WORKSHOP PROCEEDINGS OF THE 13TH INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA

EDITED BY: J. Pfeffer, R. West, J. Diesner, S. Wu, M. M. Malik, K. Mayer, K. Joseph, S. Gaito, C. Müller-Birn, Y. Mejova, A. Sala, R. Interdonato, D. Paolotti, H. Lamba, J. Seering, A. Tagarelli, S. R. Kairam, K. Kalimeri, A. Hannak, R. Chunara, E. Chandrasekharan, D. Alburez-Gutierrez, S. Gil-Clavel, E. Zagheni and S. Chancellor

PUBLISHED IN: Frontiers in Big Data

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88963-911-3
DOI 10.3389/978-2-88963-911-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

WORKSHOP PROCEEDINGS OF THE 13TH INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA

Citation: Pfeffer, J., West, R., Diesner, J., Wu, S., Malik, M. M., Mayer, K., Joseph, K., Gaito, S., Müller-Birn, C., Mejova, Y., Sala, A., Interdonato, R., Paolotti, D., Lamba, H., Seering, J., Tagarelli, A., Kairam, S. R., Kalimeri, K., Hannak, A., Chunara, R., Chandrasekharan, E., Alburez-Gutierrez, D., Gil-Clavel, S., Zagheni, E., Chancellor, S., eds. (2021). Workshop Proceedings of the 13th International AAAI Conference on Web and Social Media. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88963-911-3

Table of Contents

frontiers in Big Data

ORIGINAL RESEARCH
published: 31 May 2019
doi: 10.3389/fdata.2019.00007

You Can't See Me: Anonymizing Graphs Using the Szemerédi Regularity Lemma

Daniele Foffano[1], *Luca Rossi*[2*] *and Andrea Torsello*[1]

[1] *Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari Venezia, Venezia, Italy,* [2] *Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China*

Complex networks gathered from our online interactions provide a rich source of information that can be used to try to model and predict our behavior. While this has very tangible benefits that we have all grown accustomed to, there is a concrete privacy risk in sharing potentially sensitive data about ourselves and the people we interact with, especially when this data is publicly available online and unprotected from malicious attacks. *k*-anonymity is a technique aimed at reducing this risk by obfuscating the topological information of a graph that can be used to infer the nodes' identity. In this paper we propose a novel algorithm to enforce *k*-anonymity based on a well-known result in extremal graph theory, the Szemerédi regularity lemma. Given a graph, we start by computing a regular partition of its nodes. The Szemerédi regularity lemma ensures that such a partition exists and that the edges between the sets of nodes behave almost randomly. With this partition, we anonymize the graph by randomizing the edges within each set, obtaining a graph that is structurally similar to the original one yet the nodes within each set are structurally indistinguishable. We test the proposed approach on real-world networks extracted from Facebook. Our experimental results show that the proposed approach is able to anonymize a graph while retaining most of its structural information.

Keywords: privacy, anonymity, social networks, graph, regularity lemma

OPEN ACCESS

Edited by:
Roberto Interdonato, Télédétection et Information Spatiale (TETIS), France

Reviewed by:
Ruggero Gaetano Pensa, University of Turin, Italy
Matteo Zignani, University of Milan, Italy

***Correspondence:**
Luca Rossi
rossil@sustech.edu.cn

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *25 March 2019*
Accepted: *13 May 2019*
Published: *31 May 2019*

Citation:
Foffano D, Rossi L and Torsello A (2019) You Can't See Me: Anonymizing Graphs Using the Szemerédi Regularity Lemma. Front. Big Data 2:7. doi: 10.3389/fdata.2019.00007

1. INTRODUCTION

The beginning of the twenty-first century has been characterized by the rise of online social media and data-hungry artificial intelligence (AI). In this context, sophisticated machine learning algorithms feed off massive amounts of data produced by our digital personas to perfect the way they model and predict our behavior, both online and offline. However, the comforts of an increasingly AI-assisted life are overshadowed by the threat it poses to our privacy and freedom (Fung et al., 2010; Rossi and Musolesi, 2014; Rossi et al., 2015b; Qian et al., 2016). At the same time, the digital traces we produce, particularly interactions between users in an online social network, are often abstracted using a graph representation and made available in the form of public datasets, as they offer a unique opportunity for researchers to study real-world complex networks of interactions (Kwak et al., 2010; Chorley et al., 2016).

A common practice to protect the identity of the users whose interactions are captured by the graph is that of stripping the nodes of sensitive information (e.g., the users names), generating a random identifier to label the graph nodes. However, it has been shown that this does not guarantee that the user's privacy is preserved (Backstrom et al., 2007). Indeed, it is possible to disclose the identity of an individual participating in the network with minimal external background information. One common example is that of a user for which the number of connections in the network is known (i.e., the number of friends on Facebook) and this number happens to be unique for that individual. In other words, this piece of information alone would be sufficient to identify that user among the rest of the nodes. Most importantly, once the identity is revealed, other potentially sensitive pieces of information can be inferred. For instance, the individual may turn out to belong to a group of nodes labeled with a certain sensitive attribute, e.g., health condition.

For these reasons, the problem of anonymizing graph data is becoming an increasingly studied one (Hay et al., 2008; Liu and Terzi, 2008; Rossi et al., 2015a; Qian et al., 2016). A common anonymity model is k-anonymity, which aims to ensure that each node in a network is structurally indistinguishable from at least other k nodes. Different works have focused on different definitions of "structurally indistinguishable." Liu and Terzi (2008) considered the case of k-degree anonymous graphs, where k-degree anonymity guarantees that each node of the graph shares the same degree of at least k other nodes. Successive works attempted to reduce the total running time of Liu and Terzi (2008) to make it feasible to scale up to large networks (Hay et al., 2008). Rossi et al. (2015a), on the other hand, extended the concept of k-degree anonymity to multi-layer and time-varying graphs. Other researchers considered different structural distinguishability criteria where the attacker has increasing levels of information available to deanomymize the nodes (Hay et al., 2008; Cheng et al., 2010; Zhou and Pei, 2011), however the main issue with these approaches lies in the need to add increasing amounts of noise as increasingly complex structural information needs to be obfuscated. More recently Rousseau et al. (2018) considered the problem of anonymizing a graph maximizing the amount of preserved community information. Finally, Qian et al. (2016) and Ma et al. (2018) looked at the complementary problem of deanonymizing a graph in the case where the attacker has access to richer features as well as structural information.

While most of the previous k-anonymity approaches assume that the attacker has access only to a certain level of structural information (from the degree of a node, to its immediate neighborhood or even the whole graph), in this paper we propose a method that creates k-anonymous groups of nodes where no degree of structural information can help to break the anonymity guarantee. Our approach is based on the Szemerédi regularity lemma (Diestel, 2012), a well-known result of extremal graph theory. The Szemerédi regularity lemma has been successfully applied to several problems, from graph theory (Komlós and Simonovits, 1996) to computer vision and pattern recognition (Sperotto and Pelillo, 2007; Pelillo et al., 2017). The lemma roughly states that every sufficiently large and dense graph[1] can be approximated by the union of random-like bipartite graphs called regular pairs. Our observation is that the groups of graph nodes that form these regular pairs can be anonymized by rewiring the intra-group edges according to an Erdös-Rényi process (Erdős, 1960). Thanks to the theoretical guarantees of the Szemerédi regularity lemma, this has minimal effect on the overall graph structure and, together with the random-like behavