizations: ULP Advisor™ Software and EnergyTrace™ Technology," *Application Report SLAA603*. Texas Instruments, 2014.
- [22] T. Yoo and S. Lafortune, "NP-completeness of sensor selection problems arising in partially observed discrete-event systems," *IEEE Transactions on Automatic Control*, 2002, 47(9), pp. 1495-1499.

A Framework of Web-Based Dark Patterns that can be Detected Manually or Automatically

Ioannis Stavarakakis, Andrea Curley, Dympna O’Sullivan, Damian Gordon, Brendan Tierney

ASCNet Research Group, School of Computer Science, Technological University Dublin, Dublin, Ireland

Email: Ioannis.Stavarakakis@TUDublin.ie, Andrea.F.Curley@TUDublin.ie, Dympna.OSullivan@TUDublin.ie, Damian.X.Gordon@TUDublin.ie, Brendan.Tierney@TUDublin.ie

Abstract— This research explores the design and development of a framework for the detection of Dark Patterns, which are a series of user interface tricks that manipulate users into actions that they do not intend to do, for example, share more data than they want to, or spend more money than they plan to. The interface does this using either deception or other psychological nudges. User Interface experts have categorized a number of these tricks that are commonly used and have called them Dark Patterns. They are typically varied in their form and what they do, and the goal of this research is to explore existing research into these patterns, and to design and develop a framework for automated detection of potential instances of web-based dark patterns. To achieve this, we explore each of the many canonical dark patterns and identify whether or not it is technically possible to automatically detect that particular pattern. Some patterns are easier to detect than others, and there are others that are impossible to detect in an automated fashion. For example, some patterns are straightforward and use confusing terminology to flummox the users, e.g. “Click here if you do not wish to opt out of our mailing list”, and these are reasonably simple to detect, whereas others, for example, sites that prevent users from doing a price comparison with similar products might not be readily detectable. This paper presents a framework to automatically detect dark patterns. We present and analyze known dark patterns in terms of whether they can be either: (1) detected in an automated way (it can be partially or fully), (2) detected in a manual way (it can be partially or fully) and (3) cannot be detected at all. We present the results of our analysis and outline a proposed software tool to detect dark patterns on websites, social media platforms and mobile applications.

Keywords-Dark Patterns; User Experience; Digital Ethics; Privacy.

I. INTRODUCTION

Computers and technological applications are now central to many aspects of life and society, from industry and commerce, government, research, education, medicine, communication, and entertainment systems. Computer scientists and professionals from related disciplines who design and develop computer applications have a significant responsibility, as the systems they develop can have wide ranging impacts on society where those impacts can be beneficial but may also at times be negative, thus it cannot be argued that modern technology is value-neutral, as it is clear that it can have both planned and unplanned negative consequences on users.

In this, and previous research [1], we outline and explore the ethical limits of a technology design phenomenon known as “dark patterns”. Dark patterns are user interfaces that benefit an online service by leading users into making decisions they might not otherwise make. At best, dark patterns annoy and frustrate users. At worst, they can mislead and deceive users, e.g., by causing financial loss, tricking users into giving up vast amounts of personal data or inducing compulsive and addictive behavior in adults and children. They are an increasingly common occurrence on digital platforms including social media sites, shopping websites, mobile apps, and video games. Although they are gaining more mainstream awareness in the research community, dark patterns are the result of three decades-long trends: one from the world of retail (deceptive practices), one from research and public policy (nudging), and the third from the design community (growth hacking) [2].

The aim of our work is the development of a framework for classifying web-based dark patterns as to which are readily detectable, and which are not. The framework forms the basis of a software tool that can automatically alert users to the presence of dark patterns on websites, social media platforms and mobile applications. In developing the framework we analysed common documented types of data patterns. We present these dark patterns to the reader and classify each dark pattern using the following taxonomy: (1) A pattern that can be detected in an automated way (either partially or fully); (2) A pattern that can be detected in a manual way (either partially or fully); and (3) A pattern that cannot be detected. In this paper we outline the features and functionality of the proposed tool. This research is part of a larger research project (called Ethics4EU) whose goal is develop a repository of teaching and assessment resources to support the teaching of ethics in computer science courses, supported by the Erasmus+ programme [3].

In Section 2, a review of some of the key literature focusing on what dark patterns are, and why they are so successful. Section 3 looks at the specific collection of dark patterns that will be explored in this research. Section 4 presents the initial framework for the detect of dark patterns, looking at which patterns can be detected automatically, which manually, and which cannot be detected at all. Section 5 outlines some other dark patterns that should also be looked at, and finally, Section 6 presents some conclusions and future work about this research.

II. LITERATURE REVIEW

Since the early 1980s computer programmers have used the concept of patterns in software engineering as a useful way of categorizing different types of computer programs. The term dark patterns has been used since 2010 to refer to interface design solutions that intend to deceive users into carrying out undesirable actions [4]. Gray et al. [5] defined dark patterns as “instances where designers use their knowledge of human behavior (e.g., psychology) and the desires of end users to implement deceptive functionality that is not in the user’s best interest”.

There has been significant research done on dark patterns from the fields of Cognitive Psychology, Usability, Marketing, Behavioural Economics, Design and Digital Media. All this research has led to the abandonment of the rational choice theories for explaining decision making, particularly for matters of privacy [6] and has prompted new examinations that attribute the effectiveness of dark patterns on human cognitive limitations. However, there is still not a universal theoretical explanation of the ‘whys’ and ‘hows’ of the effectiveness of dark patterns. For example, Maier [7] argues that manipulation is closely linked to decision making and the latter can be easily influenced through one’s emotions and mood leading to decisions lacking rational thought [8].

What is more, according to Kahneman [9] humans are more intuitive than rational thinkers and most of their daily reasoning is performed by their intuition. Below are the main human psychological mechanisms being targeted or exploited by Dark Patterns [10]:

- Nudging, which is based on soft paternalism, positive reinforcement and compliance [11]. Nudging can be and has been used with good intentions in mind and has been proved effective [12][13]. However, because of its proven efficiency, nudging is one of the most common digital manipulation strategies used to mislead users into bad decisions privacy-wise.
- Persuasion techniques built on what Cialdini [14] identifies as the “six basic tendencies of human behaviour” (p. 76). These tendencies namely are: reciprocation, consistency, social validation, liking, authority and scarcity.
- Cognitive biases that fundamentally are information processing limitations of the human mind and are rooted in cognitive heuristic systems [9]. According to Waldman [15] the five most pervasive are: anchoring [16], framing [17], hyperbolic discounting [18][19][20], overchoice [21][22][23] and metacognitive processes such as cognitive scarcity [24] and cognitive absorption [25].
- Cognitive dissonance, an uncomfortable state of mind where one’s beliefs and actions are contradictory. Bösch et al. [10] (p. 247) mention “[i]n terms of privacy dark patterns, this process can be exploited by inconspicuously providing justification arguments for sugar-coating user decisions that have negatively affected their privacy”.

Although, so far, it appears that the cognitive and psychological factors play a significantly important role on users’ failure to protect their privacy when dealing with Dark Patterns, some researchers argue that contextual and social factors are important too. For example, Acquisti et al. [6] claim that incomplete or asymmetric access to information between two agents in a transaction can significantly disadvantage one party leading to problematic decisions. Furthermore, users are not always certain of what they are agreeing to share as the collection of personal data is not always apparent and therefore people remain unaware of what information is collected about them by both private and public organisations [26]. This is usually the norm in digital environments where the user has no control over the design and information processing they are being shown.

On the other hand, research has shown that users, care about their privacy [27], however, the contextual, social and cognitive aspects mentioned earlier lead users to a set of behaviours that are inconsistent to their attitudes towards privacy [15]. Norberg et al. [28] have called this the ‘privacy paradox’.

In today’s digital environment most digital platforms’ provide services seemingly for free. In order for these services to generate revenue they have become dependent on accumulating and processing users’ data, oftentimes personal data [29]. According to Zuboff [30] user data is the raw material that produces, what she calls, ‘behavioural surplus’ which has become a valuable commodity for companies. Behavioural surplus is a powerful tool for predicting user behaviour and many companies use it to influence users into providing more data which leads into a vicious cycle of user data, influence, prediction and so on [31].

Mathur et al. [32] did a meta-analysis of 11,286 shopping websites, and created a taxonomy to try to explain how dark patterns affects user decision-making by exploiting cognitive biases. Their taxonomy has the following characteristics: Asymmetric, Covert, Deceptive, Hides Information, and Restrictive. They found that 11.1% (1254 websites) of the sites had dark patterns, and recommend the development of plug-ins for browsers to help detect these patterns.

Nouwens et al. [33] discuss the growth of Consent Management Platforms (CMPs) which are software systems that manage the interaction between users and the website(s) of an organization, recording (and updating) their privacy preferences, and getting consent for recording interactions with cookies. Crucially these CMPs are compliant with GDPR (the General Data Protection Regulation) however it is still possible for a website to employ Dark Patterns to circumvent GDPR, and almost 90% of the sites with CMPs surveyed were in some way themselves breaching GDPR.

Chromik et al. [34] explore how there is potential for dark patterns to be used in Intelligent Systems. An intelligent system is computer system with an embedded artificial intelligence that can work to solve well-defined tasks, e.g. object recognition, medical diagnosis, language translation. As a consequence of GDPR, these systems must be able to provide some explanation as to how they came to

specific decisions. Some intelligent systems incorporate explanation facilities to support users in understanding decisions. However, this paper discusses the possibility of Intelligent Systems using Dark Patterns in conveying these explanations to get further data from the users. For example, the system could use a Dark Pattern to collect valuable user data under the pretext of explanation. So, the user might be forced to provide additional personal information (e.g., social connections) before receiving personalized explanations. Otherwise, the user would be left off with a generic high-level explanation.

Di Geronimo et al. [35] explore the use of Dark Patterns in mobile apps. They looked at 240 popular mobile apps and explored whether or not these apps included any dark patterns. Their analysis showed that 95% of the apps they reviewed included one or more Dark Patterns, with an average of 7.4 malicious designs per app, with a standard deviation of 5. Almost 10% of the apps included 0, 1, or 2 Dark Patterns (N=33), 37% of the apps contained between 3 to 6 Dark Patterns (N=89), while the remaining 49% included 7 or more (N=118). They also conducted an online experiment with 589 users on how they perceive Dark Patterns in such apps. Overall, the majority of our users did not spot malicious designs in the app containing Dark Patterns (55%), some were unsure (20%), and the remaining found a malicious design in the app (25%). But they found that most users did perform better in recognizing malicious designs if informed on the issue.

Grassl et al. [36] looked at cookie consent requests in the context of Dark Patterns to explore whether or not they undermine principles of EU privacy law. They undertook two online experiments where they investigated the effects of common design nudges on users' consent decisions and their perception of control over their personal data in these situations. In the first experiment (n = 228) they explored the effects of dark patterns to encourage the participants to select the privacy-unfriendly option, and the experiment revealed that most people agreed to all consent requests regardless of dark patterns. The research indicated that the dark patterns made no difference to the participants' behaviour. The first experiment, also showed that despite generally low levels of perceived control, obstructing the privacy-friendly option led to more rather than less perceived control for the participants. In the second experiment (n = 255) the participants we presented with patterns to select the privacy-friendly option (bright patterns). The bright pattern did succeed in swaying people effectively towards the privacy-friendly option. The second experiment also looked at the perceived control of the participants, and it found that it stayed the same compared to Experiment 1. Overall, the researchers concluded about Experiment 1 that whether the participants were presented with a dark pattern or not, they have been conditioned by years of practice to consent, and therefore they concluded that the EU's consent requirement for tracking cookies does not work as intended.

Dark patterns are only just beginning to emerge as a topic in the software development literature. In 2021 Kollnig et al. [37] reported in the development of a functional prototype that allows users to disable dark patterns in apps selectively. This differs from our approach where we are developing a comprehensive framework for identifying dark patterns across a range of platforms, from apps to websites.

Chugh and Jain [38] looked at dark patterns from the perspective of consumer protection as well as their impact on democratic political processes. The researchers distinguish between dark patterns and persuasive advertisements, classifying dark patterns as being manipulative, whereas persuasive advertisements merely attempt to influence people to revise their preferences. They see two major issues with dark patterns, (1) users are typically unaware that they are interacting with dark patterns, and are, therefore, unable to safeguard themselves against the effects of these patterns, and (2) market forces and market competition don't seem to be penalizing organizations for using these patterns. Therefore, they recommend that legislation and regulations are necessary to combat these patterns.

Bongard-Blanchy et al. [39] explored the impact of dark patterns on end-users by surveying 406 individuals. They found that although the participants were aware of the type of manipulative techniques that online services use to impact their online behaviour, they are nonetheless unable to combat their impact. The researchers advocate a multi-faceted approach to addressing these issues, including raising awareness and educating people about the different patterns and how they work, concomitant with this approach, the researchers propose that the users are presented positive information that will encourage them to avoid engaging with new patterns and to cease engaging with existing patterns, e.g. the user could be made aware of how much time they spend engaging with infinite scrolling systems, and they could be reminded that they could be using that time for more enjoyable activities. They also advocate targeting the educational initiatives about patterns based on age-groups and other demographics, and finally they suggest that a combination of strong legal penalties and regulations are needed, as well as new software tools to help detect and highlight the existence of these patterns. However, they do note that some patterns may be more readily detectable in an automated fashion than others.

III. PATTERN DESCRIPTIONS

A vital step in developing the web-based Dark Patterns Framework is to clearly define each pattern and to categorize the patterns into themes. In the research literature previously discussed there is some variance as to the exact meaning of each pattern, therefore below we present definitions that attempt to be as inclusive as possible to the range of definitions for each pattern, but always prioritizing the original canonical definitions developed by the pioneer of dark patterns - user experience designer Harry Brignull [4].

A. *Sneaking*

- **Sneak into Basket:** When purchasing a product, an additional item is added into the basket, usually the new product is added in because of an obscured opt-out button or checkbox on a previous page. Detection of this pattern is challenging since there may be legitimate reasons for a site to add new items into a shopping basket (e.g. taxes), therefore, automated detection may not be possible, but nonetheless it would still be possible to manually highlight changes in cost, and let the shopper decide if the additional items are valid.

- **Hidden Costs:** When reaching the last step of the checkout process, some unexpected charges have appeared in the basket, e.g. delivery charges, etc. Detection of this pattern is challenging since there may be legitimate reasons for a site to add new items into a shopping basket (e.g. taxes), therefore, automated detection may not be possible, but nonetheless it would still be possible to manually highlight changes in cost, and let the shopper decide if the additional items are valid.

B. *Misdirection*

- **Trick Questions:** Often found when registering for a new service. Typically, a series of checkboxes are shown, and the meaning of checkboxes is alternated so that ticking the first one means "opt out" and the second means "opt in". Detection of this pattern is possible at least partially because it is possible to detect pre-ticked checkboxes, and to search for phrases like "opt out" and "opt in".
- **Misdirection:** When the design purposefully focuses users' attention on one thing in order to distract their attention from another, for example, a website may have already undertaken a function and added a cost to it, and the opt out button is small. Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.
- **Confirmshaming:** This involves guilt-tripping the user into opting into something. The option to decline is worded in such a way as to shame the user into compliance, for example, "No thanks, I don't want to have unlimited free deliveries". Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.
- **Disguised Ads:** Advertisements that are disguised as other kinds of content or navigation, in order to get you to click on them, for example, advertisements that look like a "download" button or a "Next >" button. Detection of this pattern is possible at least partially because it is possible to detect buttons on a webpage. And by using either the ALT tags or OCR to determine

the purpose of the button, and then to look at whether it links internally, or to an external site.

C. *Obstruction*

- **Roach Motel:** When users find it easy to subscribe to a service (for example, a premium service), and find it is hard to get out of it, like trying to cancel a shopping account. Detection of this pattern is possible because it is possible to search for "activate" or "subscribe" links or buttons, that have no reciprocal "deactivate" or "unsubscribe" links or buttons.

D. *Forced Action*

- **Forced Continuity:** When a user gets a free trial with a service comes to an end and their credit card silently starts getting charged without any warning, and there isn't an easy way to cancel the automatic renewal. Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.

E. *Variegations*

- **Privacy Zuckering:** Tricking users into sharing more information than they intended to, for example, Facebook privacy settings were historically difficult to control. Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.
- **Price Comparison Prevention:** The retailer makes it hard for you to compare the price of an item with another item, so you cannot make an informed decision. Retailers typically achieve this by creating different bundles where it is not easy to work out the unit price of the items within the bundles. Detection of this pattern is challenging since it may not be obvious (or clearly labelled) if the products are in different bundles, but it will be possible to manually highlight packaging types, and let the shopper decide if there are any issues.
- **Bait and Switch:** The user sets out to do one thing, but a different, undesirable thing happens instead, for example, Microsoft's strategy to get users to upgrade their computers to Windows 10. Detection of this pattern is extremely challenging as there is such a significant variation in how the pattern is implemented on different sites.
- **Friend Spam:** The product asks for users for their email or social media permissions to spam all their contacts. Detection of this pattern is possible since the HTML in the website can be analyzed to determine if the site asked for email or social media permissions.

F. Beyond Brignull

UX researcher Reed Steiner [40] added six patterns:

- **Fake Activity:** On a commercial website, when the page says “three other people are viewing this item right now” this may not be a fully truthful claim. Detection of this pattern is possible at least partially because it is possible to search for phrases such as “other people are viewing this item now” and warn the shopper of this pattern.
- **Fake Reviews:** Research shows that several reviews and testimonials are fake, and exact matches with different customer names can be found on several sites. Detection of this pattern is challenging, but it may be possible to take reviews from the current site, and manually search for them on other similar sites.
- **Fake Countdown:** Some online purchases include countdown timers, in most cases countdown timers only add urgency to a sale. Detection of this pattern is possible at least partially because it is possible to search for phrases such as “offer ends in” or “countdown” and warn the shopper of this pattern.
- **Ambiguous Deadlines:** Some online purchases indicate that a product is only on sale for a limited amount of time, but don’t mention a specific deadline. Detection of this pattern is possible at least partially because it is possible to search for phrases such as “for a limited amount of time” and warn the shopper.
- **Low Stock Messages:** Sometimes sites claim that they are low on a particular item. Detection of this pattern is possible at least partially because it is possible to search for phrases such as “only” and “units left” and warn the shopper of this pattern.
- **Deceptive High Demand:** This is similar to the low stock messages. Detection of this pattern is possible at least partially because it is possible to search for phrases such as “in demand” and “in high demand” and warn the shopper of this pattern.

IV. DEVELOPING THE FRAMEWORK

With these definitions established, it becomes possible to categorize the patterns into one of three classifications:

- (1) A suspected pattern that can be detected in an automated way (partially or fully) based on the text, images or HTML in a webpage or website.
- (2) A suspected pattern that can be detected in a manual way (partially or fully) based on the text, images or HTML in a webpage or website.

- (3) A suspected pattern that cannot be detected, based on the fact that there is so much variation in either how the pattern is defined or in how the pattern is implemented.

As all of the researchers involved in this project are teaching on an MSc in Data Science, they have knowledge of a wide range of detection techniques, therefore, a Morphological Matrix approach [41] was undertaken, whereby a table was created listing all of the pattern types on the Y-axis, and listing a range of detection techniques on the X-axis (HTML Parsing, Computational Linguistics, Image Processing, Machine Learning, Data Mining, Compiler Design, Regular Expressions) and a series of three online brainstorming sessions were held to identify which patterns might be detectable using which techniques (if any). To help reach a shared understanding of the patterns, not only were definitions of each pattern shared and discussed, but also images from over 100 websites with dark patterns from the Mathur et al. [32] dataset were presented and discussed. Of all patterns discussed, there was general consensus as to which aspects of patterns could be detected, and to what extent that detection was possible. The full framework is presented below in Table 1 where each pattern presented in Section III is classified as to how it can be detected, as well as some detail as to how such a pattern can be detected (if it can) as shown in the *Rationale* column.

Patterns that can be detected automatically will typically have terms in them such as “opt-in”, “activate”, or “subscribe”. These, and other indicators such as the placement or configuration of images, or in the formulation of the HTML tags, allow for the automated detection of dark patterns. In contrast, there are some web-based activities or transactions that cannot, in and of themselves, be automatically detected, but are sufficiently indicative to suggest the presence of a dark pattern. In these cases the framework proposes the development of an ancillary (or appurtenant) window to highlight to the users that there may be something suspicious occurring in the transaction that they are undertaking. Finally, it is worth noting that, there are some patterns that cannot readily be detected, but may be reported using the reporting feature of the system.

The patterns beyond Brignull canon is the only one where it may be possible to do some form of automated detection on all of the patterns (Fake Activity, Fake Reviews, Fake Countdown, Ambiguous Deadlines, Low Stock Messages, Deceptive High Demand). This may be because these patterns focus almost exclusively on text-based enticements to encourage users to purchase content, and because they use text, it is possible to do searches for specific phrases, for example, “offer ends in”, “for a limited amount of time” or “in high demand”. The one pattern that is slightly different from the others is the Fake Reviews, where instead of searching for a particular phrase on the webpage, we use the entire review to search for that exact same review (or a similar review) on other sites.

TABLE I. DARK PATTERNS DETECTION FRAMEWORK

<i>Category</i>	<i>Pattern</i>	<i>Detection</i>	<i>Rationale</i>
<i>Sneaking</i>	Sneak into Basket	Manual (fully)	Highlight changes in cost
	Hidden Costs	Manual (fully)	Highlight changes in cost
<i>Misdirection</i>	Trick Questions	Automated (partially)	Look for phrases like “opt-in” and “opt-out”, as well as pre-ticked checkboxes
	Misdirection	Cannot be detected	There is too much variation in how this pattern is implemented.
	Confirmshaming	Cannot be detected	There is too much variation in how this pattern is implemented.
	Disguised Ads	Automated (partially)	Look for buttons (noting colour and size) and see which ones link to external sites.
<i>Obstruction</i>	Roach Motel	Automated (fully)	Look for sites with “activate” or “subscribe” links or buttons but with no “deactivate” or “unsubscribe”
<i>Forced Action</i>	Forced Continuity	Cannot be detected	There is too much variation in how this pattern is implemented.
<i>Variegations</i>	Privacy Zuckering	Cannot be detected	There is too much variation in how this pattern is implemented.
	Price Comparison Prevention	Manual (fully)	Highlight if products are displayed with different units of the product
	Bait and Switch	Cannot be detected	There is too much variation in how this pattern is implemented.
	Friend Spam	Automated (partially)	Check if the site asks for email or social media permissions, and notify users.
<i>Beyond Brignull</i>	Fake Activity	Automated (partially)	Look for phrases like “other people are viewing this item now”.
	Fake Reviews	Manual (partial)	Select the review and search for it on other sites.
	Fake Countdown	Automated (partially)	Look for phrases like “offer ends in” or “countdown”
	Ambiguous Deadlines	Automated (partially)	Look for phrases like “for a limited amount of time”
	Low Stock Messages	Automated (partially)	Look for phrases like “only” and “units left”
	Deceptive High Demand	Automated (partially)	Look for phrases like “in demand” and “in high demand”

Some patterns will have words or images that make them easy to identify (“opt in”, “offer ends soon”, “in demand”, etc.) and therefore we can say that they are automatically detectable (either partially or fully). And, in contrast, some patterns are implemented in such a range of different ways depending on the particular interface (and the definitions of some patterns vary in different research literature), that they are impossible to consistently detect, so we classify these as “Cannot be detected”. Other patterns require human judgement, such as determining if using pre-ticked checkboxes is being deceptive, or if the site is asking for security permissions, and so we classify these as being detectable manually (either partially or fully). To help recognise the patterns that can potentially be manually detected, the proposed system will allow the user to display an ancillary window that will help highlight some potential issues of concern on a given webpage or website. The new window can display things like:

- The percentage of the webpage that is visible in the browser window, to ensure the user is aware that there may be instructions or options that are not visible on the current page, but are elsewhere on the page.
- The total number of checkboxes on the page, and the number that are pre-ticked.
- The total number of radio buttons on the page, and the number that are pre-ticked.
- The shopping basket total, that will be zero if there are no items.
- A “fake review detection” tool that allows a user to select the text of a review, and to automatically search for that text elsewhere on the web.
- Highlight the number of links on the page, noting which are from text and which from images (to help detect potential Disguised Ads).

- Highlight which tick boxes or radio buttons are concerned with privacy issues, looking for words such as “privacy” or “GDPR” .
- Indicate if the current webpage or website has already been reported as having a dark pattern.

Further, to help users locate suspected dark patterns on a webpage, the system will provide two modes of operation:

- (1) where the system highlights all of the areas on that webpage to show suspected patterns on the page with suitable pointers, and
- (2) if the user clicks on a particular type of issue on the auxiliary window, only those areas on the page will be highlighted, for example, if the user selects the “Radio Buttons” section of the panel, then all of the radio buttons on the webpage will be highlighted with pointers.

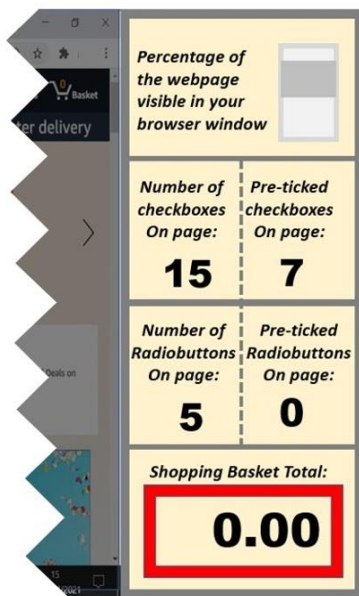


Figure 1. Appurtenant Window with Page Details

Two additional elements of the proposed system are the Reporting and Educational features:

- The *Reporting Feature* is designed to compensate for the fact that some patterns are difficult (or impossible) to detect, and it will allow users to record and report websites and webpages that they suspect have dark patterns. For example, if a user feels that they have been a victim of Forced Continuity, they can report the webpage or website, and indicate which pattern they feel is present.
- The *Educational Feature* which is designed to educate the users on each of the main dark patterns, as well as the variation among different researchers. This feature will help the users appreciate why they are being

warned about a particular feature on a website as well as giving them sufficient information to allow them to accurately categorize patterns that they encounter if they wish to report them. It is envisioned that a central part of this feature will consist of a series of videoed micro-lessons.

V. IMPLEMENTATION AND LIMITATIONS

The goal of this research is to define a collection of dark patterns, and to explore whether or not it is possible to develop a framework to detect these dark patterns - in an automated way, a manual way, or not at all. The detection process not only categorizes whether each pattern is detectable, but it also describes to what extent it is detectable, and suggests some ways it might be detected. The development process of framework was as a result of the brainstorming sessions, and these crucially categorized the patterns into three groupings:

1. Automated Detection ("Disguised Ads", "Friend Spam", "Roach Motel" and "Trick Questions")
2. Manual Detection ("Hidden Costs", "Price Comparison Prevention", "Sneak into Basket")
3. Cannot be Detected ("Bait and Switch", "Confirmshaming", "Forced Continuity", "Misdirection", "Privacy Zuckering")

To help confirm the analysis process, an initial prototype system has been developed using the Python programming language which provides ample software libraries for web crawling and web scraping, specifically the HTMLparser and URLOpen libraries were used in this case. The system was developed as a plug-in for the Google Chrome browser and was able to detect four patterns were selected to be implemented, "Trick Questions", "Roach Motel", "Friend Spam", and "Low Stock Messages" were chosen as they are the most straightforward to implement, since that have been classified as "Automated (partial)" and "Automated (fully)" in the above table. These four were implemented, and were tested using over 60 of the dark patterns from the Mathur et al. [28] dataset, and the prototype was able to successfully detect all three of these patterns, each with significant variation. Three key takeaways from the prototype development process were as follows:

1. When testing the prototype system with some users it became evident that the terminology itself was proving to be a barrier to understanding the purpose of the system. Although the participants had experienced the phenomena of being pressured into purchasing goods online, the term "Dark Patterns" was unfamiliar to them, and two of the names of the patterns: "Roach Motel" and "Friend Spam" were equally opaque to the users, proving to be moreso confusing than enlightening. Future development will change some of the terms to more descriptive one, including changing "Dark Patterns Detector" to "Online Shopping Tricks

Detector”, changing “Roach Motel” to “Hard to Unsubscribe”, and changing “Friend Spam” to “May use your addressbook”.

2. A rudimentary Optical Character Recognition (OCR) system was developed to read text off the images on webpages to determine if they have messages that could be considered to be Dark Patterns, for example, text saying “Only a Limited Amount of Stock Left”. The implementation proved to be highly effective in terms of reading text from the images, but slowed down the overall detection process significantly, and particularly for websites that had a lot of images on them, it delayed the detection process from being almost instantaneous into taking almost 10 minutes to complete the process.
3. Perhaps one of the most interesting outcomes of the prototyping process was that it allowed the researchers to interrogate their fundamental understanding of the notion of a Dark Pattern. Most websites include some forms advertising, which are not the same as dark patterns, for example, some of the test sites included phrases such as “Customers who bought this product also bought ...” which were classified as Dark Patterns by the system, as they are similar to a “Fake Activity” which might say something like “Other Customers are looking at this product”. After much discussion it became clear that this is just advertising, and in particular, it is persuasive advertising, which is similar to Dark Patterns, but they differ in that they do not rely on pressuring or confusing the customers.

In terms of the limitations of this research, perhaps the most serious one is the fact that five of the patterns (“Misdirection”, “Confirmshaming”, “Forced Continuity”, “Privacy Zuckering”, and “Bait and Switch”) have been classified as “Cannot be detected”. If these cannot be detected, it significantly limits the efficacy of the final tool, therefore a thorough exploration of the Mathur et al. [32] dataset is planned to determine if there are any implicit characteristics associated with these five patterns that can be used to detect them (either automatically or manually), as well as a number of further brainstorming sessions.

It is also worth noting that that the full implementation of this framework will result in some additional challenges, for example, some sites have a special file called Robots.txt that prohibits the use of web scraping, and it is also the case that some sites use technologies that make them more difficult to parse, for example, frames or webpages implemented in Javascript or CSS.

Finally, another consideration is that many shoppers use mobile applications instead of websites to purchase products and services, and the techniques outlined so far would be ineffective on these applications.

VI. CONCLUSIONS AND FUTRE WORK

This paper presented a framework for the detection of web-based dark patterns and an accompanying proposed software tool. It begins with a review of some of the key literature in this field, which highlights some of the reasons for the success of dark patterns, as well as their ubiquity. It follows this with an explanation of some of the key dark patterns, and a categorization of the patterns as being in one of the following three classifications:

1. A suspected pattern that can be detected in an automated way (partially or fully), in other words there is some characteristic either in the text, images or HTML of a webpage or website that indicates that it is a dark pattern.
2. A suspected pattern that can be detected in a manual way (partially or fully), in other words there is some characteristic either in the text, images or HTML of a webpage or website that indicates that there is potential for dark pattern on this page or site, but because it cannot be detected definitively, the potential pattern is highlighted to the user.
3. A suspected pattern that cannot be detected, in other words there is so much variation in either how the pattern is defined or in how the pattern is implemented, there is no direct way of detecting it just using web crawling and web scraping techniques.

This classification, in turn, leads to the design of a proposed software tool with the ability to detect patterns from category 1, and to highlight potential instances of patterns from category 2. For those patterns in category 3, even if there is no obvious way to identify them, nonetheless, it is important to deal with them in some way, therefore additional features are required for the system, a *Reporting feature* to address instances of patterns for category 3, as well as an *Educational feature* to create awareness about dark patterns in general.

Future work will focus on full implementation of the software tool and the inclusion of the Reporting and Education features. The Reporting features of the system are envisioned to work either in *stand-alone mode*, or *shared mode*. In stand-alone mode the reporting process is recorded locally on the user’s own computer as a series of XML files, whereas in shared mode, the user can share their suspicions about potential dark patterns with other users also using the system, and they can also label and add a description to the suspected pattern.

The Educational features will consist of a series of micro-lessons describing the range of dark patterns. Also, a series of pop-up windows will be developed with simple explanations (and links to examples) of a specific pattern will be developed, to remind the users about the key characteristics of each specific pattern.

Finally, the framework provides a way forward to deal with dark patterns in a comprehensive and comprehensible manner. This has become more and more important as the

number of services that have become available online continues to grow, and in many cases these services are available only exclusively online. It, therefore, becomes a matter of necessity that as many people as possible are aware of these deceitful patterns, and incumbent on IT practitioners to spread the word about these patterns.

ACKNOWLEDGMENT

The authors of this paper and the participants of the Ethics4EU project gratefully acknowledge the support of the Erasmus+ programme of the European Union. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

- [1] A. Curley, D. O'Sullivan, D. Gordon, B. Tierney, I. Stavrakakis, "The Design of a Framework for the Detection of Web-Based Dark Patterns". ICDS 2021: The 15th International Conference on Digital Society, Nice, France, 18th - 22nd, July 2021.
- [2] A. Narayanan, A. Mathur, M. Chetty, M. Kshirsagar, "Dark Patterns: Past, Present, and Future: The evolution of tricky user interfaces". Queue, 18(2), pp. 67-92, 2020.
- [3] D. O'Sullivan, D. Gordon, "Check Your Tech – Considering the Provenance of Data Used to Build Digital Products and Services: Case Studies and an Ethical CheckSheet", IFIP WG 9.4 European Conference on the Social Implications of Computers in Developing Countries, 10th–11th June 2020, Salford, UK, 2020.
- [4] H. Brignull, "Dark patterns: Deception vs. honesty in UI design". Interaction Design, Usability, 338, 2011.
- [5] C. M. Gray, Y. Kou, Y., B. Battles, J. Hoggatt, A. L. Toombs, "The dark (patterns) side of UX design". In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1-14, 2018.
- [6] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, S. Wilson "Nudges for privacy and security: Understanding and assisting users' choices online", ACM Computing Surveys (CSUR), 50(3), pp. 1-41, 2017.
- [7] M. Maier, "Dark patterns: An end user perspective". Master's thesis. Umeå University, 2019.
- [8] R. Mehta, R. J. Zhu, "Blue or red? Exploring the effect of color on cognitive task performances", Science (New York, N.Y.), 323(5918), pp. 1226-1229, 2009.
- [9] D. Kahneman, "Thinking, Fast and Slow", Penguin Books, 2011.
- [10] C. Bösch, B. Erb, F. Kargl, H. Kopp, S. Pfattheicher, "Tales from the dark side: Privacy dark strategies and privacy dark patterns". Proceedings on Privacy Enhancing Technologies, 2016(4), pp. 237-254, 2016.
- [11] A. Acquisti, "Nudging privacy: The behavioral economics of personal information". IEEE Security & Privacy, 7(6), pp. 82-85, 2009.
- [12] H. Almuhiemedi, F. Schaub, N. Sadeh, N.I. Adjerid, A. Acquisti, J. Gluck, Y. Agarwal, "Your location has been shared 5,398 times! A field study on mobile app privacy nudging". In Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 787-796, 2015.
- [13] E. Peer, S. Egelman, M. Harbach, N. Malkin, A. Mathur, A. Friuk, "Nudge me right: Personalizing online security nudges to people's decision-making styles". Computers in Human Behavior, 109, 106347, 2020.
- [14] R. Cialdini, "Influence. The Psychology of Persuasion". New York, NY: William Morrow Company, 1984.
- [15] A. E. Waldman, "Cognitive biases, dark patterns, and the 'privacy paradox'". Current opinion in psychology, 31, pp. 105-109, 2020.
- [16] D. Ariely, G. Loewenstein, D. Prelec, "Coherent arbitrariness: Stable demand curves without stable preferences". The Quarterly journal of economics, 118(1), pp. 73-106, 2003.
- [17] I. Adjerid, A. Acquisti, L. Brandimarte, G. Loewenstein, "Sleights of privacy: Framing, disclosures, and the limits of transparency". In Proceedings of the ninth symposium on usable privacy and security, pp. 1-11, 2013.
- [18] A. Acquisti, J. Grossklags "Losses, gains, and hyperbolic discounting: An experimental approach to information security attitudes and behavior". In 2nd Annual Workshop on Economics and Information Security-WEIS (Vol. 3, pp. 1-27), 2003.
- [19] J. Puauschunder, "Towards a utility theory of privacy and information sharing and the introduction of hyper-hyperbolic discounting in the digital big data age". In Handbook of research on social and organizational dynamics in the digital era, pp. 157-200, IGI Global, 2020.
- [20] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. Leon, L. F. Cranor, "I regretted the minute I pressed share" a qualitative study of regrets on Facebook. In Proceedings of the seventh symposium on usable privacy and security, pp. 1-16, 2011.
- [21] A. Chernev, U. Böckenholt, J. Goodman, "Choice overload: A conceptual review and meta-analysis". Journal of Consumer Psychology, 25(2), pp. 333-358, 2015.
- [22] S. Jilke, G. G. Van Ryzin, G. G. S. Van de Walle, "Responses to decline in marketized public services: An experimental evaluation of choice overload". J. of Public Administration Research & Theory, 26(3), pp. 421-432, 2016.
- [23] K. Nagar, P. Gandotra, "Exploring choice overload, internet shopping anxiety, variety seeking and online shopping adoption relationship: Evidence from online fashion stores". Global Business Review, 17(4), pp. 851-869, 2016.
- [24] G. A. Veltri, A. Ivchenko, "The impact of different forms of cognitive scarcity on online privacy disclosure". Computers in human behavior, 73, pp. 238-246, 2017.
- [25] T. Alashoor, R. Baskerville, "The privacy paradox: The role of cognitive absorption in the social networking activity". In Thirty Sixth International Conference on Information Systems, Fort Worth, Texas, USA, pp. 1-20, 2015.
- [26] A. Acquisti, L. Brandimarte, G. Loewenstein, "Privacy and human behavior in the age of information". Science, 347(6221), pp. 509-514, 2015.
- [27] S. Kokolakis, "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon". Computers & security, 64, pp. 122-134, 2017.
- [28] P. A. Norberg, D. R. Horne, D. A. Horne, "The privacy paradox: Personal information disclosure intentions versus behaviors". Journal of consumer affairs, 41(1), pp. 100-126, 2007.
- [29] GDPR, EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament, Article 4, 2016, <https://gdpr-info.eu/art-4-gdpr/>
- [30] S. Zuboff, "The Age of Surveillance Capitalism: The Fight for Human Future at the New Frontier of Power". London: Profile Books, ISBN 978-1-7881-6316-3, 2019.
- [31] M. Van Otterlo, "Automated experimentation in Walden 3.0.: The next step in profiling, predicting, control and

- surveillance". *Surveillance & society*, 12(2), pp. 255-272, 2014.
- [32] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, J. M. Chetty, A. Narayanan, A. "Dark patterns at scale: Findings from a crawl of 11K shopping websites". *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp. 1-32, 2019.
 - [33] M. Nouwens, I. Liccardi, M. Veale, D. Karger, L. Kagal, "Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence". In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13), 2020.
 - [34] M. Chromik, M. Eiband, S.T. Völkel, D. Buschek, "Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems", *IUI Workshops* (Vol. 2327), 2019.
 - [35] L. Di Geronimo, L. Braz, E. Fregnan, F. Palomba, A. Bacchelli, "UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception", *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 2020.
 - [36] P. Grassl, H. Schraffenberger, F. Borgesius, M. Buijzen, "Dark and bright patterns in cookie consent requests". [10.31234/osf.io/gqs5h](https://doi.org/10.31234/osf.io/gqs5h), 2020.
 - [37] K. Kollnig, S. Datta, M. Van Kleek, "I Want My App That Way: Reclaiming Sovereignty Over Personal Devices", 2021, arXiv preprint [arXiv:2102.11819](https://arxiv.org/abs/2102.11819).
 - [38] B. Chugh, P. Jain, "Unpacking Dark Patterns: Understanding Dark Patterns and Their Implications for Consumer Protection in the Digital Economy". *RGNUL Student Research Review Journal*, 7, 23, 2021.
 - [39] K. Bongard-Blanchy, A. Rossi, S. Rivas, S. Doublet, V. Koenig, G. Lenzini, "I am Definitely Manipulated, Even When I am Aware of it. It s Ridiculous!--Dark Patterns from the End-User Perspective". arXiv preprint [arXiv:2104.12653](https://arxiv.org/abs/2104.12653), 2021.
 - [40] R. Steiner, "Dark Patterns" . [Online]. Available from: <https://www.fyresite.com/dark-patterns-a-new-scientific-look-at-ux-deception/>, 2021.06.24
 - [41] R. G. Weber, S. S. Condoor. "Conceptual design using a synergistically compatible morphological matrix." In *FIE'98. 28th Annual Frontiers in Education Conference. Moving from 'Teacher-Centered' to 'Learner-Centered' Education. Conference Proceedings* (Cat. No. 98CH36214), vol. 1, pp. 171-176. IEEE, 1998.

Study for In-Vehicle-Network and New V2X Architecture by New IP

Lin Han, Lijun Dong, Richard Li

Futurewei Technologies, Inc.

Santa Clara, California, U.S.A

email: {lin.han, lijun.dong, richard.li}@futurewei.com

Abstract— For many In-Vehicle-Network (IVN) and Vehicle-to-everything (V2X) applications in the latest vehicle, the higher Quality of Service (QoS) and more deterministic networking are mandatory requirements. The paper proposes an architecture to support latency sensitive communication that is based on New IP technology. The new architecture and technologies can provide the End-to-End (E2E) Latency Guaranteed Service (LGS) and Bandwidth Guaranteed Service (BGS) for any granularity of IP flow(s). It can be used for IVN and V2X communication combined with 5G for future Internet. This paper will use IVN as an example to prove that the New IP can replace other legacy protocols and is able to provide satisfactory service in terms of the critical QoS metrics (Bandwidth, Latency, Jitter and Packet loss). The paper will analyze the challenge of latency requirements for IVN, it focuses on the design of new IVN control plane and data plane especially queuing and scheduling. The theoretical latency analysis, estimation and experimental verification are provided.

Keywords- IVN; V2X; TCP; IP; UDP; QoS; New IP; Deterministic Networking; In-band signaling; Guaranteed Service; Class Based Queueing, Priority Scheduling; Cyclic Queueing, End-to-End; Traffic Shaping; Congestion; Packet loss; Bandwidth; Latency; Jitter; eMBB; mMTC; uRLLC

I. INTRODUCTION

This paper is an extended version of [1], which investigates the latency requirements for IVN, proposes a New IP based IVN architecture, and presents a detailed study and emulations. The paper will provide more details about the New IP based V2X architecture, the algorithms, and experimental results.

Recently, a trend in vehicle industry is that electrical or hybrid motors are gradually replacing the combustion engine and power transmission. The major components of Electrical Vehicle (EV) are battery and electrical motors. They are simpler, more modular, and easier to be manufactured with standard and thus lower the manufacturing threshold and cost. This results in tougher competitions in other areas, such as Tele-driving, Self-driving, Infotainment System, etc. All those advanced futures are computing driven and require advanced networking technologies in following two areas:

- In-Vehicle-Network (IVN): this is the network inside vehicle to connect different electronic devices, such as Sensors, Actuators, Electrical controller unit (ECU), GPS, Camera, Radar, LiDAR, Embedded computer, etc.
- Vehicle-to-Everything (V2X): This is a technology that allows moving vehicle to communicate with other moving vehicles, the traffic control system along roads, and everything in Internet, such as Cloud, home, environment,

people, etc. The traditional V2X term only represents the wireless technologies DSRC defined in IEEE802.11p [2], and C-V2X defined in 3GPP [3]. DSRC is a modification of Wi-Fi and allows wireless devices communicate directly without intermediate device. C-V2X supports two modes: Direct C-V2X (Devices communicate directly) and Indirect C-V2X (Device communicate via wireless network). In this paper, V2X is defined as a general term that is End-to-End communication between any applications within a car and another application running outside of that car, that application could be running in another car, in a cell phone, in cloud or in Internet.

There are different types of applications using IVN or V2X. Based on the requirements for network, traffic can be categorized as three types:

- The time sensitive: For this type of communication, the latency requirement is stringent, but the data amount is limited. This includes the communication for sensor data, control data, such as the control for powertrain system, braking system, security system, etc. The data rate is up to Mbps per flow. This type of traffic normally could be within a car on top of In-Vehicle-Network (example a), it could also be between applications in a car and remote applications on device outside the car using V2X (example b):
 - a. For Self-driving car, some critical sensor data and control data are very time sensitive, the IVN must provide the guaranteed service for shortest E2E latency and zero packet loss.
 - b. Tele-driving system will control a car remotely by human being, or by an automatic AI system in cloud. The feedback data from a car and associated control signal from remote site must experience the shortest latency.
- The bandwidth sensitive: For this type of communication, the latency requirement is not stringent, but the data amount is higher. It includes GPS display, Radar, LiDAR data feeding. The data rate could be up to tens of Mbps per flow. Like the 1st type traffic, some of this type of traffic is within a car, but some is between a car and a remote application.
- Best-Effort: This is the traditional IP traffic that is not belonging to above two types. Network will deliver the traffic to destination without any guarantee.

For above three types of traffic, the 1st one is the most challenging to support by the current technologies for V2X and Internet. This is because the current V2X only addresses the wireless technologies by DSRC or C-V2X but does not

consider other wired network segments. From the perspective of E2E effects for V2X, the latency, jitter and packet loss happened in the segments of wired network are not negligible. Since the IP network can only provide the Best-Effort service, the queuing latency and packet loss due to congestion in IP network is very normal.

The paper proposes to use New IP technology for new architecture of IVN and V2X. New V2X architecture will integrated 5G and New IP to obtain the true E2E guaranteed service in terms of bandwidth, latency, jitter, and packet loss. The remained paper has three parts:

- 1st part discusses the basics of New IP. Section II introduces the New IP. Section III will talk about New IP based V2X architecture.
- 2nd part focus on the new IP based IVN details that includes Sections IV to IX. Section IV reviews the current technologies for IVN. Sections V, VI and VII will discuss the basics, architecture for control plane, and data plane respectively. Section VIII addresses the latency analysis and estimation. Section IX describes the network modeling and experiments.
- 3rd part is in Section X that will describe the conclusions.

II. NEW IP INTRODUCTION

New IP is a broad technology set dedicated to solving requirements from future Internet, it is still in research stage and not mature. It was first proposed in ITU [4], and some research papers were published [5][6][7].

Compared with the existing IPv4 and IPv6, New IP has many forward-looking visions and will support some new features, such as

- Free Choice Addressing. Different size of IP address can be used for different use case. For the scenario that packet header overhead is a concern, such as in IOT network, a shorter than IPv4 or IPv6 address can be selected. For the extreme secured environment, invisible source address or longer than 128-bit randomized address can be used. This paper will not discuss this feature in detail. We still assume to use IPv4 for IVN. For IVN experiment in Sections VIII to IX, 32-bit IPv4 address is used for simulation.
- Deterministic E2E IP service. It can provide the guaranteed service to satisfy the pre-negotiated Service Level Agreement (SLA). New IP can be used for IVN and E2E V2X since both have very strict QoS requirements especially in bandwidth, latency, jitter, and packet loss that the current IP technology cannot meet.

New IP can coexist with other technologies in Internet, the traditional IP packets can still be processed and delivered in New IP networks. The interworking between New IP and IP networks can be easily provided by a proper gateway device between different networks. Migration to New IP network can take step by step gradually, we only need to upgrade the

network required to support new services that traditional IP network cannot support, so, the cost is limited.

As a summary, New IP is for Future Internet to provide services that the current Internet cannot provide. It is like the New Radio (NR) for 5G [8] in objectives, solutions, and technologies, see TABLE I for comparison.

TABLE I. 5G NR for 5G and New IP for Future Internet

	5G	Future Internet
Purpose and Requirements	<ul style="list-style-type: none"> • eMBB [9] • mMTC [9] • uRLLC [9] 	<ul style="list-style-type: none"> • Ultra-high throughput • All things connected • High Precision Communication
Solutions	<ul style="list-style-type: none"> • New Radio (5G NR) • Service Based Architecture (SBA) [10] 	<ul style="list-style-type: none"> • New IP
Technologies	<ul style="list-style-type: none"> • New spectrum • MIMO [8] • New protocol stack at UE • 5G NR QoS [8] • Grant Free Dynamic Scheduling 	<ul style="list-style-type: none"> • Flexible addressing • Network Layer Multiple path • New protocol stack at host and UE • In-band signaling • New queuing and scheduling

There could be different technologies developed for New IP for different use cases. The paper [7] has proposed key New IP technologies to realize the E2E guaranteed service for Internet, details are as following:

- In-band signaling. This is a control mechanism to provide a scalable control protocol for flow level guaranteed service. The key part of In-band signaling is that the control messages are embedded into the user data packets. With such binding, when the user data packets travel through a network, the control messages can be fetched by each network device on the path and control the behaviors of expected devices accordingly. Since all QoS metrics (bandwidth, latency, jitter, packet loss) are majorly determined by each network device on how user data packets are processed, accurately control network devices on path is the best way to achieve the best service a network can provide to applications. In traditional way, such controls are provided by separate protocols (sometimes called out-of-band signaling), the complexity is high and the scalability are limited. Through in-band signaling, the QoS path setup, SLA negotiation, Resource Reservation, QoS forwarding state report and control are accomplished without running extra control protocol like RSVP [11] for IP, or Stream Reservation Protocol (SRP) [12] for TSN [13]. The details of In-band signaling is described in [7].
- Class based queuing and scheduling. It uses the concept of Class as defined in Differentiated Service (DiffServ) [14] to identify different types of traffic. Different class of traffic is queued into different queues for differentiated service. Priority Queuing (PQ) combined with Deficit Weighted Round Robin (DWRR) or any other Weighted

Fair Queuing (WFQ) are used. Compared with other algorithms, this is the simplest to be implemented in high-speed hardware, and can achieve very satisfactory QoS in bandwidth, latency, jitter, and packet loss ratio. It also solves the scalability issue in Integrated Service (IntServ) [15] where the per-flow queueing was used.

- New TCP/UDP transport stack for end devices. The current TCP/UDP transport protocol stack was designed based on the best-effort service from IP. Enhanced protocol stacks are expected to obtain the benefits if the network can provide guaranteed service while keep the backward compatibility.

Above technologies set can be used by different combinations for IVN and V2X. For V2X, all technologies could be used. But for IVN, control methods (such as SDN controller) other than In-band signaling can also be used.

III. NEW IP BASED V2X ARCHITECTURE

5G has defined that the End-to-End latency (uRLLC) is the Round-Trip Time (RTT) of IP packets transmitted from User End Device (UE) to the N6 interface in the 5G network [16]. The N6 interface is the reference point between UPF and Data Network (DN). It is obvious that uRLLC does not include the latency occurred in UE and DN.

The latency in UE is that when IP packet left application, it takes some time before the scheduler will send the packet to outgoing physical interface, this delay is significant when the UE has multiple applications running since different IP flow will compete the resource to get service from Operating System.

The latency for Data Network is the time spent for IP packets traveling from N6 interface to the IP (IPv4 or IPv6) destination. The destination can be any IP address in Internet, for example, a server inside a cloud. Normally, this latency is significant and is much bigger than inside a 5G network.

Same behaviors will apply to other QoS characters. The insufficient bandwidth, waiting for resource, and resulted jitter and packet loss happened in DN is normal and significant.

The root cause of above QoS degradation in data network for IP is because all IP packets are treated equally on the path the packet is traveling. Every IP packet is competing for the network resource, this will result in unexpected congestion, queue built up and even packet loss when queue is full. Even there are many technologies to mitigate or fix the problem, such as different congestion avoidance algorithms studied for long time [17], TSN [13], L4S [18], MPLS traffic engineering [19], etc. All these solutions are only working in a specific network but cannot be applied to Internet from real end to end (IP source to IP destination). It is insufficient to solve E2E latency issue in Internet if only considering specific network segments, such as wireless access network by 5G uRLLC [16] or Ethernet network by TSN [13].

The paper proposes to combine New IP technologies with 5G wireless technologies for the new architecture of future V2X communication.

To minimize the latency in UE, a new IP protocol stack is needed for UE. Figures 1 and 2 illustrate these stacks in wired and wireless device. The major changes for the new protocol stack are new socket or API. It is introduced for applications that require new service which is different with the traditional best-effort service using traditional socket. The new socket will pass application's service expectation to the network. The different flow with different service expectation will be queued to different queues, Latency Guaranteed Service (LGS) queue, Bandwidth Guaranteed Service (BGS) queue, or Best-Effort (BE) queue. System scheduler will serve different queue based on the priority and resource. Signaling Process module is to process the setup and forwarding state for in-band signaling. M-path control is for the multi-path support, it could split one flow into different network path, or replicate one flow couple of times to send to multiple network path. Multi-path feature can either increase the total bandwidth for application or compensate the packet loss due to the physical failure on one path. For a wireless device, an extra module will provide the interworking between New IP and New Radio (Figure 2). This module will coordinate the mapping between L3 multi-path and multiple Bearer introduced in 5G NR.

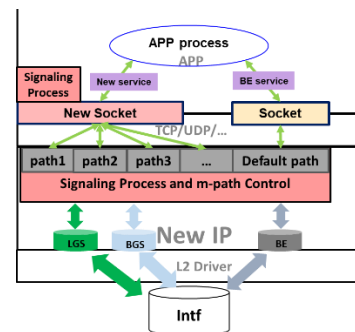


Figure 1. The New IP protocol stack for a computer or ECU.

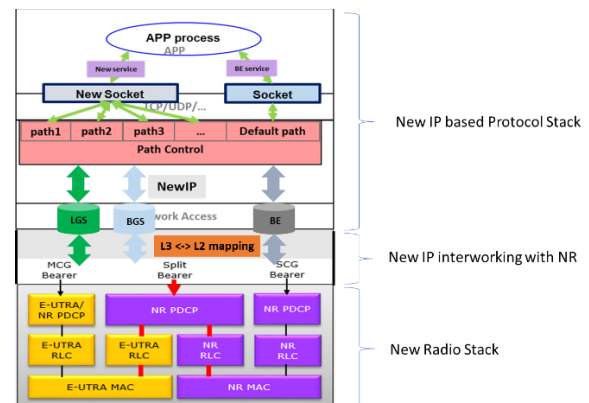


Figure 2. The New IP protocol stack integrated with 5G New Radio (NR) protocol stack for wireless device.

To minimize the latency in Data Network, the in-band signaling initiated from UE can pass through all network and reach the IP destination. This mechanism provides a simpler and more scalable control mechanism to provision a true end-to-end guaranteed service for any IP based application. When encountering a heterogeneous network (Ethernet, MPLS or other types), the in-band signaling carried in IP packet can be retrieved and used to interwork with other protocols, such as SRP for TSN, RSVP-TE for MPLS, etc.

Figure 3 illustrates New IP enabled V2X architecture in future Internet where IVN, 5G and wired data network in Internet all need New IP enabled, with such architecture, the true E2E deterministic service can be realized. It should be noted, for the case of directly communication (DSRC or Direct C-V2X), the architecture will only have IVN and V2X.

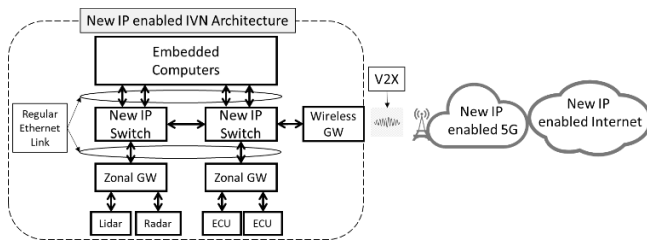


Figure 3. New IP enabled IVN architecture in future Internet.

Compared to the traditional V2X architecture which only address the wireless technologies, the new architecture shown in Figure 3 has New IP enabled networks including IVN, 5G and Internet. Only after the integration of those new IP enabled network, the true E2E service can be guaranteed for new applications.

In above picture, how to use New IP for each segment of network has many technical details. Due to the space limit, the paper cannot go to details for each, but will only focus on the case that New IP is used for IVN. We select IVN as an example is because the traditional IVN did not use IP, it normally uses some legacy protocols because of the stringent latency and packet loss requirement. The paper will demonstrate and prove that the New IP can provide the satisfactory deterministic service that the traditional IP cannot provide, and this service will satisfy the latency requirement of IVN.

IV. REVIEW OF CURRENT IVN TECHNOLOGIES

The section will brief the networking protocols used in current IVN and analyze the latency requirement for IVN.

A. Network technologies in current IVN

Most of the current IVN uses the legacy protocols, such as Local Interconnect Network (LIN) [20], Controller Area Network (CAN) [21], FlexRay [22]. These are specifically L2 technologies, they use the special designed physical media, signaling to manage strictly and timely for data to satisfy the requirements for communications inside car.

When more and more IP based applications come to IVN, the disadvantage of above legacy protocols is obvious. Its cost is normally higher than the TCP/IP plus Ethernet based network, IP based application must re-write the interface with new underlayer network if it is not Ethernet. AutoSAR [23] has proposed all IP based interface for IVN, and IP based IVN was proposed in [24][25].

However, without special technology, traditional TCP/IP and Ethernet cannot satisfy the requirement of IVN in terms of QoS. That is why IEEE TSN [13] was also proposed for IVN [26].

B. Requirement for IVN

The most important requirement in terms of QoS for IVN is the communication latency, jitter, and packet loss ratio.

The latency is crucial to the safety of vehicle and will determine if a new technology can be used in IVN. So far, there is no industry standard or requirement for the latency for IVN. Below are some existing publications about the topic:

- From the perspective of fastest human reaction time, the IVN latency must not be slower than that. It is said the fastest human reaction time is 250ms [27]. Some papers gave lower values but not shorter than 100ms if human brain is needed to process the input signal.
- The paper [26] mentioned the latency for control data must be less than 10ms. The papers [24] and [28] said the latency for control data must be less than 2.5ms.

Based on all available analysis, it is safe to assume that the qualified IVN must support the E2E latency not bigger than 2.5ms. During this short time, a car with a speed of 200 km/s will only move 0.138m.

There is no requirement for the jitter from current research. Theoretically, jitter can be removed by buffering technology when the maximum latency is within the target.

The zero-packet-loss is expected for control data. In a packet network (Ethernet or IP), the packet loss is normally caused by two factors: (1) the congestion in network (2) physical failure, such as link, node, hardware. The 1st factor has much higher occurrence probability and higher packet loss ratio than the 2nd factor. Thus, it must be eliminated for control data in New IP based IVN. The loss by 2nd factor can be mitigated or eliminated by sending the same data to two or multiple disjointed paths to reach the same destination, and/or, sending the same data more than one time as long as the time period is chosen below the upper bound of the latency.

V. THE IVN ARCHITECTURE - INTRODUCTION

The new architecture of IVN is based on New IP technologies and consists of Control plane and Data Plane. This section will discuss some basics for architecture.

A. Topologies

The topologies of new IVN can be any type, but to reduce the complexity and to provide a redundant protection, the paper proposes to use two topologies, one is the Spine-Leaf

topology, and another is Ring topology. They are shown in Figure 4 and Figure 5, respectively.

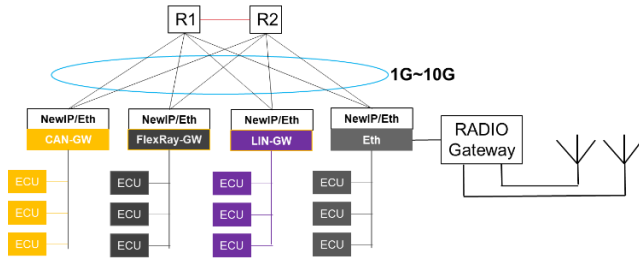


Figure 4. The Spine-Leaf IVN topology.

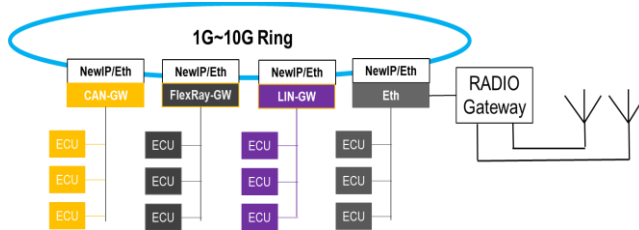


Figure 5. The Ring IVN Topology.

In the topologies illustrated in Figures 4 and 5, there are always two disjointed physical paths between any network devices. Also, the Ethernet Bus is supported. The advantages of such design are:

- The protection of physical link. Any failure of any link does not completely stop the communication.
- The higher reliability for zero packet loss. Multiple paths of New IP can be used to transport critical packet to two paths to compensate possible packet loss due to temporary failure or fault in one physical transmission media.
- Ethernet Bus can make the plug-and-play possible for most of sensors, ECU, computers, etc.

B. Network Device and Link

The network device can be either IP Router or Ethernet Switch. IP router is more powerful to provide more features in networking, such as more flexibility in routing and network state changes, higher link utilization, secured communication, etc.

When Ethernet Switch is selected, DPI (Deep Packet Inspection) should be configured to check the IP level information (address, port, protocol, DSCP values) for admission control for IP flows.

The Physical Link and protocol can be any type of Layer 2 link, Normal Ethernet or IEEE802.1 with the speed higher than 100 Mbps is minimum, and 1G ~10G is better to achieve a shorter latency. There is no need to select any special IEEE802.1Q serials, such as TSN. This is one of the advantages of the new architecture compared with TSN and other legacy protocols (LIN, CAN, FlexRay, etc). It not only provides more flexibility in device development and technology selection, but also save the cost for V2X

applications, since IP is more general technology that fits most of existing application's interfaces. In addition to that, IP device is normally cheaper than legacy device especially in higher speed.

C. Backward Compatibility

The legacy protocol LIN, CAN and FlexRay are still supported in the new IVN architecture. As shown in Figures 4 and 5, legacy ECUs used for legacy protocols can still be attached to the legacy bus. The New IP based network node will have an interworking function to support the legacy protocols. Figure 6 illustrates a Gateway board with two interfaces: Ethernet and FlexRay, and another board only has Ethernet interface. Two board can be connected by Ethernet interface. The ECU attached to the FlexRay bus can communicate with any application running in both boards on top of New IP.

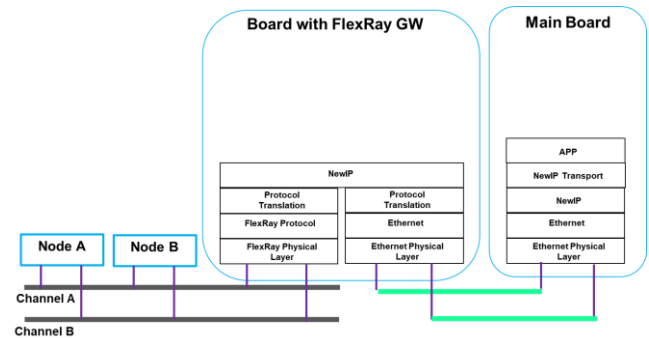


Figure 6. Interworking between Ethernet and FlexRay.

D. New Service

The new service provided by New IP based IVN is "E2E flow level guaranteed service for bandwidth, latency, jitter and packet loss". Following is detail about the new service:

- The E2E is defined as "From Application(s) of one end-user device to other Application(s) of another end-user device. For IVN, the end-user device is any device connected to IVN that supports TCP/IP protocols, and application is running on top of TCP/IP, such as TCP/IP capable ECU, Embedded computer, Infotainment system, Mobile device, etc.
- The Flow can be any granularity, for example, it can be an IP flow defined by 5 tuples (source/destination address, source/destination port number, protocol), or a group of flows defined by less tuples, such as source/destination address.
- The Guaranteed service means that the service provided by system will go through some crucial steps like Service Level Agreement (SLA) negotiation or provisioning, admission control and user traffic conformity enforcement, etc. After all procedures are accomplished, the promised service will meet the negotiated bandwidth, latency, jitter, and packet loss defined in SLA.

- Different application may need different guaranteed service. For example, critical sensor and control data may need the guaranteed service for both bandwidth and latency. The new service is like the service for Scheduled traffic and Real-time traffic defined in FlexRay [22]. For these types of traffic, the strictest service is needed to achieve the minimum latency, jitter, and packet loss ratio. almost all other type of data does not need any guaranteed service, the best-effort service is good enough. For any application, weather it needs the new service is case by case and up to the application's requirement from the networking.

VI. ARCHITECTURE- CONTROL PLANE

This section discusses the aspects of control plane for new IVN architecture including the Control Plane Candidates, and Control Plane Functions.

A. Control Plane Candidates

The control plane could select the following candidates:

- Central controller: such as SDN controller or network management controller. For IVN, it is normally a controller's responsibility to provision some basic function for IVN, such as address assignment, routing protocol configuration (for dynamic routing) and static routing table installation (for fast and simple system boot up). Central controller can also be used for the static provisioning for the guaranteed service, such as scheduled and real-time traffic configuration on ECUs,
- In-band signaling protocol [7] is an alternative control method distributed to all network nodes. It can be used for connections between IVN and cloud for critical data in V2X scenario, it can also be used in IVN for dynamic service state report, network state OAM and network problem diagnosis. In-band signaling is not mandatory for communication within IVN.

B. Control Plane Functions

In addition to the static provisioning from a central controller described in A, another key function for the control plane to achieve the guaranteed service support is the Admission Control. All flows requesting new service, except the Best Effort, must obtain the approve for the admission from central controller or from in-band signaling process. This includes three steps:

- An application requesting new service specifies the expectation of service type (BGS, LGS), the traffic pattern (rate specification) and expected End-to-End latency.
- System (Central controller or the network device) will process the request and try to reserve the resource for the flow, and notify the application about the CIR (Committed Information Rate), PIR (Peak Information Rate), bounded end-to-end latency and jitter values, packet loss ratio, etc.
- The application agreed the offered service will send traffic according to the system notification, i.e., send traffic no

more than CIR, and monitor the notification from network to adjust the traffic pattern accordingly.

VII. ARCHITECTURE - DATA PLANE

This section discusses the aspects of data plane for new IVN architecture including the Protocol Selection, Queuing and Scheduling Algorithm, Traffic shaping, Latency estimation.

A. Protocol Selection

As new IVN is IP based, IPv4 is proposed to be the basic protocol for New IP, a protocol extension is needed if in-band signaling is used [29]. All data process, such as forwarding, traffic classification, traffic shaping, queuing, and scheduling, are for IPv4 data. It is noted that New IP's "Free address choice" feature can provide address shorter than IPv4 that can benefit the latency, but it is not discussed here.

B. Traffic Classification and Services

This paper will propose to classify all IVN traffic as four types:

- Scheduled traffic (ST). This type of traffic has fixed data size, exact time of when the data is starting and what is the interval of the data. Normally, all sensor data report and control data belong to this type. Typically, IVN can configure the polling mechanism for all sensors to make use of this type of traffic. The service associated with this type of traffic will get LGS. This type of traffic is classified as EF class in DSCP value defined in DiffServ.
- Real-Time Traffic (RT). This type of traffic has fixed data size, but the time of the data starting, and the data rate is unknow. Normally, all urgent sensor data report and control data belong to this type. IVN can configure the critical sensors to send data to controller in the situation of emergency and the polling mechanism did not catch the latest data changes. The service associated with this type of traffic is also LGS. But the latency and jitter might be a little bigger than for the ST depending on the algorithm and burst of RT. This type of traffic is classified as AF4x class in DSCP value.
- Bandwidth Guaranteed Traffic. This type of traffic has special requirement from the network bandwidth, but not the latency, jitter, and packer loss ratio. Normally, the IVN software update from cloud, diagnosis data uploading to cloud, on-line gaming and streaming for infotainment system, etc., belong to this type. It can be classified as any AFxy class (other than EF and AF4x) in DSCP value.
- Best-Effort Traffic. This is a default class of traffic, all applications that do not require any special treatment from network perspective can be classified as this type of traffic, Best Effort Class is used.

There are four types of services in IVN corresponding to the above four type of traffic. TABLE II shows QoS Characters and Use Case for different type of services. Both

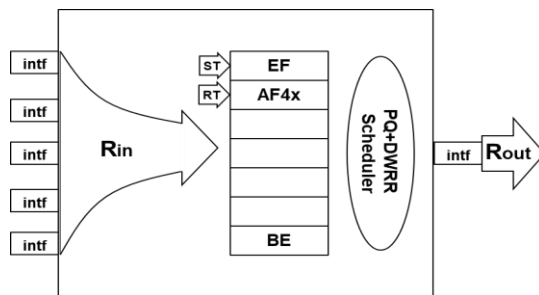
Scheduled Traffic (ST), Real-Time Traffic (RT) are treated by Latency Guaranteed Service (LGS) as described in [7]. The traffic that only needs the bandwidth guarantee is treated by Bandwidth Guaranteed Service (BGS). Other types of traffic are treated by Best-Effort Service (BES)

TABLE II. FOUR TYPE OF SERVICE AND QOS CHARACTERS

Service Type	QoS Characters	Use Case
LGS for Scheduled Traffic	Bandwidth: Network guarantees the bandwidth is within (CIR, PIR) Latency: Most precise. Network guarantees E2E bounded latency Jitter: Approximately zero Packet Loss: Almost Zero <ul style="list-style-type: none"> Congestion-free Lossless queuing Multi-path to prevent drop from physical failure 	Asynchronous or Synchronous communication: Critical sensor and control data
LGS for Real Time Traffic	Bandwidth: Network guarantees the bandwidth is within (CIR, PIR) Latency: Minimized. Network guarantees E2E bounded latency Jitter: 1/2 of E2E bounded latency Packet Loss: Minimized <ul style="list-style-type: none"> Congestion-free Lossless queuing Only drop when physical failure 	Asynchronous communication: Critical sensor and control data
BGS for bandwidth sensitive traffic	Bandwidth: Network guarantees the bandwidth is within (CIR, PIR) Latency: Less important Jitter: Less important Packet Loss: Don't care	Un-critical data
BES for other type of traffic	Don't care	Other data

C. Queuing and Scheduling Algorithm

The paper proposes two types of algorithms illustrated in Figures 7 and 8. One is for asynchronous environment that there is no clock sync for network. Another is synchronous environment that clock is synced with certain accuracy for IVN including all devices. Below are details, also, the experiment section is based on the two algorithms discussed here.



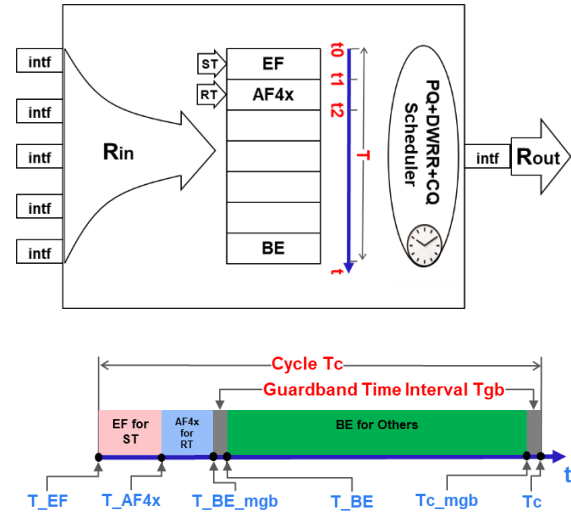
Algorithm 1: Asynchronous Class Based Scheduler

```

Packet *pkt scheduler () {
//Scheduler function. The EFQ has the highest priority, AF4xQ and other Q
have lower priority and are served by DWRR
1. while EFQ.length() > 0 do //serve the EF queue
2.   EFQ.dequeue(pkt)
3.   return(pkt)
4. while AF4xQ.length() > 0 do //serve the AF4x queue
5.   W_AF4x -> W_AF4x' //update weight W for AF4x
6.   if W_AF4x' < W_AF4x then //updated W < assigned W for AF4x
7.     AF4xQ.dequeue(pkt)
8.     return(pkt)
9.   else
10.    continue
11. while BEQ.length() > 0 do //serve the BE queue
12.   BEQ.dequeue(pkt)
13.   return(pkt)
14.}

```

Figure 7. 1st Algorithm and psudo code: Asynchronous Solution.



Timer:

T_{EF}: The time when EF class is started to be served

T_{AF4x}: The time when AF4x class is started to be served

T_{BE}: The time when BE class is started to be served

T_c: Cycle time interval

T_{gb}: Time interval for Guard-band

T_{BE_mgb} = T_{BE} - T_{gb}

T_{c_mgb} = T_c - T_{gb}

Algorithm 2: Synchronous Class Based Scheduler

```

Void timerProcess (TIMER ExpiredTimer) {
//Timer process function, process events when a timer expired. When a timer
expired, the associated gate is open, then the scheduler can schedule the traffic
for the class. Example only shows three classes.
1. if ExpiredTimer == T_EF then //Timer for T_EF is expired
2.   openGate = EF //Open the gate for EF
3.   if isTimerRunning() != true then
4.     startTimer(T_AF4x) //Start next timer for T_AF4x
5. else if ExpiredTimer == T_AF4x then //Timer for T_AF4x is expired
6.   openGate = AF4x //Open the gate for AF4x
7.   if isTimerRunning() != true then
8.     startTimer(T_BE_mgb) //Start next timer for T_BE_mgb

```

```

9.  else if ExpiredTimer == T_BE_mgb then
10.                                     //Timer for T_BE_mgb is expired
11.    openGate == NONE                                     //Close the gate for all
12.    isTimerRunning() != true then
13.        startTimer(T_BE) /                                     //Start next timer for T_BE
14.    else if ExpiredTimer == T_BE then //Timer for T_BE is expired
15.        openGate = BE                                     //Open the gate for BE
16.        isTimerRunning() != true then
17.            startTimer(Tc_mgb) //Start next timer for Tc_mgb
18.    else if ExpiredTimer == Tc_mgb then
19.                                     //Timer for Tc_mgb is expired
20.        openGate = NONE                                     //Close the gate for all
21.        isTimerRunning() != true then
22.            Increment all timer by Tc //Increase all timer by one Tc
23.            startTimer(T_EF) //Start next timer for T_EF
24.}
25.
26. Packet *pkt scheduler () { //Scheduler function
27.     while EFQ.length() > 0 and
28.         openGate == EF do //serve the EF queue
29.         EFQ.dequeue(pkt)
30.         return(pkt)
31.     while AF4xQ.length() > 0 and
32.         openGate == AF4x do //serve the AF4x queue
33.         W_AF4x >= W_AF4x' //update weight W for AF4x
34.         if W_AF4x' < W_AF4x then //updated W < assigned W for AF4x
35.             AF4xQ.dequeue(pkt)
36.             return(pkt)
37.         else
38.             continue
39.     .... //serve other queues
40.     while BEQ.length() > 0 and
41.         openGate == BE do
42.         BEQ.dequeue(pkt) //serve the BE queue
43.         return(pkt)
44.}

```

Figure 8. 2nd Algorithm and psudo code: Synchronous Solution.

- For asynchronous environment, Priority Queuing (PQ) combined with Deficit Weighted Round Robin (DWRR) or any type of Weighted Fair Queuing (WFQ) are used. It is called the 1st Algorithm in the document thereafter. Normally, the time sensitive flows, i.e., scheduled traffic (EF class) and real-time traffic (AF4x class) are put into the 1st and 2nd priority of the queue, and other classes of traffic, BGS and Best Effort class of traffic, are put into the lower priority queues. For admission control and scheduler configuration, the total CIR for LGS class, and the WEIGHT values of BGS class can be calculated from the sum of CIR of all flows in the same class. This algorithm has already deeply analyzed in [7].
- For synchronous environment, above asynchronous PQ+DWRR algorithm is combined with Cyclic Queuing (CQ). It is called the 2nd Algorithm in the document thereafter. Each class of traffic has a dedicated time window to be served by the scheduler. The service time is associated with the sum of CIR of all flows in the same service. The Scheduler will calculate and adjust the serving time window for each class when a flow's state is changed, such as new flow is added, or old flow is removed. The guard-band is added for lower priority classes to guarantee the EF class traffic, when served, is not blocked by lower priority traffic on wire. In another word, when EF class is served, the wire is always available

for transmission. The guard-band timer interval can be calculated as the required time to transmit one maximum size of packet on wire speed.

D. Traffic Shaping

Traffic shaping is used to absorb the overflow and burst of the traffic in the class and its objectives are: (1) the packet in the class is never built up, thus reducing the latency (2) traffic in lower priority class is never starved by higher priority traffic. Existing Single Rate Three Color Marker [30] or Two Rate Three Color Marker [31] could be used for traffic shaping. Other type shaping like leaky bucket shaping can also be used. Traffic shaping deployment is very flexible. It can be configured in both ingress and egress interface. It can be per flow based, or per class based.

Flow-level traffic shaping in ingress interface can also be used as the policy enforcement module, it will check the user's traffic to see if it is allowed to pass or trigger some policy, such as discard or put into lower priority to process.

VIII. LATENCY ANALYSIS AND ESTIMATION

To provide the Latency Guaranteed Service (LGS) for ST and RT, the network must be able to estimate the latency for a network path and offer to user in the provisioning stage. This is the requirement for SLA negotiation. This section will analyze all factors that can result in network latency and discuss some basic formulas.

A. The Latency Analysis for IP Network

In this paper, the latency estimation is for E2E from the perspective of user's application. The latency must include all delay occurred in network and hosts. This is illustrated in Figure 9. The formula for the latency is as in (1) and (2). The superscript "LGS" denotes LGS packet.

$$D_{e2e}^{LGS} = PD + \sum_{i=1}^n (OD_i^{LGS} + QD_i^{LGS}) + \sum_{s=1}^m SD_s^{LGS} = t_1 - t_0 \quad (1)$$

$$SD_s^{LGS} = L^{LGS} * 8/R_{out} \quad (2)$$

- t_0 : the time the 1st bit of a pack is leaving the application process on the source host.
- t_1 : the time the 1st bit of the pack is received by the application process on the destination host.
- PD : Propagation delay, this delay is limited by the speed of signaling in a physical media. For example, it is approximately 200k KM/s in optical fiber.
- OD_i : The other delays (pack process, deque, de-capsulation, lookup, switch, L2-rewrite, encapsulation, etc.) at the i -th hop and source host. This delay is related to the Forwarding Chip and hardware, it is normally and relatively steady for a specified router or switch and can be easily measured. This delay is insignificant compared with QD and SD described below.

- QDi : The queuing delay at the i -th hop and source host.
- SDs : The serialization delay at the s -th link segment, it can be calculated by the formula (2). L^{LGS} is the packet length (byte) for the LGS flow. R_{out} is the link speed.

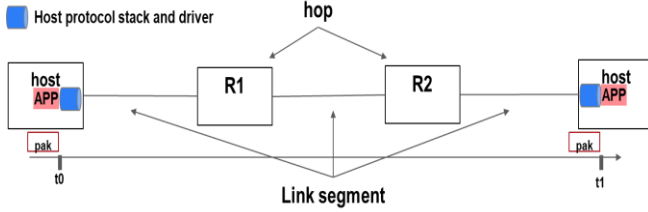


Figure 9. The End-to-End Latency for IP Applications.

B. Estimation for the Queuing Latency (QD)

The formulas for the queueing latency estimation (for the same packet size) have been derived in [7] for the 1st Algorithm. In this paper, different packet size for two class is used, thus formulas are different as in [7]. The maximum number of packet and queuing time for a queue (EF or AF4x) under the worst scenario for a hop are shown in equations from (3) to (8).

$$N_{max}^{EF} = \lceil R_{in}^{EF} / R_{out} * (L_{max}^{LOW} / L_{max}^{EF} + 1) + 1 \rceil \quad (3)$$

$$D_{max}^{EF} = N_{max}^{EF} * L^{EF} * 8 / R_{out} \quad (4)$$

$$N_{max}^{AF4x} = \lceil R_{in}^{EF} / R_{out} * (L_{max}^{LOW} / L_{max}^{EF} + 1) + 1 \rceil + \lceil (R_{in}^{AF4x} / R_{out} * (L_{max}^{LOW} / L_{max}^{AF4x} + 1) + 1) * (R_{in}^{AF4x} / R_{out}) \rceil \quad (5)$$

$$D_{max}^{AF4x} = N_{max}^{AF4x} * L^{AF4x} * 8 / R_{out} \quad (6)$$

$$R_{in}^{EF} = r_{EF} \sum_{i=1}^m cir_i^{EF} \quad (7)$$

$$R_{in}^{AF4x} = r_{AF4x} \sum_{i=1}^n cir_i^{AF4x} \quad (8)$$

For the 2nd Algorithm, the packet in any queue is served on a pre-allocated time window, and this will guarantee that flows will not be interfered by any packets in other queues. So, it is easy to estimate that the maximum number of packets in a queue is as in (9), (10). The associated queuing time is the same as in (4) and (6). However, for the worst scenario when a packet is out of the allocated window for some reason, the maximum latency will be as the (11).

$$N_{max}^{EF} = \lceil R_{in}^{EF} / R_{out} + 1 \rceil \quad (9)$$

$$N_{max}^{AF4x} = \lceil R_{in}^{AF4x} / R_{out} + 1 \rceil \quad (10)$$

$$D_{max}^{EF} = D_{max}^{AF4x} = T \quad (11)$$

The symbols and parameters in the formulas above are described as below,

- The symbol “ $\lceil \quad \rceil$ ” is the rounding up operator.
- N_{max}^{EF} : the maximum queue depth for EF queue.

- N_{max}^{AF4x} : the maximum queue depth for AF4x queue.
- D_{max}^{EF} : the maximum queueing time for EF queue.
- D_{max}^{AF4x} : the maximum queueing time for AF4x queue.
- R_{in}^{EF} : the ingress rate for EF queue.
- R_{in}^{AF4x} : the ingress rate for AF4x queue.
- cir_i^{EF} : the Committed Information Rate (cir) for the i -th flow for EF queue.
- cir_i^{AF4x} : the Committed Information Rate (cir) for the i -th flow for AF4x queue.
- r_{EF} : the burst coefficient for the traffic of EF queue.
- r_{AF4x} : the burst coefficient for the traffic of AF4x queue.
- T : the cycle time for the scheduler when CQ is used.

IX. NETWORK MODELING AND EXPERIMENTS

To verify and analyze the New IP based IVN architecture can meet the requirements of IVN, OMNeT++ [32] is used to simulate the network, the detailed bandwidth, E2E latency, pack loss, etc., can be retrieved from tests. OMNeT++ is very popular to simulate time driven events and activities involved in networking technologies, it can accurately calculate and simulate the life of each individual packet traveling from source to destination via different intermediate devices. So, its results in QoS metrics are very close to the theoretical estimations.

A. Network Topology

The network is illustrated in Figure 10. It is a ring topology but with the cut of another ring link to focus on the latency simulation under the worst scenario (longer latency). All links speed is 100 Mbps. The network consists of ECU, computers, and routers. ECU is to simulate the sensors with control connected on Ethernet Bus. It has a full TCP/IP stack and is responsible for the ST and RT generation and process. The ST and RT are simulated by UDP packets. Computers are simulating the generation and process of Best-Effort traffics (TCP and UDP) that are used to interfere ST and RT between ECUs.

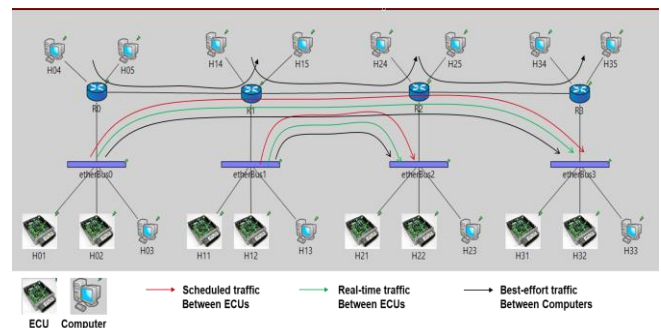


Figure 10. Network Topology and traffic.

The purpose of simulation is to illustrate the new architecture can provide the E2E guaranteed service for ST and RT flows when the network is severely congested and interfered by the Best-Effort traffic. The E2E guaranteed service includes three criteria: (1) bounded latency (2) bounded jitter (3) congestion free and lossless. Moreover, the tested latency and jitter for ST and RT should be close to the estimated latency described in section VIII.

B. Network Devices

Each router consists of Ingress Modules, Switch Fabric and Egress Modules that are illustrated in Figure 11. The Ingress Modules simulate the traffic classification and ingress traffic shaping functions; The Egress Modules simulate the egress traffic shaping, queuing, and scheduling functions. The Switch Fabric Modules simulate the IP lookup, switching and L2 re-writing functions. Two types of schedulers are used. Only class level traffic shaping is used for ST for ingress and egress.

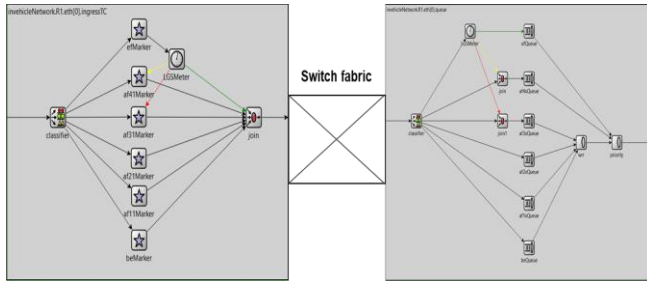


Figure 11. Router structure.

C. Traffic Configuration

To simulate the worst scenario, very heavy traffic for the IVN simulation is configured as below:

- There is total 100 ST flows and 100 RT flows using UDP, each flow has the packet size 254 bytes (200 bytes data, 54 bytes of UDP and Ethernet header), the send interval is 10ms. So, each flow has a rate of 203.2 Kbps. Both rate for ST flows and RT flows are 20.32Mbps, it means the remained bandwidth for BGS, and BE is about 60Mbps.
- 50 ST flows and 50 RT flows are from ECU H01 and H02 to H31 and H32, these flows' results are checked and compared with the estimation. 50 ST flows and 50 RT flows are from ECU H11 and H12 to H21, H22.
- There is total 250 interference flows configured between other computers. The interference flows will cause all links between routers congested, R1 link Eth[0] is the most severely congested router and link. All flows packet size are 200 bytes or 1500 bytes. Both TCP and UDP are configured for interference flows.

D. Cyclic Queueing and Scheduler Configuration

For the 2nd algorithm, the detail of the cyclic queuing is configured as in Figure 12.

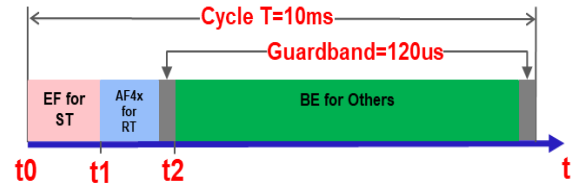


Figure 12. The Cyclic Queueing Configuration.

- o The cycle T for all router and hosts are 10ms.
- o A guard-band of 1500 bytes or 120 us are configured for both AF4x and BE classes. 120 us is the time to transmit 1500 bytes packet on 100M bps link.
- o The time window size for EF and AF41 are 22% and 32% of the cycle T respectively.

E. Experiment Results and Analysis for E2E Latency/Jitter

This sub-section will analyze the E2E latency/jitter for different type of traffic, compare the experiment results with the theoretical estimation made in Section VIII.

TABLE III shows the detailed calculation for the E2E latency estimation. First, estimate the maximum number of packets queued in each egress link of all routers on the path, then calculate the maximum queuing delay. The minimum E2E latency means there is no queuing latency in each hop, so it is determined by the sum of all link segment's serialization latency on the path. Each 100M link will have 20.3 us serialization latency for 254 bytes ST or RT traffic. The burst coefficient for each case is also shown in Table III. Higher coefficients for router R0 and R1 are selected since there are aggregation of the traffic for the routers. For other routers, the coefficient is selected as 1, or no burst effect.

TABLE III. THE E2E DELAY ESTIMATION OF ST AND RT FLOWS

Algorithm	Class and traffic	Estimated max number of packet in Egress Q					Estimated Total Queuing Latency (us)	Calculated Total Serialization Delay (each hop has 20 us)	Estimated Total E2E Delay (us)
		Host	R0	R1	R2	R3			
PQ+DWRR	EF for ST	0	3 ($r_{EF}=2$)	6 ($r_{EF}=4$)	3 ($r_{EF}=1$)	3 ($r_{EF}=1$)	305	100	405
	AF4x for RT	0	4 ($r_{AF4x}=2$)	6 ($r_{AF4x}=4$)	4 ($r_{AF4x}=1$)	4 ($r_{AF4x}=1$)	365	100	465
PQ+DWRR+CC	EF for ST	0	2 ($r_{EF}=1$)	2 ($r_{EF}=1$)	2 ($r_{EF}=1$)	2 ($r_{EF}=1$)	162	100	262
	AF4x for RT	0	2 ($r_{AF4x}=1$)	2 ($r_{AF4x}=1$)	2 ($r_{AF4x}=1$)	2 ($r_{AF4x}=1$)	162	100	262

TABLE IV shows the Min/Max E2E Delay for the worst performed flow, and estimation values also compared. The worst performed flow is defined as that the flow's Max E2E delay is the biggest in all flows in the same class.

Jitter is not shown in the table, but it can be easily calculated by the variation of mean and Min/Max value, the mean value can be simply calculated by the average of Min/Max values.

TABLE IV. THE COMPARISON OF EXPERIMENT RESULT AND ESTIMATION

Algorithm	Min/Max E2E Delay (us) for the worst performed flow carrying ST between H01/H02 to H31/H32		Min/Max E2E Delay (us) for the worst performed flow carrying RT between H01/H02 to H31/H32	
	Experiment	Estimation	Experiment	Estimation
PQ+DWRR	108/391	100/405	278/542	100/465
PQ+DWRR+CQ	109/152	100/262	169/169	100/262

Figures 13-16 illustrate the E2E delay changes with time for the worst performed flows shown in TABLE IV.

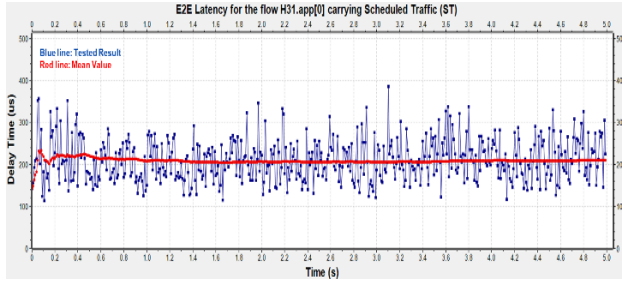


Figure 13. The 1st Algo: The E2E Latency (min=108us, max=391us) for the worst performed ST flow.

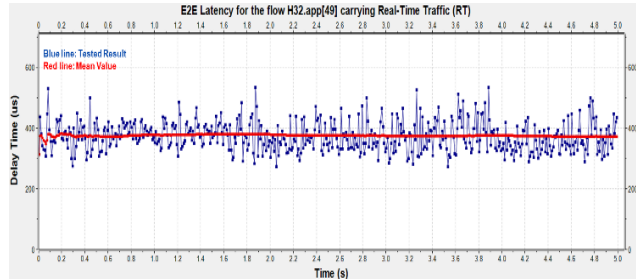


Figure 14. The 1st Algo: The E2E Latency (min=278us, max=542us) for the worst performed RT flow.

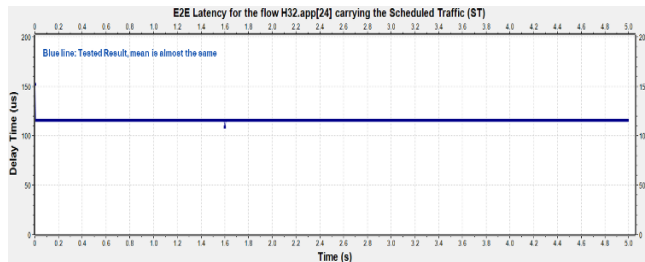


Figure 15. The 2nd Algo: The E2E Latency (min=109us, max=152us) for the worst performed ST flow.

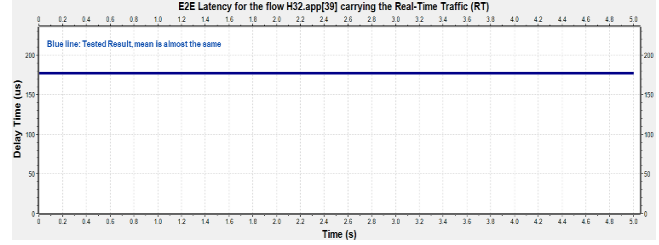


Figure 16. The 2nd Algo: The E2E Latency (min=169us, max=169us) for the worst performed RT flow.

F. The Receiver's Instantaneous Bandwidth and Packet Loss Verification

This sub-section will verify there is no bandwidth loss for every flow. "No bandwidth loss" is verified by checking if receiver's instantaneous rate or bandwidth is similar to the sender's rate for every flow.

The receiver's Instantaneous Bandwidth (B) is calculated for each received packet at receiver side by the formulars (12) to (13), there are three scenarios :

- When there is only one packet received:

$$B = 0 \quad (12)$$

- When there are two packets received with different size in byte. At t_0 , received a packet and its size is L_{t_0} . At t_1 , received a packet and its size is L_{t_1} :

$$B = 0.5 * (L_{t_0} + L_{t_1}) * 8 / (t_1 - t_0) \quad (13)$$

- When there are more than two packets received with different size in byte. Three packets are sampled for calculation: At t_0 , received a packet and its size is L_{t_0} . At t_1 , received a packet and its size is L_{t_1} . At t_2 , received a packet and its size is L_{t_2} :

$$B = (0.5 * L_{t_0} + L_{t_1} + 0.5 * L_{t_2}) * 8 / (t_2 - t_0) \quad (14)$$

For the test for Algorithm 1, five ST flow's sending rate are set differently at source, two have constant rate and three have variable rate.

For the test for Algorithm 2, five ST flow's sending rate are constant. It is hard to set the rate to be variable for algorithm 2 since if a packet is not sending at its allocated time window, there will be extra delay of time cycle. This will impact the analysis for the instantaneous bandwidth.

The paper only demonstrates the bandwidth for ST flows. The results for RT flows are similar.

Figures 17 to 20 illustrate the instantaneous rate or bandwidth for the five ST flows for two algorithms respectively. It is obvious that each flow for two algorithms has almost same wave shape. It indicates that the receiver's instantaneous rate is almost the same as the sender's rate, so there is no bandwidth loss for the network.

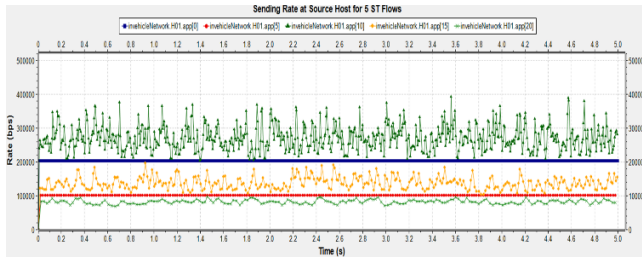


Figure 17. The 1st Algo: The Sender's Instantaneous Bandwidth for 5 ST flow.

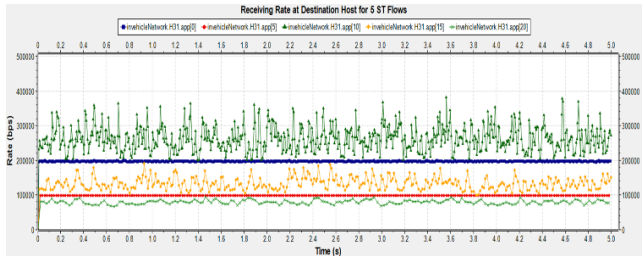


Figure 18. The 1st Algo: The Receiver's Instantaneous Bandwidth for 5 ST flows.

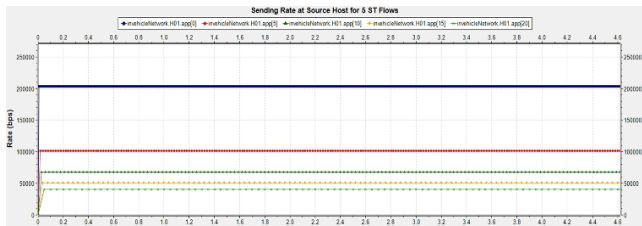


Figure 19. The 2nd Algo: The Receiver's Instantaneous Bandwidth for the worst performed ST flow.

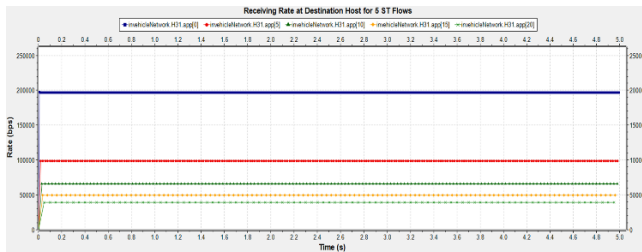


Figure 20. The 2nd Algo: The Receiver's Instantaneous Bandwidth for the worst performed RT flow.

This sub-section will also verify there is no packet drop from queuing and congestion. To demonstrate the lossless and congestion-free for ST and RT flows, Figure 21 shows the statistics of all queues in R1 for two algorithms. No packet dropped in EF and AF4x queues while there are packets dropped in BE queue. This is as expected, congestion should only happen for BE traffic, ST and RT flows are not impacted and are lossless and congestion-free. R1 is the most severely congested, other Router's queues also have similar pattern. No packet drops for EF and AF4x.

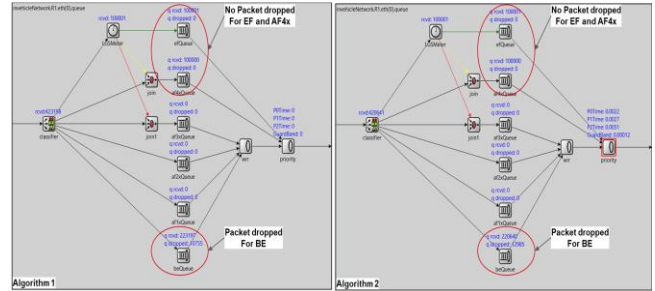


Figure 21. The statistics for all Queues for two algorithms

Here is a summary from the test results:

- The queuing latency of higher priority queues by PQ is very short and is not impacted by the congestion of lower priority class of traffic. E2E Maximum latency estimation in Section VIII can be used as the rough prediction for almost all traffic's real maximum E2E latency.
- Lossless and congestion free can be achieved for ST and RT flows if the admission control is done for the flows. When the total rate for ST and RT flows are below the CIR of service expectation has claimed, there will be no packet drop caused by queue overflow.
- The E2E latency shown in the experiment does not include "Other Delay" and "Propagation Delay" described in Section VIII. "Propagation Delay" is very trivial in IVN, but "Other Delay" should be considered and added up if they are significant compared with the final queueing latency. For most of forwarding chip, "Other Delay" is very small and below hundred microseconds, but for x86 based virtual router, it might not be true depending on the forwarding software design.
- The latency per hop is inversely proportional to the link speed. For example, the experiment using 100M link with 4 hops network can achieve hundreds microsecond for E2E latency. It is expected that the corresponding latency for the same network is about tens of microsecond and couple microseconds for 1G and 10G link, respectively. Higher link rate will not only reduce latency, but also provide more bandwidth for non-time-sensitive applications. So, the paper proposes to use at least 1G link for the IVN in the future.

X. CONCLUSIONS

The paper has proposed a new architecture for future V2X communication, that is based on the integration of New IP and 5G Technologies. Unlike the 5G uRLLC that is only limited in wireless network for its end-to-end definition, The new V2X architecture can provide a real end-to-end guaranteed service for bandwidth, latency, jitter and packet loss. The "real end-to-end" will cover all segments of network including user end device (UE) associated with IP source, wireless access, wireless core network, data network and to another user end device or computer in Internet associated with IP destination.

The paper also analyzed the detailed requirements for the In-Vehicle-Network in terms of QoS characters. The paper proposed to use New IP for future IVN. Class based queueing

and scheduling plus traffic shaping can provide per-hop LGS and BGS. Combined with Central Controller or In-band Signaling, the E2E guaranteed service for new IVN can be achieved by enforcing the per-hop guaranteed service on all network devices on the IP forwarding path. The solution is backward compatible as the existing IP traffic and traditional best effort service can coexist with the new classes of traffic and new services.

To prove the concept, the paper also discussed in detail about the experiments of network modelling on New IP based IVN. The simulation has demonstrated that the New IP can satisfy very stringent QoS requirements for IVN. The results indicate the future IVN can obsolete diversified legacy protocols and unify to one protocol: New IP. This will dramatically reduce the cost of IVN.

The paper investigated two algorithms for scheduling, asynchronous and synchronous solutions. If the accurate clock can be provided, the synchronous solution by using CQ could improve the latency and jitter significantly. But it must be noted that costs of synchronous solution are not trivial, following tasks are mandatory:

- The crucial requirement of using CQ is the clock sync in the IVN, this is a different topic, and the paper does not address it. Basically, a central controller or distributed protocol, such as IEEE1588 can be used to sync all device clock with a certain accuracy.
- Cycle value selection. The cycle value and the clock accuracy requirement depend on each other, both will determine the granularity of the served packet size, the link utilization, the maximum latency, and the cost of the scheduler design.
- Time window allocation for different flows with different constraints in bandwidth and latency. The optimized solution needs complicated math and cause an overhead for the solution.

As a conclusion, the New IP based IVN can satisfy very well the requirements for the communications of different applications. It opens the door for future IVN and V2X.

Further research is still needed in the following areas:

- Burst effect analysis: The burst coefficient value (Section VIII) will directly impact the accuracy of queuing latency estimation at each hop and will finally determine the accuracy of E2E latency estimation. More study is needed for the burst analysis. A better and more accurate quantitative estimation to the queueing behavior by burst traffic is expected.
- TCP congestion control: The congestion control for different service is expected to be different. New algorithms are critical for application to effectively utilize the new guaranteed service provided by network.
- Algorithm for network resource planning and allocation for synchronous solution, such as optimized cycle number, fast and efficient time slot allocation, scheduler management, etc.

- Simpler method than preemption is needed to eliminate the extra latency and jitter for higher priority traffic caused by a lower priority packet on hardware that is in transmission. This unfinished packet is the root cause of jitter for high priority traffic. Preemption is hard to realize in hardware. Without preemption, the only way to eliminate such effect is to use CQ, but CQ has to sacrifice the link utilization.

REFERENCES

- [1] L. Han, L. Dong, R. Li, "A Study of In-Vehicle-Network by New IP", INTERNET 2021, The Thirteenth International Conference on Evolving Internet, ISBN: 978-1-61208-880-8
- [2] "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments". IEEE 802.11p published standard. IEEE. July 15, 2010.
- [3] 3GPP, "Cellular Vehicle-to-Everything (V2X)," https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_73/Info_for_workplan/revised_WID_22/RAN1_6/RP-161894.zip
- [4] S. Jiang, S. Yan, L. Geng, C. Cao, and H. Xu, "New IP, Shaping Future Network: Propose to initiate the discussion of strategy transformation for ITU-T", TSAG C-83
- [5] R. Li, A. Clemm, U. Chunduri, L. Dong, and K. Makhijani, "A New Framework and Protocol for Future Networking Applications," ACM Sigcomm NEAT workshop, 2018, pp 21–26.
- [6] L. Han, Y. Qu, L. Dong and R. Li, "Flow-level QoS assurance via IPv6 in-band signalling," 2018 27th Wireless and Optical Communication Conference (WOCC), 2018, pp. 1-5, doi: 10.1109/WOCC.2018.8372726..
- [7] L. Han, Y. Qu, L. Dong and R. Li, "A Framework for Bandwidth and Latency Guaranteed Service in New IP Network," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2020, pp. 85-90, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162747.
- [8] "3GPP specification series: 38series". 3GPP. Retrieved 2018-10-31.
- [9] "Minimum requirements related to technical performance for IMT-2020 radio interface(s)", Report ITU-R M.2410-0.
- [10] "System Architecture for the 5G System", 3GPP TS 23.501 version 15.2.0 Release 15
- [11] R. Braden, L. Zhang., S. Berson, S. Herzog, and S. Jamin, "RFC 2205: Resource ReSerVation Protocol (RSVP)-Version 1 Functional Specification", IETF, Sept. 1997.
- [12] "Stream Reservation Protocol (SRP)", IEEE 802.1Qat
- [13] "IEEE 802.1 Time-Sensitive Networking Task Group".
- [14] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "RFC 2475: An Architecture for Differentiated Services," IETF, Dec. 1998.
- [15] R. Braden, D. Clark, and S. Shenker, "RFC 1663: Integrated Services in the Internet Architecture: an Overview," IETF, Jun. 1994.
- [16] "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Management and orchestration; 5G end to end Key Performance Indicators (KPI)", 3GPP TS 28.554
- [17] P. Yang, J. Shao, W. Luo, L. Xu, J. Deogun and Y. Lu, "TCP Congestion Avoidance Algorithm Identification," in

- IEEE/ACM Transactions on Networking, vol. 22, no. 4, pp. 1311-1324, Aug. 2014, doi: 10.1109/TNET.2013.2278271.
- [18] B. Briscoe, K. Schepper, M. Bagnulo, and G. White, "Low Latency, Low Loss, Scalable Throughput (L4S) Internet Service: Architecture", Work in Progress, Internet-Draft, draft-ietf-tsvwg-l4s-arch-08, 15 November 2020, <<http://www.ietf.org/internet-drafts/draft-ietf-tsvwg-l4s-arch-08.txt>>.
- [19] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, "RFC3209: RSVP-TE: Extensions to RSVP for LSP Tunnels", IETF, Sept. 2001
- [20] LIN, "ISO/AWI 17987-8"
- [21] CAN: "Road vehicles - Controller area network (CAN) - Part 1: Data link layer and physical signalling", ISO 11898-1:2003
- [22] FlexRay: ISO 17458-1 to 17458-5
- [23] AUTOSAR: AUTomotive Open System ARchitecture
- [24] H. Lim, L. Völker, and D. Herrscher, "Challenges in a future IP/Ethernet-based in-car network for real-time applications", 48th ACM/EDAC/IEEE Design Automation Conference (DAC), 2011
- [25] R. Steffen, R. Bogenberger, J. Hillebrand, W. Hintermaier, A. Winckler, and M. Rahmani, "Design and Realization of an IP-based In-car Network Architecture", Proceedings of "1st International ICST Symposium on Vehicular Computing Systems", 2008
- [26] "P802.1DG – TSN Profile for Automotive In-Vehicle Ethernet Communications". 1.ieee802.org.
- [27] E. Ackerman, "Upgrade to Superhuman Reflexes Without Feeling Like a Robot" <<https://spectrum.ieee.org/enabling-superhuman-reflexes-without-feeling-like-a-robot#toggle-gdpr>>
- [28] S. Tuohy, M. Glavin, C. Hughes, E. Jones, M. Trivedi and L. Kilmartin, "Intra-Vehicle Networks: A Review," in IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 534-545, April 2015, doi: 10.1109/TITS.2014.2320605.
- [29] "Supporting internet protocol version 4 (IPv4) extension headers", United States Patent, 10,742,775.
- [30] J. Heinanen and R. Guerin, "RFC 2697: A Single Rate Three Color Marker", IETF, Sept. 1999.
- [31] O. Aboul-Magd and S. Rabie, "RFC 4115: A Differentiated Service Two-Rate, Three-Color Marker with Efficient Handling of in-profile Traffic", IETF, Jul. 2005.
- [32] "OMNeT++ Discrete Event Simulator"

A Topic Modeling Framework to Identify Online Social Media Deviance Patterns

Thomas Marcoux, Esther Mead, Nitin Agarwal

COSMOS Research Center

University of Arkansas at Little Rock

Little Rock, AR, USA

email: {txmarcoux, elmead, nxagarwal}@ualr.edu

Abstract—Following the COVID-19 pandemic and the subsequent vaccine related news, the information community has seen the emergence of unique misinformation narratives in a wide array of different online outlets, through social media, blogs, videos, etc. Taking inspiration from previous COVID-19 and misinformation detection related works, we expanded our topic modeling tool. We added filtering capabilities to the tool to adapt to more chaotic social media datasets and create a chronological representation of online text content. We curated a corpus of 543 misinformation pieces whittled down to 243 unique misinformation narratives, and collected two separate sets of 652,120 and 1,664,123 YouTube comments. From our corpus of misinformation stories, this tool has shown to accurately represent the ground truth of COVID misinformation stories. This highlights some of the misinformation narratives unique to the COVID-19 pandemic and provides a quick method to monitor and assess misinformation diffusion, enabling policy makers to identify themes to focus on for communication campaigns. To expand previous publications and further explore the potential of topic streams in understanding online misinformation, we propose a framework used as a filter to help whittle down big data corpora and identify latent misinformation within. This could be scaled and applied to very large social networks to highlight misinformation.

Keywords—misinformation; disinformation; topic models; topic streams; COVID-19; misinfodemic; narratives.

I. INTRODUCTION

Social media is characterized as a powerful online interaction and information exchange medium. However, it has given rise to new forms of deviant behaviors, such as spreading fake news, misinformation, and disinformation. For this reason, we began this research in our previous publication [1] and are now introducing this extended version. Due to afforded anonymity and perceived diminished personal risk of connecting and acting online, deviant groups are becoming increasingly common. Online deviant groups have grown in parallel with Online Social Networks (OSNs), whether it is black hat hackers using Twitter to recruit and arm attackers, announce operational details, coordinate cyber-attacks [2], and post instructional or recruitment videos on YouTube targeting certain demographics; or state/non-state actors and extremist groups (such as the Islamic State of Iraq and Syria) savvy use of social communication platforms to conduct phishing operations, such as viral retweeting of messages containing harmful URLs leading to malware [3].

More recently, there is a surge in misinformation and scam cases pertaining to COVID-19. The problem of misinformation is actually worse than the pandemic itself. That is why it is called infodemic or more specifically, misinfodemic. Like the pandemic, misinformation cases are also rising exponentially. These cases are more difficult to track than the epidemic, as they can originate in the dark corners of the Internet. To make matters worse, we cannot enforce lockdown on the Internet to stop the spread of this infodemic. This is in part because, during crises, the Internet is usually the first mode of communication and source of information. Although there are some quarantine efforts, for instance from social media companies, such as Facebook, YouTube, and retail companies like Amazon are doing their best to block such content, by suspending bad actors or scammers who are spreading misinformation to further their political agenda or to try to profit off of this adversity. But such cases are simply too many and growing too fast. What makes this problem worse is the fact that the information spreads like a wildfire on the Internet, especially the false or misinformation. Many studies have concluded that misinformation travels faster than its corrective information, and the more questionable the misinformation is the faster it travels. This is simply because on social media people usually have a lot more virtual friends than they do in their real life. So, if they share or retweet some misinformation, wittingly or unwittingly, they expose all their virtual friends to the misinformation.

There are similarities between misinformation about COVID-19 and other misinformation cases that we have studied for NATO, US, EU, Singapore, and Canada, etc. Like in other cases, the motivation for spreading COVID-19 misinformation is monetization or to provoke hysteria. Bad actors or scammers are spreading misinformation to further their political agenda or simply trying to profit off of this adversity. For instance, there exists many cases of scammers selling fake masks, fake cures, using fake websites to ask for private/sensitive information from people by posing as government websites. However, there is a significant difference between COVID-19 and other misinformation campaigns that we have studied before. Being a global and rapidly evolving crisis, the nature of misinformation is also extremely diverse and super-fast. Other misinformation campaigns were specific to an entity, event, region, elections, military exercises.

However, misinformation about COVID-19 has both global as well as regional narratives. While fake masks, fake cures, etc., affect a global audience, the regional narratives include promoting medicines for bovine coronavirus as cure for human coronavirus affecting rural/agriculturalist regions. Moreover, the misinformation about COVID-19 ranges from health to policy to religion to geopolitical affairs, i.e., highly topically diverse. Given the volume, velocity, and variety of COVID-19 related misinformation, research is warranted to study such campaigns and their organization. As resources are stretched too thin, government and other regulatory bodies cannot afford to investigate all the misinformation campaigns and scams. Such research could help prioritize investigation of misinformation campaigns and scams.

Therefore, we propose a study of the themes and chronological dynamics of the spreading of misinformation about COVID-19. Our scope focuses on misinformation geographically relevant to us (Arkansas, USA), as well as some global stories, with our main corpus is a collection of unique misinformation stories manually curated by our team. In collaboration with the Arkansas Attorney General, we have shared our findings with their office and made all reports and misinformation stories publicly available online [4]. In addition, we have collected a variety of YouTube video titles and comments. This allows us to compare a curated corpus to a data set more chaotic and true to life. To highlight and visualize these misinformation themes, we use topic modeling, and introduce a tool to visualize the evolution of these themes chronologically.

In addition, to expand our previous work [1], we introduce a manual node-based design to filter very large datasets and identify information of interest within, while avoiding the bias that can come with artificial intelligence methods. This framework is tested with a set of 1,664,123 YouTube comments and is built to introduce further feature detection, such as commenting behavior, or even inorganic video engagement behavior, tackling the issue of multimedia misinformation.

The rest of this study is structured as follows. First, we will discuss the work done by other researchers in comparable research in Section 2, describe our methodology in Section 3, including data collection, processing, and topic modeling methodology. Then, in Section 4, we will discuss our results and the subjective findings of our misinformation team with the scientific topic streams visualizations that support them. Finally, we briefly introduce our free online resource where the misinformation stories used here can be found, before presenting our conclusions in Section 5.

II. LITERATURE REVIEW

In this section, we first argue of the importance of this field as it can directly relate to public safety, followed by the efforts of the research community to combat this issue. We then introduce the significance of the YouTube platform and argue our choice of using YouTube comments for this study, finishing this section with the relevant literature on our primary analysis technique: topic models.

A. The Significance of Misinformation

The information community has been tackling the issue of misinformation surrounding the COVID-19 pandemic since early in the outbreak. We base the claims found in this paper on the findings that misinformation spreads in a viral fashion and that consumers of misinformation tend to fail at recognizing it as such [5]. In addition to this, we believe this research is essential as rampant misinformation constitutes a danger to public safety [6]. We also believe this research is helpful in curbing misinformation since researchers have found that simply recognizing the existence of misinformation and improving our understanding of it can enhance the larger public's ability to recognize misinformation as such [5]. In order to better understand the misinformation surrounding the pandemic, we look at previous research that has leveraged topic models to understand online discussions surrounding this crisis. Research has shown the benefits of using this technique to understand fluctuating Twitter narratives [7] over time, and also in understanding the significance of media outlets in health communications [8]. Studies on information propagation [9] establish entire mathematical models around the diffusion of misinformation and emphasize that early detection is essential to allow a proper response.

B. Misinformation Detection

Because of the severity of the threat of misinformation campaigns and the need to quickly discover such efforts, we concern ourselves with detection models to help systematically recognize inorganic or concerted information operations. Because misinformation spreads so quickly and deals long lasting damage, we consider developing scalable models to quickly identifying misinformation a critically important endeavor. Of course, because of the severity of this public issue, there are a great many efforts within the information community striving to propose solutions. The state of the art in fake news detection could be roughly described as being divided between three main ideas. One is artificial intelligence models, where researchers will use traditional machine learning techniques [10, 11], multinomial Bayesian models [12, 13], or deep-learning [14, 15, 16]. Another school of thought in misinformation detection leveraging natural language processing processing technique. Some researchers, for example, focus on text features and experiment with natural language processing techniques, such as sentiment analysis [17]. The authors of this publication propose the use of this extra dimension as a source of auxiliary features. Finally, an emerging technique is the use of a combination of the previous two [18, 19].

While proponents of natural language processing point out that deep learning models tend to produce inexplicable black boxes that may lead to biased outputs [14], which is sometimes echoed by proponents of machine learning [18], the same researchers [18] rightfully point out that the bag-of-words nature of topic models impedes such methods from capturing features based on the sequential ordering of words. This is a weakness of note and why topic models should not be used alone when attempting to systematically detect

misinformation, especially considering the more difficult to detect subtle pieces of misinformation. The authors also classified misinformation detection methods as belonging to either traditional machine learning models, topic models, or deep learning models.

Researchers agree that the fake news detection problem is a complex one and has not yet seen a perfectly appropriate solution.

Some approaches attempt to model claims as binary true or false and run into issues of representing further nuance and complexity. For this reason, we will steer our research to rely on a score and focus on detecting suspicious or inorganic behavior rather than real or fake claims. Other works [16] use multi-platform datasets and attempt to model complex information structures by classifying claims between specific categories (here: “fake news”, “news bias”, “rumors”, and “clickbait”), and rely on annotations to build predictive models based on headline linguistic features, achieving an average effectiveness of 70.27%. Some researchers address the issue with classifying “realness” by representing both certainty and uncertainty [14] and accounting for user response and engagement. The authors found promising results and, as many other studies did [13], encouraged the use of wider arrays of features when attempting to detect social media misinformation. Researchers [14] also correctly point out many challenges of fake news detection. Such as multilingualism when relying on textual approaches, which has some researchers relying on meta-data or networking only approaches. Particularly challenging and effective misinformation also includes items which featured subtly inserted falsehood or half truths. The Multimedia nature of misinformation is another challenge.

Others use wide and deep models [18], relying on memorizing and generalizing information, which somewhat inspired our natural language processing based contribution, to advance interpretability and reduce unknown bias. These researchers also propose a framework model combining multiple design principles and detection methods. Although this particular study uses datasets of a slightly different nature: deceptive reviews and fraudulent emails

Using a self-constructed twitter dataset of 1,300 entries, researchers have been able to achieve an impressive near-real time 93% accuracy in detecting misinformation [15]. Twitter being a very prized source of data for such studies due to the wide array of metadata available [20]. One concern however is how scalability and ability to detect a very wide range of misinformation may become a hurdle for this model as it could detect merely dubious information. As opposed to our approach, these researchers ignored textual content and focused on networking and linguistic features. In contrast, other authors [21] found 49.2% accuracy with a much larger dataset of 34,918 claims. These claims were crawled from fact checking websites and include metadata, such as the creator of the misinformation, the checker, etc. This approach is more suited to predict performances for fact checking websites.

C. The Role of YouTube

From third party public resource and web traffic reports [22], we know that YouTube is the second most popular website, ceding the first spot to Google, and accounts for 20.4% of all search traffic. According to official YouTube sources [23], 1 billion hours of videos are watched each day. Another study by Cha et al. [24] found that 60% of YouTube videos are watched at least 10 times on the day they are posted. The authors also highlight that if a video does not attract viewership in the first few days after upload, it is unlikely to attract viewership later on. YouTube provides an overwhelming amount of streaming data: over 500 hours of videos are uploaded every minute on average. A number which was “only” 300 in 2013 [25]. In previous publications [26, 27] we identified YouTube as a potential vehicle of misinformation. We proposed the use of YouTube metadata for understanding and visualizing these phenomena by observing data trends. We also proposed the concept of movie barcodes as a tool for video summarization clustering [28]. In this publication, we present the movie barcode tool as a part of VTracker, as well as new video characterization tools. Previous research [29] has looked into engagement patterns of YouTube videos and highlighted the related videos engagement trends, later designated as the “rabbit hole effect” where users will be recommended increasingly relevant videos. In some cases, where the subject matter is a very polarizing one, this effect has been shown to be a contributing factor in user radicalization [30]. This last study takes the example of vaccine misinformation, which has attracted much interest from the information community. With some research highlighting that while users turn to YouTube for health information, many of the resources available failed to provide accurate information [31, 32], and public institutions should increase their online presence [33] to make reliable information more accessible. Recent research on the same subject leverages advanced NLP techniques on text entities, such as video comments [34] but we could find little work available on the video content itself.

D. Topic Modeling

To implement topic modeling, we use the Latent Dirichlet Allocation (LDA) model. Within the realm of Natural Language Processing (NLP), topic modeling is a statistical technique designed to categorize a set of documents within a number of abstract “topics” [35]. A “topic” is defined as a set of words outlining a general underlying theme. For each document, which in this case, is an individual item of misinformation in our data set, a probability is assigned that designates its “belongingness” to a certain topic. In this study, we use the popular LDA topic model due to its widespread use and proved performances [36]. One point of debate within the topic modeling community is the elimination of stop-words: i.e., analysts should filter common words from their corpus before training a model. Following recent research claiming that the use of custom stop-words adds few benefits [37], we followed the researchers’ recommendation and removed common words **after** the model had been trained.

Our model choice has seen use in previous research using LDA for short texts, specifically for short social media texts, such as tweets [38, 39, 40]. Some other social media research using homogeneous social media sources, such as tweets or blog posts use associated hashtags to provide further context to topic models [41]. We expand this research on social media corpora by focusing one of the largest information propagator on the web: YouTube.

In this paper, we propose to leverage topic models to understand the main underlying themes of misinformation and their evolution over time using a manually curated corpus of known fake narratives.

As a secondary goal, we observe the performances of different topic models for understanding online discourse. To accomplish this, we repeated our methodology on a secondary data set using a Hierarchical Dirichlet Process (HDP) model [42]. For our purposes, the major difference between the two models is that LDA models require a number of topics prior to training and will actively attempt to fit that number to the corpus, potentially leading to biased results. On the other hand, the HDP model infers the number of topics present in the corpus during training.

III. METHODOLOGY

This study uses a two-step methodology to produce relevant topic streams. First, through a manual curating process, we aggregate different misinformation narratives for later processing. We consider misinformation narratives, any narrative pushed through a variety of outlets (social media, radio, physical mail, etc.) that has been or is later believably disproved by a third party. This corpus constitutes our input data. Secondly, we use this corpus to train an LDA topic model and to generate subsequent topic streams for analysis. We describe these two steps in more details in the next sections.

A. Collection of Misinformation Stories

This is the set referred to as **Dataset-1**. Initially, the misinformation stories in our data set were obtained from a publicly available database created by EUvsDisinfo in March of 2020 [43]. EUvsDisinfo's database, however, was primarily focused on "pro-Kremlin disinformation efforts on the novel coronavirus". Most of these items represented false narratives that were communicating political, military, and healthcare conspiracy theories in an attempt to sow confusion, distrust, and public discord. Subsequently, misinformation stories were continually gleaned from publicly available aggregators, such as POLITIFACT, Truth or Fiction, FactCheck.org, POLYGRAPH.info, Snopes, Full Fact, AP Fact Check, Poynter, and Hoax-Slayer. The following data points were collected for each misinformation item: title, summary, debunking date, debunking source, misinformation source(s), theme, and dissemination platform(s). The time period of our data set is from January 22, 2020 to July 22, 2020, which is the COVID-19 breakout period. The data set is comprised of 543 total stories and 243 unique misinformation narratives. For many of the items, multiple platforms were used to spread the

misinformation. For example, oftentimes a misinformation item will be posted on Facebook, Twitter, YouTube, and as an article on a website. For our data set, the top platforms used for spreading misinformation were websites, Facebook, Twitter, YouTube, and Instagram, respectively. All the stories found by our team are made public through our partnership with the Arkansas Attorney General Office and can be found on our website.

B. Collection of YouTube Data

In order to observe results in uncontrolled, relevant social media environments, we also gathered YouTube data. We chose YouTube because it is a principal vector of information and communication between users and is heavily understudied. Using the official YouTube API, we performed separate searches for the following keywords on April 19th 2020: "Coronavirus, Corona, Virus, COVID19, COVID, Outbreak". The result is a set of the most popular videos at that time, as determined by YouTube's algorithm. From this search, we collected a total of 7,727 videos ranging mostly from January 1st to April 19th 2020. For this particular study, in order to focus on the most relevant videos possible, we selected only videos published between March 1st and March 31st (included). Like the previous set, this is a key month of the COVID-19 breakout period. This totals 444 videos, which is comparable to the number of narratives studied. For the purposes of this study, we will only look at the video titles. After selecting this corpus, we used the same API to collect comments posted in these videos and gathered a total of 652,120 comments. This is **Dataset-2**.

Based on a manual qualitative analysis of known alt-right public figures active on social media, a set of specific actors was identified and selected as seeds for preliminary data collection. YouTube data for our set of key actors was collected using the YouTube Data API according to the methodology described by Kready et al. [44]. During post-processing, the dataset was filtered to focus in on the two months prior and post the January 6, 2021 U.S. Capitol riot event, resulting in a timeframe of analysis of November 1, 2020 to March 1, 2021. We chose this period because that is where most discussion revolving around vaccines can be found. This is **Dataset-3**. In order to comply with YouTube's terms of service, this data cannot be made public.

C. Topic Modeling

In order to derive lexical meaning from this corpus, we built a pipeline executing the following steps. First, we processed each document in our text corpus. All that is needed is a text field identified by a date. Because in most cases of word of mouth or social media it is impossible to pinpoint the exact date the idea first emerged, we use the date of publication of the corresponding third party "debunk piece". We trained our LDA model using the Python tool Gensim, with the methodology and pre-processing best practices as described by its author [45] as well as best stop words practices as described earlier [37]. In this study, we found that generating

20 different topics best matched the ground truth as reported by the researchers curating the misinformation stories.

Still using Gensim, we also trained an alternative topic model using HDP [42]. The process is the same except for the number of topics. HDP infers the number of topics in a corpus (with a default threshold of 150). Therefore, we only select the first 20 topics, ordered by α , the weight of each document to topic distribution.

Once the models have been trained, we ordered the documents by date and created a numpy matrix where each document is given a score for each topic produced by the model. This score describes the probability that the given document is categorized as being part of a topic, i.e., if a probability score is high enough (more details below), the document is considered to be part of the topic. Through manual observations, we noticed that many documents retain "noise probability", giving them a probability to be in every topic of around 1% to 5%. For this reason, we set the probability threshold to a comfortable 10% and noticed consistent results. This allowed us to leverage the Python Pandas library to plot a chronological graph for each individual topic. We averaged topic distribution per day and used a moving average window size of 20 unless otherwise specified. This helped in highlighting the overarching patterns of the different narratives. Note, however, that this process hides some early and late data in our set as there are less data points around that time.

IV. RESULTS

In this section, we discuss the thoughts of our data collection team and the ground truth as they were observed, and compare these with the results obtained through our topic modeling visualization tool.

A. Prominent Misinformation Themes Over Time

Although a variety of misinformation themes were identified, particular dominant themes stood out, changing over time. These themes were considered as dominant based on a simple sum of their frequency of occurrence in our data set. During the month of March, the prominent misinformation theme was the promotion of remedies and techniques to supposedly prevent, treat, or kill the novel coronavirus. During the month of April, the prominent themes still included the promotion of remedies and techniques, but additional prominent themes began to stand out. For example, several misinformation stories attempted to downplay the seriousness of the novel coronavirus. Others discussed the anti-malaria drug hydroxychloroquine. Others promoted the idea that the virus was a hoax meant to defeat President Donald Trump. Others consisted of various attempts to attribute false claims to high-profile people, such as politicians and representatives of health organizations. Also in April, although first signs of these were seen in March, the idea that 5G caused the novel coronavirus began to become more prevalent. During the month of May, the prominent themes shifted to predominantly false claims made by high-profile people, followed by attempts to convince citizens that face masks are either more harmful

than not wearing one, or are ineffective at preventing COVID-19, and how to avoid rules that required their use. The number and variety of identity theft phishing scams also increased during May. Misinformation items attempting to attribute false claims to high-profile people continued throughout May. Also becoming prominent in May were misinformation items attempting to spread fear about a potential COVID-19 vaccine, and items promoting the use of hydroxychloroquine. During the month of June, the prominent theme shifted significantly to attempts to convince citizens that face masks are either more harmful than not wearing one, and how to avoid rules that required their use. Phishing scams also remained prominent during June. During the month of July, the dominant themes of the misinformation items shifted back to attempts to downplay the deadliness of the novel coronavirus. Another prominent theme in July was the proliferation of attempts to convince the public that COVID-19 testing is inflating the results.

B. Topic Streams

After using the tool described in Section III-C, we generated the graphs and tables described and discussed in this section. Our data for this step contained 243 unique misinformation narratives spanning from January 2020 to June 2020, when we stopped data collection. The data was curated by our research team through the process described in the methodology. Each entry contains, among other fields, a "date" used as a chronological identifier, a "title" describing the general idea the misinformation is attempting to convey, and a "theme" field putting the story in a concisely described category. For example, a story given the title "*US Department of Defense has a secret biological laboratory in Georgia*" is categorized in the following theme: "*Western countries are likely to be purposeful creators of the new virus.*" Each topic was represented by an identification number up to 20 and a set of 10 words. We picked the three most relevant words that best represented the general idea of each topic. Notably, obvious words, such as *covid* or *coronavirus* were removed from the topic descriptions since they are common for every topic.

In Tables I and II, we described some of the twenty topics found by each of our LDA models. These topics were chosen because they each described a precise narrative and have a low topic distribution (or proportion within the corpus). A low proportion is desirable because this indicates the detection of a unique narrative within the corpus; as opposed to an overarching topic including general words, such as "world", "outbreak", or "pandemic". Do note that topic inclusiveness is not exclusive and documents can be part of multiple topics. This becomes apparent in Table I: from our topic model, we found a dominant topic encompassing 68% of narratives. It includes words such as "Trump", "outbreak", "president", etc. Some other narratives also included words such as "flu", "news", or "fake". Because the evolution of these narratives are consistent across the corpus and show little temporal fluctuation, we chose not to report on them further. For these reasons, the narratives we focused on below show a low percentage of distribution (Tables I & II).

TABLE I
MOST FREQUENT DOMINANT TOPICS FROM TITLES.

Topic ID	Word 1	Word 2	Word 3	Proportion
10	china	chinese	spread	2%
12	scam	hydroxy...	health	2%
17	state	donald	trump	2%
18	vaccine	gates	bill	5%

TABLE II
MOST FREQUENT DOMINANT TOPICS FROM THEMES.

Topic ID	Word 1	Word 2	Word 3	Proportion
3	fear	spread	western	2%
9	predicted	pandemic	vaccine	2%
16	phishing	hydroxy...	vaccine	2%

1) *Using narrative titles as a corpus - Dataset-I*: The general narratives described by the topics were thus:

- Topic 10 described the narratives related to the Chinese government and its responsibility in the spread of the virus. These stories represented an estimated 2% of the 243 stories collected.
- Topic 12 described the narratives related to personal health and scams or misinformation, such as the benefits of hydroxychloroquine. These stories represented an estimated 2% of the 243 stories collected.
- Topic 17 described the narratives related to the response of Donald Trump and his administration. These stories represented an estimated 2% of the 243 stories collected.
- Topic 18 described the narratives related to the involvement of Bill Gates in various conspiracies, mostly linked to vaccines. These stories represented an estimated 4% of the 243 stories collected.

Related studies have found that finger-pointing narratives usually lead to negative sentiment and toxicity in online communities [38, 46, 39].

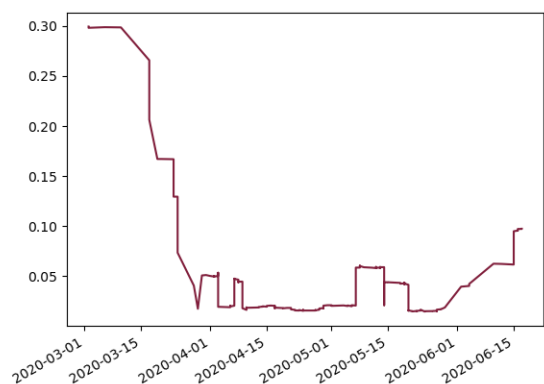


Fig. 1. Topic's probability distribution of titles for topic 10 (keywords: china, chinese, spread) over time (LDA model)

Figure 1 shows the evolution of Topic 10, the topic de-

scribing China-related narratives. It shows that these narratives were already in full force from the beginning of our corpus and slowly came to a near halt during the month of April. We notice a short spike again towards the end of the corpus during the month of June. This is consistent with the ground truth of online narratives that focused on the provenance of the virus during the early stages.

Figure 2 shows the evolution of Topic 12, the topic describing narratives related to health, home remedies, and general hoaxes and scams stemming from the panic. We can see it was consistent with the rise of cases in the United States and panic increased as with the spread of the virus. It is interesting to note that this figure roughly coincides with the daily number of confirmed cases for this time period [47].

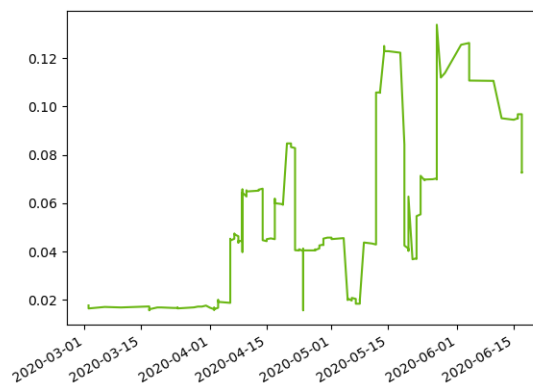


Fig. 2. Topic's probability distribution of titles for topic 12 (keywords: hydroxychloroquine, health, scam) over time (LDA model)

Figure 3 shows the evolution of Topic 17. This topic described stories related to Donald Trump and his administration. These stories generally referred to claims that the virus was manufactured as a political strategy, or claims that various public figures were speaking out against the response of the Trump administration.

Figure 4 shows the evolution of Topic 18. This topic described stories such as Bill Gates and his perceived involvement with a hypothetical vaccine, and other theories describing the virus' appearance and spread as an orchestrated effort. As with Figure 1, these narratives were especially strong early on (albeit this narrative remained active for a slightly longer time), before coming to a near halt.

We notice that, as theories about the origins of the virus slowed down, hoaxes and scams increased - as shown on Figure 2. This includes attempts at identity theft, especially toward senior citizens, and attempts to sell miracle cures and miracle personal protection items.

2) *Using narrative themes as a corpus*: For this section, we inputted narrative themes as the corpus. Note that the topic IDs are independent from the previous set of topics using titles. Similarly to Section IV-B1, we found a dominant topic

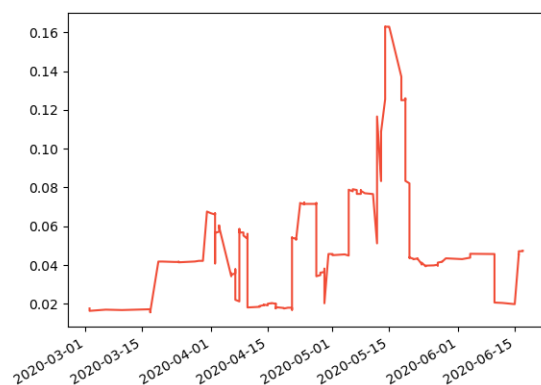


Fig. 3. Topic's probability distribution of titles for topic 17 (keywords: donald, trump, state) over time (LDA model)

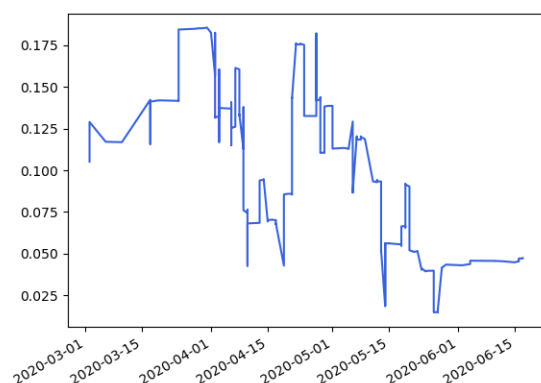


Fig. 4. Topic's probability distribution of titles for topic 18 (keywords: bill, gates, vaccine) over time (LDA model)

encompassing 68% of narratives as well. This time including words such as “attempt”, “countries”, and “purposeful”. As for section IV-B1, we chose not to report on that topic as well as other smaller but general topics showing little fluctuation. Therefore, the narratives we focused on below show a low percentage of distribution. The general narratives described by the topics are thus:

- Topic 3 described the narratives related to the speculations on the spread of the virus, especially in an international relations context. These stories represented an estimated 2% of the 243 stories collected.
- Topic 9 described the narratives related to stories claiming the creation and propagation of the virus were either designed or predicted, along with voices claiming a vaccine already exists. These stories represented an estimated 3% of the 243 stories collected.
- Topic 16 described the narratives related to personal health and scams or misinformation such as the bene-

fits of hydroxychloroquine. These stories represented an estimated 2% of the 243 stories collected.

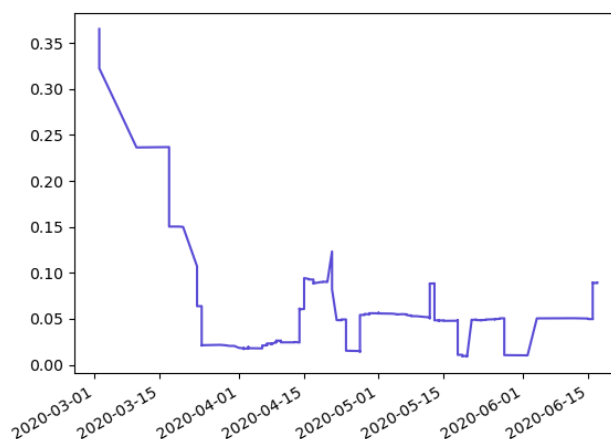


Fig. 5. Topic's probability distribution of themes for topic 3 (keywords: fear, spread, western) over time (LDA model)

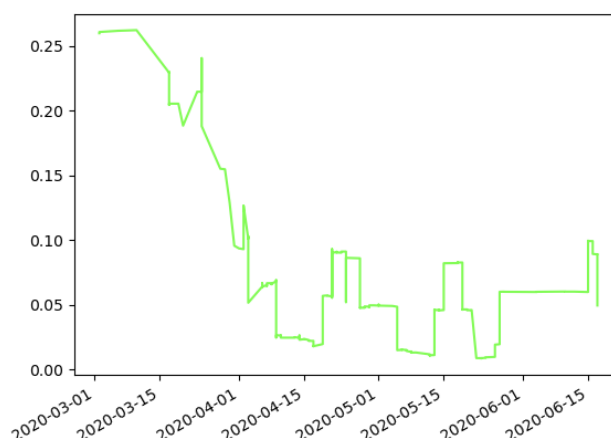


Fig. 6. Topic's probability distribution of themes for topic 9 (keywords: predicted, pandemic, vaccine) over time (LDA model)

Figure 5 shows the evolution of Topic 3. It is linked to early fear of the virus and presented narratives as opposing the western block with the East, notably China. It matched closely with Figure 1 and its China-related narratives. In both cases, we see an early dominance of the topic followed by a near halt as the virus touched the United States.

Figure 6 describes the evolution of narratives claiming the virus was predicted or even designed. This figure is consistent with the results shown by Figure 4 which shows claims regarding Bill Gates, early vaccines, etc. They both showed stories of early knowledge of the virus and peaked early,

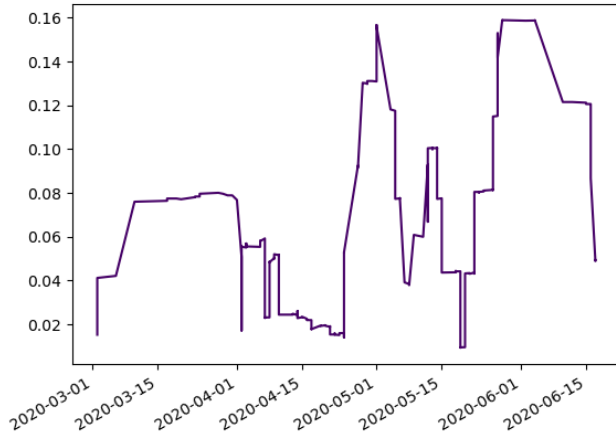


Fig. 7. Topic's probability distribution of themes for topic 16 (keywords: hydroxychloroquine, vaccine, phishing) over time (LDA model)

appearing more or less sporadically as time goes on and as cases increased.

Figure 7 is parallel to Figure 2. Both showed hoax stories promoting scams and health-related misinformation. We noticed an early rise in Figure 7, most likely due to the inclusion of the keyword “vaccines” in the topic, which caused some overlap with Topic 9 as shown in Figure 6.

C. YouTube Data

In this section, we explore how different topic models affect our YouTube data set. We focus on a subset of data published during the month of March to limit the number of comments to process.

1) *YouTube videos - Dataset-2*: The first observation for this set is that our HDP model did not perform as well as the LDA model. Our HDP model identified one dominant topic present in 87% of videos, with seemingly unrelated identifying keywords (“cases”, “hindi”, “nyc”, “italy”). While the rest of the topics are present in around 1% of the videos. The second most dominant topic (1.8% of documents) also features contradicting words such as “plandemic” and “hospitals”. One would expect language connected to the plandemic narrative in this topic, such as mentions of “Bill Gates” like we saw in the previous sets, but it is missing. There are two possible explanations for this. One is that performance may be due to the size of the set (more in the next section) as there were only 444 video titles processed. The other is that the set features numerous multilingual titles, which may skew results.

Our LDA model, however, behaved as expected and was able to identify major topics, mostly news videos (Topics 0 & 17), as well as what we suspect to be a vehicle of misinformation (Topic 6). As described in Table III and visualized in Figure 8. Figure 8 has been smoothed with a moving average equal to 15% of the total data set size (67) in order to improve legibility and reveal patterns. Due to most

TABLE III
RELEVANT TOPICS FROM VIDEO TITLES (LDA MODEL)

Topic ID	Word 1	Word 2	Word 3	Proportion
0	news	update	live	12.4%
17	outbreak	doctor	cases	7.6%
6	plandemic	dempanic	dem	2.7%

of the videos being published late in March, this has removed some granularity towards early March from the plot. However, we notice news topics staying fairly consistent while Topic 6 sees a decline, possibly as the number of covid cases makes maintaining the “fake pandemic” narrative more difficult and other misinformation narratives take over, such as various scams and hoaxes as seen in section IV-B1.

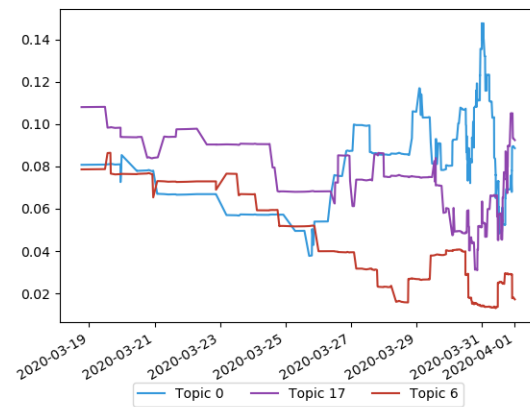


Fig. 8. Topic's probability distribution of topics 0, 17 & 6 over time (LDA model)

2) *YouTube comments - Dataset-2*: Contrary to the previous section, this is a much larger data set of 652,120 comments. This led to better performances, but still inferior to the LDA model. Our HDP model was able to identify non-English comments (11.4% German, 4.5% Spanish, 1.6% French). More importantly, the HDP model identified a topic that could be described as polarizing discourse, some of the most frequent terms including “Trump”, “China”, and “virus”. This topic accounts for 6.6% of the corpus. The evolution of this topic is shown by Figure 9 where we notice that topic is on an upward trend. A moving average equal to 3% of the set size is applied to better identify patterns.

On this very large set, our HDP model somewhat outperformed LDA for our purposes as it was able to identify a probable topic for misinformation. When applied to our comments set, our LDA model mostly found general terms while also successfully isolating non-English comments. The model did identify a topic with some toxic language and some that could be used in a hostile way or communicate sinophobic sentiments (Topic 7 & 17). See Table IV. While discussion of China has so far been on a downward trend since the start

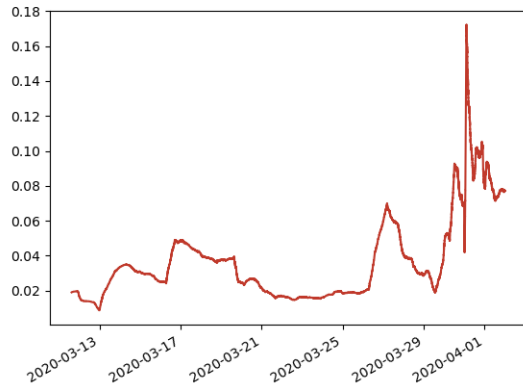


Fig. 9. Topic's probability distribution of Topic 4 over time (HDP model)

of the pandemic, the mention of the term “virus” along with “china” suggests toxic behavior. See Figure 10.

TABLE IV
RELEVANT TOPICS FROM FROM DATASET-2 COMMENTS (LDA MODEL).

Topic ID	Word 1	Word 2	Word 3	Proportion
7	china	virus	made	3.5%
17	trump	dumb	bats	3.3%

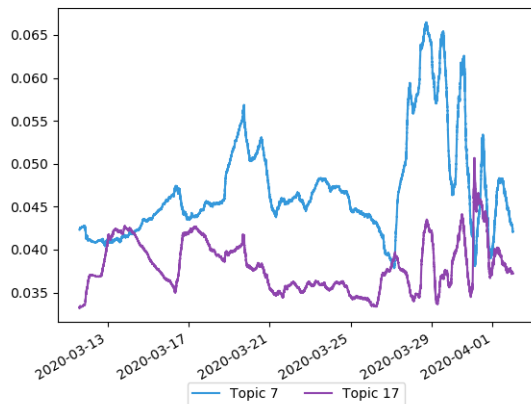


Fig. 10. Topic's probability distribution of Topic 7 & 17 over time (LDA model)

3) *YouTube comments - Dataset-3*: This larger set of 1,664,123 comments comes from efforts relating to content liked with the January 6, 2021 U.S. Capitol riot [48]. Due to its larger size, this set is our test bed for our new Pipeline Framework.

As is illustrated in Figure 11, this architecture is a node-based system where the framework first reads raw data, then have each node ingest filtered or annotated data from the previous one. These nodes can be chained in any order but, in

this study, we demonstrate what could be labelled as the data filtering layer. As was suggested in our previous publication [1], we are now using the more objective HDP model to divide a corpus into topics and then identify which topic to filter and send to our LDA model to identify latent narratives.

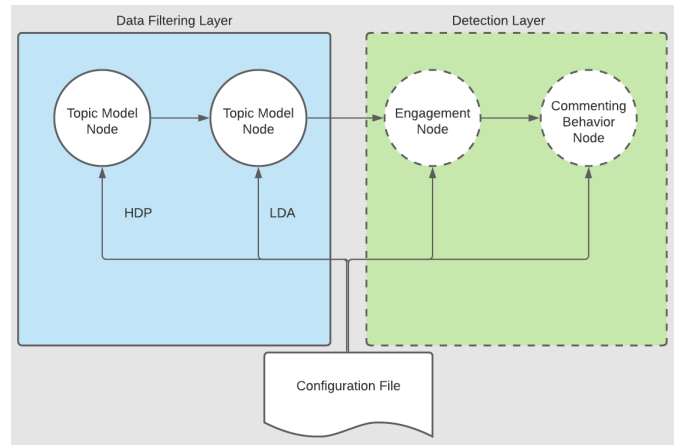


Fig. 11. Pipeline Framework

TABLE V
RELEVANT TOPICS FROM FROM DATASET-3 COMMENTS (HDP MODEL).

Topic ID	Word 1	Word 2	Word 3	Word 4	Proportion
3	gender	women	men	man	8.2%
2	covid	vaccine	even	know	6%
5	trump	ben	think	biden	3.1%

From Table V, which shows some of the most relevant words from the 20 topics we retained (in order of prominence within the dataset), we notice that Topic 2 is especially relevant to our subject at hand. For this reason, the comments belonging (where “belongingness” is characterized by a probability superior to 0.3 of belonging to a given topic) to that topic are sent to the next node where our LDA model is then retrained on these comments. The resulting main topics of interest and their descriptive keywords are described in Table VI.

TABLE VI
RELEVANT TOPICS FROM DATASET-3 COMMENTS (LDA MODEL)

Topic ID	Word 0	Word 1	Word 2	Word 3	Proportion
15	leftist	welcome	tears	change	5%
8	rumble	back	joined	parler	4.3%
1	trump	address	back	party	2.9%

Table VI and its temporal visualizations tell give us the following insight: From the keywords described in Topic 15, there seems to be a celebration of some event perceived as a victory over the opposing party. This event is represented within the graph in Figure 12 by a very obvious peak.

Topic 8 shown on Figure 13 aggregates keywords discussing other apps focused on free speech and anonymity. Interestingly, this type of speech has seen a very big revival shortly

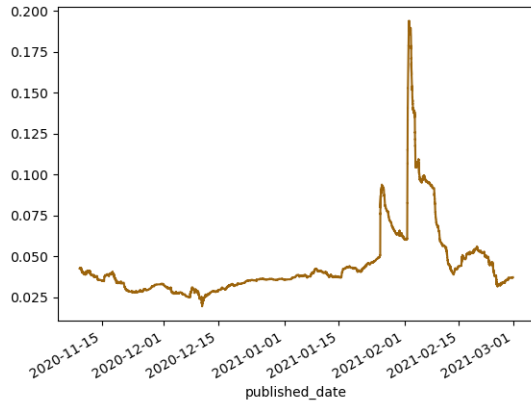


Fig. 12. Topic's probability distribution of Topic 15 over time (LDA model)

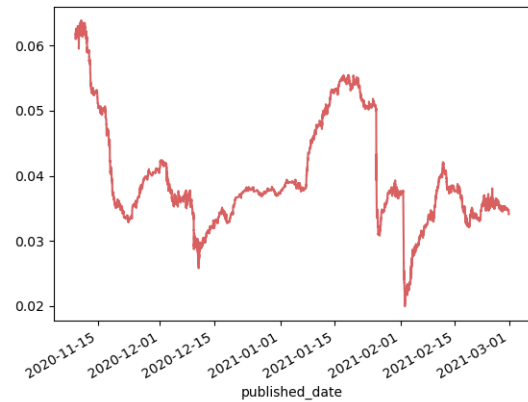


Fig. 14. Topic's probability distribution of Topic 1 over time (LDA model)

before the events on January 6th, and then another spike directly after with periodic movement following. This may suggest some level of organization or at least a desire to move away from mainstream platforms that could have been a factor in the Capitol riots.

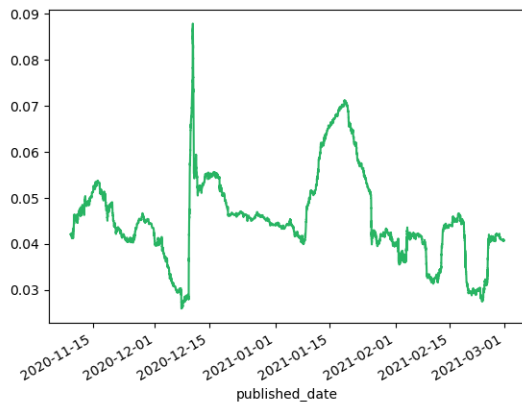


Fig. 13. Topic's probability distribution of Topic 8 over time (LDA model)

Finally, Topic 1 shown on Figure 14 shows discourse surrounding Donald Trump and his appearances. Unsurprisingly, the popularity of this topic has been on the decline since the 2020 presidential elections and then saw a revival around the January 6th riots. We also notice some periodicity.

Chaining topic models to help filter larger data sets has shown good results that are explainable by real world events and is a promising start to further enrich our framework for deviant behavior detection. Unlike deep learning networks, every node and features is strictly defined, reducing risk for bias. Of course, one limitation of such method becomes the bias of human experts designing features and also the risk of models becoming outdated. To address these weak points, we

will further expand the pipeline to accept fully modular and interchangeable nodes.

D. Future Works

As shown in Figure 11, our framework will be appended with more nodes whose goal is to annotate and “detect” misinformation by providing score based on commenting behavior as well as engagement behavior in the source video of the comment. This is one way to tackle multimedia misinformation as video misinformation has presented a significant challenge, and threat, especially due to the popularity of such video content. The design of the framework aims to allow for chaining nodes in any order, and one other goal will be to automate this process to obtain and measure the most accurate results, but also to let researchers contribute their own nodes.

E. Public Website and Citizen Science

We have put together a website with known cases of misinformation about COVID-19. As of January 2021, we have documented close to 600 cases that we identified from numerous sources (social media - Facebook, YouTube, Twitter, blogs, fake websites, robocalls, text/SMS, WhatsApp, Telegram, and an array of such apps) - see Figure 15 [4]. The principal difference between our effort and other similar efforts by Google and social media companies is that we are paying special attention to cases of misinformation and scammers that are affecting our region, while also including global cases. We update the database periodically with newly detected cases. Moreover, we have put together a list of over 50 tips on the website for people to learn how to spot misinformation. We have also provided a feature for people to report fake websites or scams that are not currently in our database.

Our website uses a three-pronged approach:

- We identify new cases of fake websites, misinformation content, and bad actors. We use social network analysis and cyber forensic methodologies to identify such cases.
- We believe in educating people to be self-reliant because we might not be able to detect all possible cases of



Fig. 15. COVID-19 Website Front page - Showing the latest misinformation stories

misinformation. Therefore, we go through identified cases and prepare a list of common telltale signs to detect whether a piece of information is genuine or not.

- For the cases that are not in our database and people cannot distinguish, we provide a way for people to submit cases of misinformation that we have not captured in our database.

The database of known misinformation cases and scams is publicly available for the research community to use [4]. We envision a tremendous value of this research database to various disciplines. The website is available for regulatory bodies (Arkansas Office of the Attorney General) and any citizen, which serves as an invaluable resource to not only educate people of the misinformation and scams about COVID-19 but also assisting legal authorities in taking action against malicious actors and groups. We are assisting the Arkansas' Attorney General's office by providing reports on cyber forensic evidence about scam/fake websites reported by people - see Figure 16. The study presented in this paper will be developed into the system as a real-time campaign tracking feature. We will continue to work with Arkansas' Attorney General's office to assist in their effort to combat COVID-19 misinformation and scams to protect Arkansans.

V. CONCLUSION

In this study which expands our last publication [1], we have highlighted some of the narratives that surfaced during the COVID-19 pandemic. From January 2020 to July 2020, we collected 243 unique misinformation narratives and proposed a tool to observe their evolution. We have shown the potential of using topic modeling visualization to get a bird's eye view of the fluctuating narratives and an ability to quickly gain a

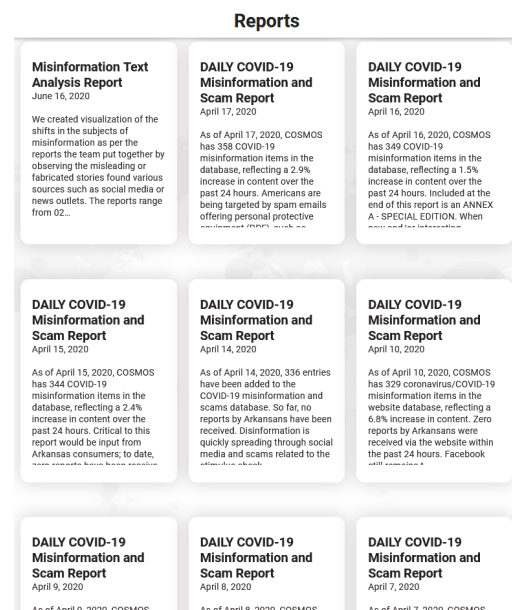


Fig. 16. COVID-19 Website Reports page - Showing all reports made to the Arkansas Attorney General Office

better understanding of the evolution of individual stories. We have seen that the tool is efficient to chronologically represent actual narratives pushed to various outlets, as confirmed by the ground truth observed by our misinformation curating team and independent international organizations. Working with the Arkansas Office of the Attorney General, this study illustrates a relatively quick technique for allowing policy makers to monitor and assess the diffusion of misinformation on online social networks in real-time, which will enable them to take a proactive approach in crafting important theme-based communication campaigns to their respective citizen constituents. We have made most of our findings available online to support this effort.

In addition to these results, we have introduced much larger datasets, one of 652,120 YouTube comments, and another of 1,664,123 more comments. To accommodate these sets, we introduce a new node-based framework which functions as a pipeline where nodes can be interchangeably used to filter and annotate documents. At this current stage, the framework supports topic model nodes based on the LDA and HDP model. By feeding into our LDA model documents belonging to specific topics as identified by our HDP model, we are able to focus on specific communities of interest and reveal latent patterns and events within those communities. The future of this tool is in the addition of more nodes that will examine wider features, such as commenting behavior and engagement behavior with videos and channels where comments are posted to detect suspicious behavior.

ACKNOWLEDGEMENT

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933,

ACI-1429160, and IIS-1110868), 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, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). 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] T. Marcoux, E. Mead, and N. Agarwal. "Studying the Dynamics of COVID-19 Misinformation Themes". en. In: *2021 Seventh International Conference on Human and Social Analytics (HUSO)*. Nice, France, July 2021.
- [2] S. Al-khateeb et al. *Exploring Deviant Hacker Networks (DHM) on Social Media Platforms*. Vol. 11. Journal of Digital Forensics, Security and Law, 2016, pp. 7–20. DOI: 10.15394/jdfsl.2016.1375. URL: <https://commons.erau.edu/jdfsl/vol11/iss2/1>.
- [3] M. Calabresi. *Inside Russia's Social Media War on America*. 2017. URL: <https://time.com/4783932/inside-russia-social-media-war-america/>. (accessed: 01.19.2021).
- [4] COSMOS. *COSMOS - COVID-19 Misinformation Tracker*. 2021. URL: <https://cosmos.ualr.edu/covid-19>. (accessed: 06.15.2021).
- [5] G. Pennycook et al. "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention". In: vol. 31. 7. eprint: <https://doi.org/10.1177/0956797620939054>. 2020, pp. 770–780. DOI: 10.1177/0956797620939054. URL: <https://doi.org/10.1177/0956797620939054>.
- [6] R. Kouzy et al. "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter". eng. In: vol. 12. 3. Publisher: Cureus. Mar. 2020, e7255–e7255. DOI: 10.7759/cureus.7255. URL: <https://pubmed.ncbi.nlm.nih.gov/32292669>.
- [7] H. Sha et al. *Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives*. 2020. arXiv: 2004.11692 [cs.LG].
- [8] Q. Liu et al. "Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach". In: vol. 22. 4. Apr. 2020, e19118. DOI: 10.2196/19118. URL: <http://www.jmir.org/2020/4/e19118/>.
- [9] H. Zhang et al. "Detecting misinformation in online social networks before it is too late". In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2016, pp. 541–548. DOI: 10.1109/ASONAM.2016.7752288.
- [10] B. Al Asaad and M. Erascu. "A Tool for Fake News Detection". In: *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. 2018, pp. 379–386. DOI: 10.1109/SYNASC.2018.00064.
- [11] S. Lyu and D. C.-T. Lo. "Fake News Detection by Decision Tree". In: *2020 SoutheastCon*. 2020, pp. 1–2. DOI: 10.1109/SoutheastCon44009.2020.9249688.
- [12] A. Jain and A. Kasbe. "Fake News Detection". In: *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECs)*. 2018, pp. 1–5. DOI: 10.1109/SCEECs.2018.8546944.
- [13] S. Yu and D. Lo. "Disinformation Detection Using Passive Aggressive Algorithms". In: *Proceedings of the 2020 ACM Southeast Conference*. ACM SE '20. event-place: Tampa, FL, USA. New York, NY, USA: Association for Computing Machinery, 2020, pp. 324–325. ISBN: 978-1-4503-7105-6. DOI: 10.1145/3374135.3385324. URL: <https://doi.org/10.1145/3374135.3385324>.
- [14] Q. Zhang et al. "Reply-Aided Detection of Misinformation via Bayesian Deep Learning". In: Feb. 2019. DOI: 10.1145/3308558.3313718.
- [15] L. van de Guchte et al. "Near Real-Time Detection of Misinformation on Online Social Networks". In: *Disinformation in Open Online Media*. Ed. by M. van Duijn et al. Cham: Springer International Publishing, 2020, pp. 246–260. ISBN: 978-3-030-61841-4.
- [16] N. Lee et al. "On Unifying Misinformation Detection". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 5479–5485. DOI: 10.18653/v1/2021.naacl-main.432. URL: <https://aclanthology.org/2021.naacl-main.432>.
- [17] M. Alonso Pardo et al. "Sentiment Analysis for Fake News Detection". In: *Electronics* 10 (June 2021), p. 1348. DOI: 10.3390/electronics10111348.
- [18] Y. Chai et al. "Disinformation Detection in Online Social Media: An Interpretable Wide and Deep Model". en. In: *SSRN Electronic Journal* (2021). ISSN: 1556-5068. DOI: 10.2139/ssrn.3879632. URL: <https://www.ssrn.com/abstract=3879632> (visited on 09/14/2021).
- [19] K. Pelrine, J. Danovitch, and R. Rabbany. "The Surprising Performance of Simple Baselines for Misinformation Detection". In: *Proceedings of the Web Conference 2021*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 3432–3441. ISBN: 978-1-4503-8312-7. URL: <https://doi.org/10.1145/3442381.3450111>.
- [20] F. Pierri, C. Piccardi, and S. Ceri. "A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter". In: *EPJ Data Science* 9.1 (Nov. 2020), p. 35. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-020-00253-8. URL: <https://doi.org/10.1140/epjds/s13688-020-00253-8>.
- [21] I. Augenstein et al. "MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4685–4697. DOI: 10.18653/v1/D19-1475. URL: <https://aclanthology.org/D19-1475>.
- [22] *Youtube.com Traffic, Demographics and Competitors - Alexa*. URL: <https://www.alexa.com/siteinfo/youtube.com> (visited on 09/15/2021).
- [23] *How YouTube Works - Product Features, Responsibility, & Impact*. URL: <https://www.youtube.com/intl/en-GB/howyoutubeworks/> (visited on 09/15/2021).
- [24] M. Cha et al. "Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems". In: *IEEE/ACM Transactions on Networking* 17.5 (Oct. 2009), pp. 1357–1370. ISSN: 1558-2566. DOI: 10.1109/TNET.2008.2011358.
- [25] J. Hale. *More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute*. May 2019. URL: <https://www.youtube.com/trending>

- <http://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/> (visited on 06/13/2021).
- [26] M. N. Hussain et al. "Analyzing Disinformation and Crowd Manipulation Tactics on YouTube". In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '18. event-place: Barcelona, Spain. IEEE Press, 2018, pp. 1092–1095. ISBN: 978-1-5386-6051-5.
- [27] T. Marcoux et al. "Understanding Information Operations Using YouTubeTracker". In: *IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume*. WI '19 Companion. Thessaloniki, Greece: Association for Computing Machinery, 2019, pp. 309–313. ISBN: 9781450369886. DOI: 10.1145/3358695.3360917. URL: <https://doi.org/10.1145/3358695.3360917>.
- [28] R. Erol et al. "YouTube Video Categorization Using Moviebarcode". en. In: *The Sixth International Conference on Human and Social Analytics (HUSO 2020)*. Porto, Portugal, 2020.
- [29] X. Cheng, C. Dale, and J. Liu. "Statistics and Social Network of YouTube Videos". In: *2008 16th International Workshop on Quality of Service*. 2008, pp. 229–238. DOI: 10.1109/IWQOS.2008.32.
- [30] L. Tang et al. "Down the Rabbit Hole" of Vaccine Misinformation on YouTube: Network Exposure Study". In: *J Med Internet Res* 23.1 (Jan. 2021), e23262. ISSN: 1438-8871. DOI: 10.2196/23262. URL: <http://www.ncbi.nlm.nih.gov/pubmed/33399543>.
- [31] C. H. Basch et al. "Preventive Behaviors Conveyed on YouTube to Mitigate Transmission of COVID-19: Cross-Sectional Study". In: *JMIR Public Health Surveill* 6.2 (Apr. 2020), e18807. ISSN: 2369-2960. DOI: 10.2196/18807. URL: <http://www.ncbi.nlm.nih.gov/pubmed/32240096>.
- [32] H. O.-Y. Li et al. "YouTube as a source of information on COVID-19: a pandemic of misinformation?" In: *BMJ Global Health* 5.5 (May 2020), e002604. DOI: 10.1136/bmjgh-2020-002604. URL: <http://gh.bmj.com/content/5/5/e002604.abstract>.
- [33] G. Donzelli et al. "Misinformation on vaccination: A quantitative analysis of YouTube videos". In: *Human Vaccines & Immunotherapeutics* 14.7 (2018). Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/21645515.2018.1454572>, pp. 1654–1659. DOI: 10.1080/21645515.2018.1454572. URL: <https://doi.org/10.1080/21645515.2018.1454572>.
- [34] J. C. Medina Serrano, O. Papakyriakopoulos, and S. Hegelich. "NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube". In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics, July 2020. URL: <https://www.aclweb.org/anthology/2020.nlpCOVID19-acl.17>.
- [35] D. M. Blei, J. D. Lafferty, and A. N. Srivastava. *Text Mining: Classification, Clustering, and Applications*. CRC Press, 2009, pp. 71–88.
- [36] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation". In: vol. 3. *Journal of Machine Learning Research*, 2003, pp. 993–1022.
- [37] A. Schofield, M. Magnusson, and D. Mimno. "Pulling Out the Stops: Rethinking Stopword Removal for Topic Models". In: *15th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 2. Association for Computational Linguistics. 2017, pp. 432–436.
- [38] A. Abd-Alrazaq et al. "Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study". In: vol. 22. 4. 2020, e19016. DOI: 10.2196/19016. URL: <http://www.ncbi.nlm.nih.gov/pubmed/32287039>.
- [39] R. Chandrasekaran et al. "Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study". In: vol. 22. 10. Oct. 2020, e22624. DOI: 10.2196/22624. URL: <http://www.jmir.org/2020/10/e22624/>.
- [40] Y. Zhang, W. Mao, and J. Lin. "Modeling Topic Evolution in Social Media Short Texts". In: *2017 IEEE International Conference on Big Knowledge (ICBK)*. 2017, pp. 315–319.
- [41] M. H. Alam, W.-J. Ryu, and S. Lee. "Hashtag-based topic evolution in social media". In: vol. 20. 6. Nov. 2017, pp. 1527–1549. DOI: 10.1007/s11280-017-0451-3. URL: <https://doi.org/10.1007/s11280-017-0451-3>.
- [42] Y. W. Teh et al. "Hierarchical Dirichlet Processes". In: vol. 101. 476. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1198/016214506000000302>. 2006, pp. 1566–1581. DOI: 10.1198/016214506000000302. URL: <https://doi.org/10.1198/016214506000000302>.
- [43] EUvsDisinfo. *EUvsDisinfo. March 16, 2020. The Kremlin and Disinformation About Coronavirus*. 2020. URL: <https://euvsdisinfo.eu/the-kremlin-and-disinformation-about-coronavirus/>. (accessed: 01.19.2021).
- [44] J. Kready et al. "YouTube Data Collection Using Parallel Processing". In: *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 2020, pp. 1119–1122. DOI: 10.1109/IPDPSW50202.2020.00185.
- [45] R. Řehůřek and P. Sojka. "Software Framework for Topic Modelling with Large Corpora". In: May 2010, pp. 45–50. DOI: 10.13140/2.1.2393.1847.
- [46] H. Budhwani and R. Sun. "Creating COVID-19 Stigma by Referencing the Novel Coronavirus as the "Chinese virus" on Twitter: Quantitative Analysis of Social Media Data". In: vol. 22. 5. May 2020, e19301. DOI: 10.2196/19301. URL: <http://www.ncbi.nlm.nih.gov/pubmed/32343669>.
- [47] H. Ritchie et al. *United States: Coronavirus Pandemic - Our World in Data*. 2020. URL: <https://ourworldindata.org/coronavirus/country/united-states?country=~USA>. (accessed: 07.29.2020).
- [48] *U.S. Capitol Riot - The New York Times*. URL: <https://www.nytimes.com/spotlight/us-capitol-riots-investigations> (visited on 09/15/2021).

Twitter Search Interface for Looking Back at TV Dramas

Taketoshi Ushiyama

Faculty of Design

Kyushu University

Fukuoka, Japan

email: ushiyama@design.kyushu-u.ac.jp

Haruka Nagai

School of Design

Kyushu University

Fukuoka, Japan

email: 1ds16408n@s.kyushu-u.ac.jp

Abstract—In recent years, while TV dramas are being broadcast, many comments and discussions about the dramas are posted on Twitter. These tweets are called “live tweets,” and after watching a drama, users can search for live tweets about scenes of interest to them, enjoy the impressions of other viewers, and deepen their thinking from a different perspective. However, in the current Twitter search function, even if the user searches for a keyword of the target scene, the tweets including the keyword are only presented in sequential order of posting. It takes time for users to find the live tweets of the scene they are interested in. This paper proposes an interface that can efficiently look back at dramas by visualizing the similarity distribution of specific keywords by time for live tweets posted during the drama. In this paper, we propose two Word2Vec-based methods and one TF-IDF-based method to calculate the similarity between keywords and live tweets posted during segments of the drama for visualization. From the results of the evaluation experiments, we found that TF-IDF-based method is the most suitable method for calculating the similarity between keywords and situation segments for visualization. In addition, the results of a usability survey of subjects using the prototype system showed that the proposed interface was able to capture the characteristics of TV drama scenes and was an effective way to look back at TV dramas.

Index Terms—Twitter; social viewing; live-tweeting; TV drama; looking back.

I. INTRODUCTION

In recent years, social networking services (SNSs) have become widespread worldwide. In particular, Twitter is considered to be one of the most popular SNSs and is used on a daily basis for a variety of purposes, including the dissemination of opinions and communication.

In this context, social viewing, where people post live tweets while watching a TV program, is becoming increasingly popular. Live tweets are tweets posted while the poster is watching a TV program and include real-time reactions to the program, such as comments and opinions. By posting live tweets, SNS users can discuss the same programs with other users via Twitter, just as they normally do with their family and friends while watching TV programs [1]–[4].

Social viewing is not only fun for users who post live tweets but also for the users who only view the tweets rather than posting them. This paper focuses on live tweet searching after watching TV dramas, where viewers may want to know what others thought about a scene that left a strong impression on them or a scene that they have questions about. In such cases, they can look at the live tweets of other viewers of the scene

and relate with the viewers that have similar opinions or gain new knowledge by seeing tweets with a different perspective.

Viewing live tweets can allow viewers to review the content of the drama and enjoy their reactions to the program more deeply. However, many live tweets can be posted about TV programs, and it is necessary to search through them to find the live tweets for the desired scene. This paper proposes an interface for finding the live tweets of TV dramas [1]. The term “TV drama review search” refers to the search for actual tweets for a specific scene in order to look back on the content of a drama after the initial viewing.

In the conventional Twitter search function, live tweets can be retrieved using hashtags. Hashtags are tags that begin with a “#” and classify posts by a specific topic. Many live tweets are tagged with the title of the program or its abbreviation, and hence people can search by hashtag to see live tweets posted by other people. However, whereas this search function is ideal for viewing real-time tweets about a scene being broadcast, it poses some problems when viewing past tweets, such as when the user wants to view tweets about an earlier scene after watching a TV program or when the user wants to record a TV program after it has aired. There are three problems users encounter when browsing past live tweets.

- 1) The number of live tweets of TV programs is huge, and it takes a lot of effort to check each result obtained by the tweet search function and to go back to the tweets of the scene that the user is interested in.
- 2) The contents of live tweets are often very brief. It can be difficult to tell from the tweet alone which scene the comment is about.
- 3) Users can also narrow down the tweets by searching for keywords that are characteristic of the target scene along with the title of the program or abbreviated hashtag, but only the tweets that match these keywords will be displayed, and hence if the keywords are ambiguous, users will not be able to obtain the tweets they want.

In this paper, we propose a tweet search interface that enables the efficient review of TV dramas to overcome these problems. This interface helps users efficiently discover live tweets of interest. In this system, the user inputs a tweet of interest, and the number of live tweets related to that keyword in the drama are visualized as a graph. Using this graph, the user can efficiently discover the time interval related to the interest and easily access the tweets of the scene the user is

interested in.

The contributions of this paper are as follows:

- 1) we propose a user interface suitable for viewing the opinions of TV dramas posted on Twitter, and
- 2) we demonstrated the effectiveness of the proposed interface through user experiments.

This paper is organized as follows. Section 2 positions this research with respect to related studies. Section 3 gives an overview of the system proposed in this paper, and Section 4 describes the details of the proposed method. Section 5 shows the results of the experiments, and Section 6 presents a summary and future work.

II. RELATED WORK

There have been many studies about TV programs and live-tweeting on Twitter.

Nakazawa et al. [5] proposed a method for detecting important scenes from tweets related to TV programs, estimating the main characters and events in each scene, and assigning them with labels representing the scenes for the efficient viewing of recorded TV programs. Lanagan et al. [6] proposed a method for identifying events of interest within the video of live sports broadcasts. Ushijima et al. [7] focused on social viewing of TV dramas using Twitter and characterized TV dramas by “development pattern” by extracting the features of scenes in the drama’s chronological order using live tweets posted during the drama broadcast. Vranić et al. [8] proposed a method for extracting drama patterns from viewer responses about TV dramas posted on social networking sites.

In these studies, the features of the scene and the sentiment of the tweets were extracted and visualized based on the live tweets. In this study, we further extract the engagement for keywords entered by the user and present them in chronological order.

Tsukuda et al. [9] proposed a method for estimating the scenes in which characters in a video attract the attention of viewers and estimating the degree of activity of each character in each scene using comments posted on Nico Nico Douga. In this method, the attention-grabbing scenes are estimated by focusing only on the characters. In contrast, in the method proposed in this paper, the attention-grabbing scenes are estimated not only using the names of characters but also using the keywords entered by users.

III. PROPOSED METHOD

The purpose of this study is to develop an interface that allows users to find live tweets related to the desired scene with simple operations in order to efficiently review TV drama programs.

A. System overview

Live tweets of TV drama programs represent the real-time responses of users who are watching the drama in question. Live tweets are considered to strongly reflect the content of the scene being broadcast at that time [7]. We assume that the scenes associated with the keywords specified by the user

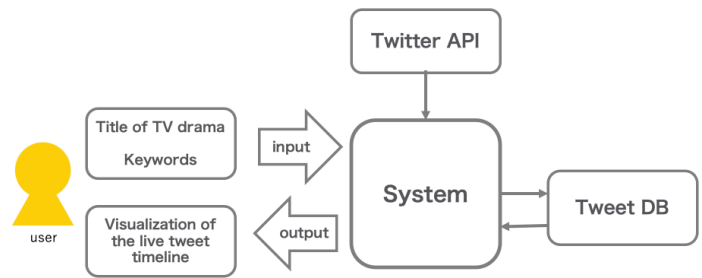


Figure 1. Overview of the proposed system.

have many live tweets with high similarity to the keywords. The relevance of the keyword to the scene is then estimated using the content of the live tweets associated with the scene. Specifically, the timeline consisting of live tweets posted during the drama broadcast time is divided into segments, and the relevance between segments and keywords is determined based on the similarity between the tweets and keywords in each segment. Then, by visualizing the transition of the relevance, users can easily find the segment they are interested in. When a user specifies a segment of interest, the user can then access the tweets contained in the segment.

Figure 1 shows an overview of the proposed system, and the procedure of the system is described as follows:

- 1) The system collects live tweets about TV drama programs using the Twitter application program interface (Twitter API). Specifically, tweets that include the title of the TV drama program hashtag posted during the broadcast time of the target TV drama program are collected and stored in the tweet database (tweet DB). Retweets and replies are excluded from the stored tweets.
- 2) The tweets of the TV drama program specified by the user are retrieved from the tweet DB, and the timeline of the collected tweets is divided into segments according to time in order to obtain the characteristics of the tweets over time.
- 3) Morphological analysis is performed on the tweets in the segment.
- 4) The tweets and keywords in the segment are vectorized.
- 5) The cosine similarities of the vectors are calculated. The similarity between each segment and the keyword is also calculated.
- 6) The similarity of each segment is visualized and presented to the user.

B. Modeling situations of TV dramas

The aim of the proposed method is to estimate and visualize the excitement related to keywords for each unit of time according to the progress of the TV drama program. We divide the timeline of collected live tweets into segments of a certain time interval. The set of tweets in the segmented time interval is called a situation segment, and each situation segment is considered to strongly reflect the characteristics of the scene broadcasted at that time.

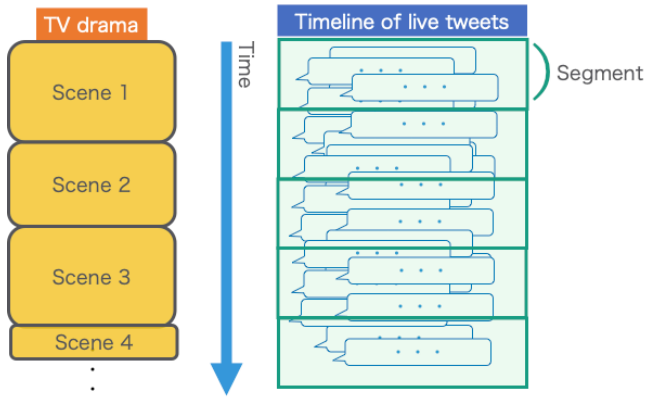


Figure 2. Relationship between scenes of a TV drama and situation segments in a timeline.

Figure 2 illustrates the relationship between the scenes of a TV drama and situation segments in a timeline.

We represent the timeline tl of live tweets as a series $tl = (tw_1, tw_2, \dots, tw_n)$ using tweets tw_i . By denoting the time of posting a tweet tw as $\text{time}(tw)$, any two tweets tw_i, tw_j in the timeline will satisfy $\text{time}(tw_i) < \text{time}(tw_j)$ if $i < j$.

This study introduces the concept of situation segmentation to describe the real-time content targeted by live tweets. A situation segment is a time interval in the targeted real-time content, which is defined as $s(tl, st, et)$. Here, tl represents the target timeline, st represents the start time of the segment, and et represents the end time.

In this study, we divide the targeted real-time content into situation segments of equal length (unit situation segments) using a time window and model the features as a unit. To generate a unit situation segment, we apply a time window of length m to the real-time content, move it by $m/2$ width, and allow the windows to overlap halfway so that we can also properly model the boundaries of the segment. When a unit situation segment is defined for the target real-time content, the state of the real-time content can be represented as a series of unit situation segments. Hereafter, unless otherwise specified, the term “situation segment” refers to a unit situation segment.

For each situation segment s , we consider the corresponding live tweet series $TW(s)$, which represents a subseries of the timeline targeted by the situation segment s .

C. Visualization

In this method, we provide a user interface that visualizes and displays the obtained similarity of each segment as a graph. The visualization approach is illustrated in Figure 3. The user first enters a keyword of interest q into the system. The system then calculates the similarity $\text{sim}(q, s)$ of the entered query keyword q and the situation segment s in the target timeline. A single situation segment is represented in a bar graph with one horizontal bar, where the length of the bar represents the similarity. By looking at the graph, the user can determine the time the scene related to the keyword was

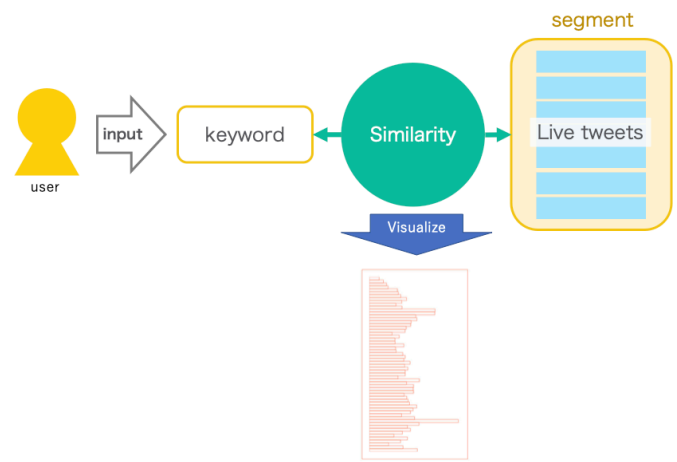


Figure 3. Visualization approach.

broadcasted, and by moving the mouse over the graph, the user can view the live tweets posted at that time. The right half of Figure 4 shows an example of timeline visualization.

To calculate the similarity $\text{sim}(q, s)$ between a keyword q and a situation segment s , several methods can be considered. In this paper, in Section IV, we propose three methods for computing the similarity and evaluate their performance in an evaluation experiment.

D. User interface

The proposed system provides an interface that enables users to view many tweets about scenes of interest using a visualization based on the similarity between keywords and situation segments. Figure 4 shows an example of the interface provided by the proposed system. A generated bar graph is shown on the left side of the interface. Users can click on any part of the graph, and tweets posted at the time represented by that location are displayed on the right side. The color of the background of each tweet indicates how well it matches the user’s query. The closer the background is to red, the more similar the tweet is to the user’s query.

IV. COMPUTATIONAL METHODS FOR QUERIES AND SITUATION SEGMENTS

In the proposed system, the similarity between the user’s query and the situation segment is calculated and used for visualization. There are several possible methods to calculate this similarity. In this section, we propose three similarity calculation methods. The performance of each method is evaluated based on the experimental results presented in Section V.

A. W2VE method

We propose the W2VE method as the first similarity calculation method. This method is based on the Word2Vec [10], which is a word vectorization method that uses a neural network consisting of two layers for text processing. By learning the weights of the neural network using a corpus,

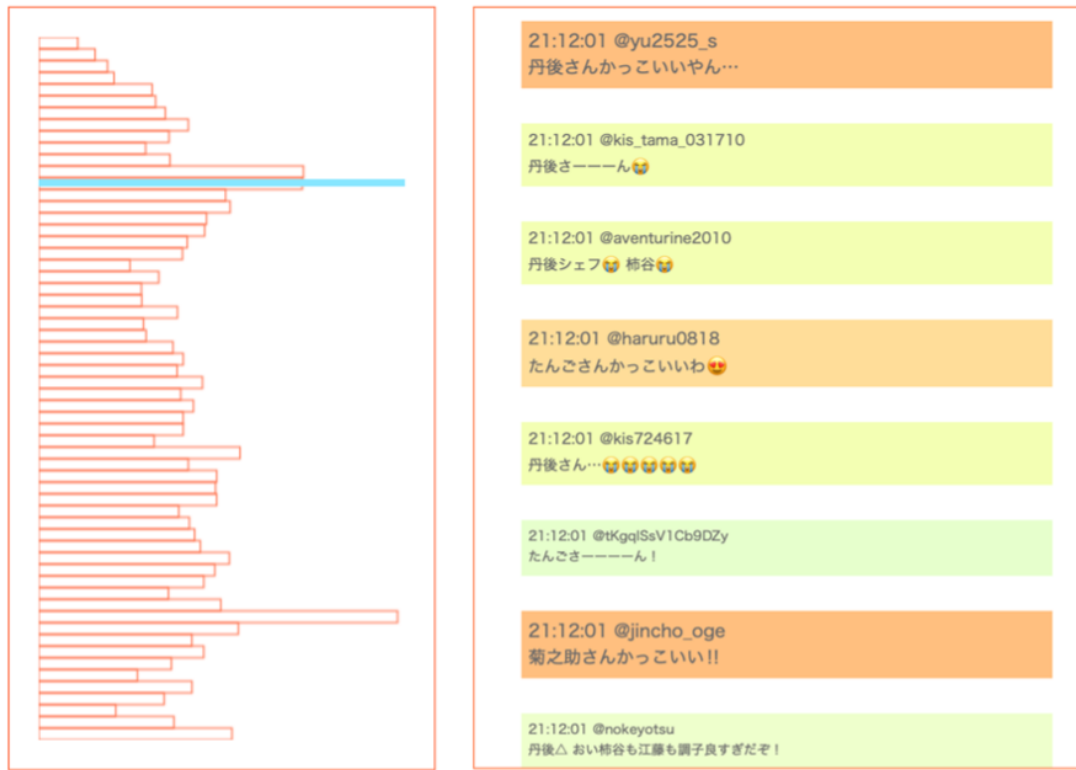


Figure 4. Screenshot of the user interface of the proposed system.

a vector representation of words can be obtained. To calculate the similarity between a situation segment and a keyword, Word2Vec is used to calculate the similarity between the query and each keyword in the segment.

In this method, the tweets and keywords in the segment are vectorized using the Word2Vec model learned by the above method, the cosine similarity with respect to the keywords is calculated for each tweet, and the average is used as the final similarity for the segment. The similarity of the W2VE method is defined as follows:

$$W2VE(q, s) = \frac{1}{|TW(s)|} \sum_{i \in TW(s)} \text{csim}(\mathbf{w2v}(q), \mathbf{w2v}(i)) \quad (1)$$

where q is the query keyword, s is the situation segment, $\mathbf{w2v}(q)$ is a function that vectorizes the query keyword q based on the Word2Vec method, and $\text{csim}(\mathbf{a}, \mathbf{b})$ represents the cosine similarity between vectors \mathbf{a} and \mathbf{b} .

B. W2VS method

We propose the W2VS method as the second similarity calculation method. The W2VS method is a calculation method that also uses the Word2Vec method. In the first method, the average of the cosine similarities of vectorized queries and tweets is obtained by Word2Vec. In contrast, in this method, the vector of the situation segment is obtained by vectorizing all the tweets in the target situation segment using Word2Vec and calculating their average. Then, the cosine similarity

between the query vector and the vector of the situation segments is calculated. The W2VS method is formally defined as follows:

$$W2VS(q, s) = \text{csim}(\mathbf{w2v}(q), \mathbf{avg}(s)) \quad (2)$$

$$\mathbf{avg}(s) = \frac{1}{|TW(s)|} \sum_{i \in TW(s)} \mathbf{w2v}(i) \quad (3)$$

C. TFIDF method

The third similarity calculation method proposed in this paper is the TFIDF method. The TF-IDF [11], [12] method calculates the importance of a word in a document based on the frequency of occurrence (TF) of the word in the target document and the inverse document frequency (IDF) of the word. The TF-IDF method has been proposed in the field of information retrieval and is currently used for various purposes. In this paper, we propose a method that calculates the importance of a word in each situation segment using situation segments instead of documents in the general TF-IDF method.

The TF value of t in s is defined by the following equation, where $\text{freq}(t, S)$ is the frequency of occurrence of a word t in the target situation segment s .

$$\text{tf}(t, s) = \frac{\text{freq}(t, s)}{\sum_i \text{freq}(i, s)} \quad (4)$$

$$\text{idf}(t) = \log \left(\frac{|S|}{1 + |\{s | \text{freq}(t, s) \geq 1, s \in S\}|} \right) \quad (5)$$

By multiplying the TF and IDF values calculated above, the importance weight(t, s) of a word t in a situation segment s is defined by the following.

$$\text{weight}(t, s) = \text{tf}(t, s) \text{idf}(t) \quad (6)$$

Using the above weights, we define the similarity TFIDF(q, s) between query keyword q and situation segment s as follows:

$$\text{TFIDF}(q, s) = \text{csim}(\mathbf{h}(q), \mathbf{w}(s)) \quad (7)$$

where $\mathbf{h}(q)$ represents the one-hot vector of the keyword q , and $\mathbf{w}(s)$ represents the feature vector of the situation segment s , which is constructed using the word weights weight(t, s).

V. EVALUATION

This section presents the experiments we conducted to evaluate the effectiveness of the proposed method and its results. We evaluated our method with respect to the following two issues:

- 1) the performance of the three similarity calculation methods proposed in this paper, and
- 2) the usability of the proposed system.

A. Dataset, preprocessing, and prototype

The dataset used for the evaluation consists of live tweets about TV dramas collected using the Twitter API. We collected live tweets for 16 TV dramas (125 episodes) broadcasted on Japanese TV stations from July to September 2019, and a further 15 TV dramas (111 episodes) broadcasted from October to December 2019. The hashtags of the respective TV drama titles were used to collect the live tweets for the TV dramas during the broadcast times of the target dramas. Retweets and replies were excluded from these data. These tweets were written in Japanese.

Figure 5 shows an overview of the preprocessing required for this dataset. From the tweets included in the dataset, the hashtags and URLs of TV drama titles used in the collection were removed from the text because they could act as noise when obtaining the characteristics of the tweets. The other hashtags were not excluded because they can contain information such as the names of the actors in the current scene and thus become features of the scene.

All the tweets in the dataset were split into morphemes by MeCab [13], a major Japanese morphological analysis engine. For the MeCab dictionary, we used the mecab-ipadic-NEologd dictionary [14], which covers a wide range of Eigen expressions, collapsed notations commonly used on the web, and new words. Of the segmented morphemes, only nouns, verbs, adjectives, and adverbs were used, and for conjugated words, the original form of the word was used.

We implemented a prototype of the proposed system for the experiments. This system runs as a web application. PHP

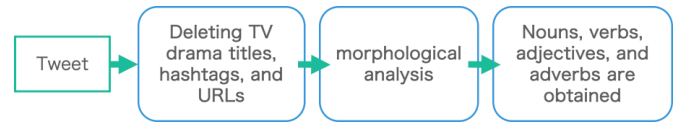


Figure 5. Overview of the preprocessing.

and JavaScript were used for its development, Apache was used as the webserver, and MySQL was used as the database management system. The Gensim library [15] was used to calculate Word2Vec, and the Twitter dataset described above was used as the corpus for training the Word2Vec model.

B. Performance comparison of the similarity calculation methods

1) *Experimental method:* In this paper, we proposed the W2VE, W2VS, and TFIDF methods to determine the similarity between the query keywords given by the user and the situation segments. We compared the performance of these three methods through experiments. For live tweets related to the target TV dramas, we determined the query keywords related to those TV dramas and calculated the similarity between each keyword and the situation segment. The number of target TV dramas was three. Ten query keywords were selected from each of the adjectives and nouns frequently found in the live tweets of each drama and used in the experiment. To create the ground-truth data, subjects were asked to read the tweets included in the target situation segment and give them a score from 0 to 10 on how similar their contents were to the keywords. The ground-truth data and the similarities derived by each method were normalized so that the maximum value was 1, and the error was calculated. The mean average error (MAE) was used as the measure of error.

2) *Results:* As an example, the results of the experiment in which the TFIDF method was used to calculate the similarity for a TV drama are shown in Figure 6. In this figure, the vertical axis represents the similarity and the horizontal axis represents the elapsed time after the start of the drama. The red line represents the calculated similarity, the green line represents the ground truth, and the blue dashed line represents the error.

The MAE values for each method are shown in Table I and the distribution of MAE for each keyword is shown in Figure 7. These results reveal that the TFIDF method yields the lowest MAE. We also analyzed whether there is a dominant difference in the MAE of each method using t-test. As a result, there was a significant difference between the W2VE and TFIDF results and between the W2VS and TFIDF results, whereas there was no significant difference between the W2VE and W2VS results. This indicates that TFIDF obtained the best performance.

C. Usability evaluation

1) *Experimental method:* To evaluate the effectiveness of the proposed method, we asked 20 male and 20 female users in their 20s to use the interface of the proposed method

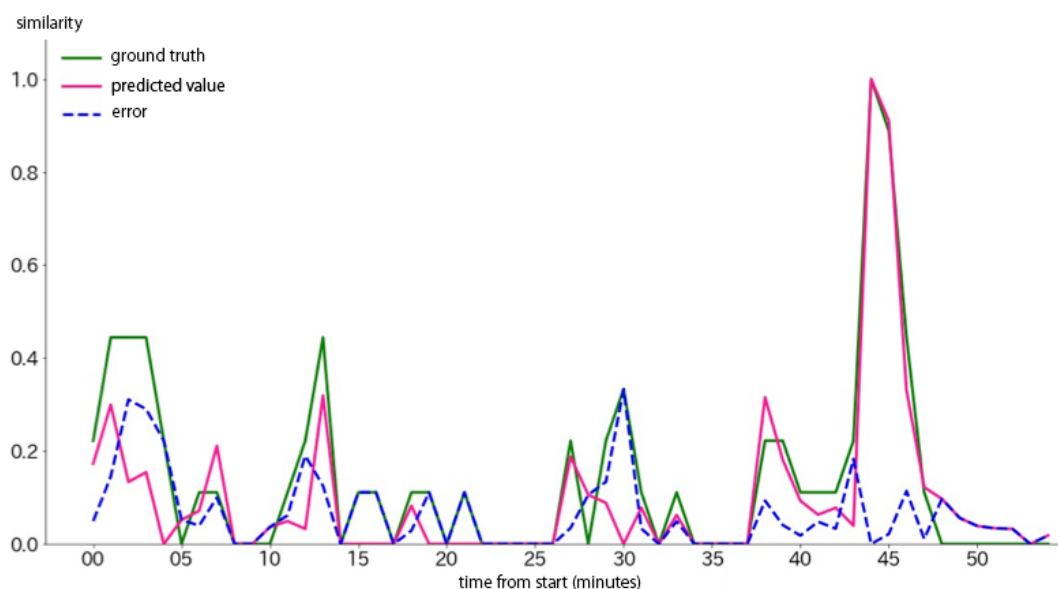


Figure 6. Example of timeline visualization.

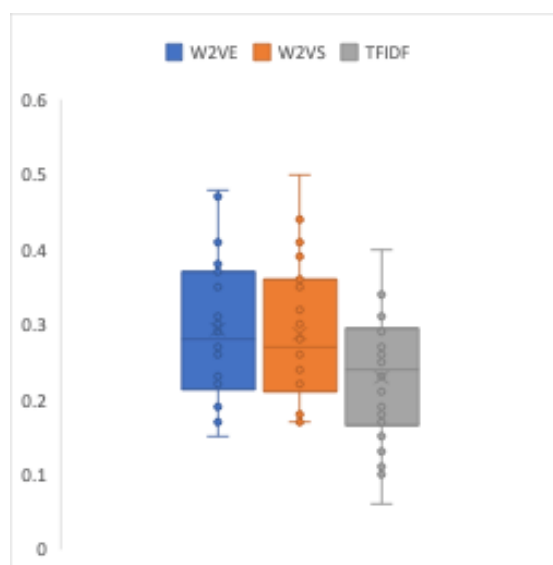


Figure 7. MAEs for the three methods.

TABLE I
AVERAGE OF MAE.

W2VE	W2VS	TFIDF
0.293	0.287	0.229

(developed using the TFIDF method) and to answer a questionnaire. The subjects were asked to enter a number of keywords for their favorite dramas, view live tweets, and answer the questionnaire. Each subject responded to each question on a five-point Likert scale from 1 to 5. 1 represents strong disagreement, and 5 represents strong agreement. The

TABLE II
RESULTS OF THE USABILITY QUESTIONNAIRE FOR THE PROPOSED INTERFACE.

Question	Average Score
Q1	4.18
Q2	4.36
Q3	3.81
Q4	4.09
Q5	4.18
Q6	4.00

following are the questions in the questionnaire.

- Q1: Were the graphs presented by the proposed interface able to represent the characteristics of the TV drama scenes?
- Q2: Compared to browsing live tweets on a typical Twitter search interface, did you find it easier to find live tweets for scenes you were interested in using the proposed interface?
- Q3: Was the proposed interface easy to use?
- Q4: Was the visual appearance of the proposed interface good?
- Q5: Is the proposed interface useful for looking back on TV dramas?
- Q6: Would you like to use the proposed interface in the future?

2) *Results:* The results of the above questionnaire administered to the subjects are shown in Table II. This table shows the averages of the users' responses to each question.

For questions Q1, Q2, Q4, Q5, and Q6, the mean values were 4 or higher, which indicates that the proposed interface is an effective way to review TV dramas. The score for question Q3 is 3.81, which indicates that usability needs to be improved

in the future.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed an interface that allows users to efficiently view live tweets for the desired scene in order to review TV dramas. The interface divides the live tweets posted during the broadcast of a TV drama into situation segments by time interval and calculates the similarity between the tweets and keywords in each segment to visualize the changes in the excitement related to the keywords of the drama. In this paper, we proposed the W2VE, W2VS, and TFIDF methods to calculate the similarity between keywords and situation segments for visualization.

From the results of evaluation experiments, we found that TFIDF is the most suitable method for this task. In addition, the results of a usability survey conducted by subjects using the prototype system showed that the proposed interface was able to capture the characteristics of TV drama scenes and was an effective approach for looking back on TV dramas.

The following is a list of issues to be tackled in the future.

- 1) Sometimes, a time lag exists between when a user posts a tweet and when it appears on the timeline. It will be necessary to develop a function to compensate for the user's posting time.
- 2) Some live tweets may contain tweets that are not directly related to the TV drama scene; we need to develop a function to filter out tweets that are not related to the TV drama content.
- 3) The proposed interface may be applicable to domains other than TV drama reviews. We plan to extend the interface so that it can be applied to other purposes, such as viewing public opinion on the news.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 19H04219.

REFERENCES

- [1] T. Ushima and H. Nagai, "Interface for looking back at tv dramas on twitter," in *WEB 2021, The Ninth International Conference on Building and Exploring Web Based Environments*. IARIA, 2021, pp. 1–2.
- [2] *EuroITV '11: Proceedings of the 9th European Conference on Interactive TV and Video*. New York, NY, USA: Association for Computing Machinery, 2011.
- [3] *EuroITV '12: Proceedings of the 10th European Conference on Interactive TV and Video*. New York, NY, USA: Association for Computing Machinery, 2012.
- [4] M. Doughty, D. Rowland, and S. Lawson, "Who is on your sofa? tv audience communities and second screening social networks," ser. *EuroITV '12*. New York, NY, USA: Association for Computing Machinery, 2012, p. 79–86. [Online]. Available: <https://doi.org/10.1145/2325616.2325635>
- [5] M. Nakazawa, M. Erdmann, K. Hoashi, and C. Ono, "Social indexing of tv programs: Detection and labeling of significant tv scenes by twitter analysis," in *2012 26th International Conference on Advanced Information Networking and Applications Workshops*, 2012, pp. 141–146.
- [6] J. Lanagan and A. F. Smeaton, "Using twitter to detect and tag important events in sports media," in *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, L. A. Adamic, R. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011. [Online]. Available: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2821>

- [7] D. Minami, M. Ushijima, and T. Ushima, "How do viewers react to drama?: Extraction of scene features of dramas from live commentary tweets," in *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, IMCOM 2018, Langkawi, Malaysia, January 05-07, 2018*. ACM, 2018, pp. 87:1–87:4. [Online]. Available: <https://doi.org/10.1145/3164541.3164616>
- [8] V. Vranić and A. Vranić, "Drama patterns: Extracting and reusing the essence of drama," in *Proceedings of the 24th European Conference on Pattern Languages of Programs*, ser. EuroPLop '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: <https://doi.org/10.1145/3361149.3361153>
- [9] K. Tsukuda, M. Hamasaki, and M. Goto, "Smartvideoranking: Video search by mining emotions from time-synchronized comments," in *IEEE International Conference on Data Mining Workshops, ICDM Workshops 2016, December 12-15, 2016, Barcelona, Spain*, C. Domeniconi, F. Gullo, F. Bonchi, J. Domingo-Ferrer, R. Baeza-Yates, Z. Zhou, and X. Wu, Eds. IEEE Computer Society, 2016, pp. 960–969. [Online]. Available: <https://doi.org/10.1109/ICDMW.2016.0140>
- [10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 3111–3119. [Online]. Available: <https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>
- [11] K. Church and W. Gale, "Inverse document frequency (idf): A measure of deviations from poisson," in *Natural language processing using very large corpora*. Springer, 1999, pp. 283–295.
- [12] T. Roelleke and J. Wang, "Tf-idf uncovered: A study of theories and probabilities," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 435–442. [Online]. Available: <https://doi.org/10.1145/1390334.1390409>
- [13] "MeCab: Yet Another Part-of-Speech and Morphological Analyzer," <https://taku910.github.io/mecab/>, [accessed: 2021-12-09].
- [14] "mecab-ipadic-NEologd : Neologism dictionary for MeCab," <https://github.com/neologd/mecab-ipadic-neologd>, [accessed: 2021-12-09].
- [15] "Gensim: Topic modelling for humans," <https://radimrehurek.com/gensim/>, [accessed: 2021-12-09].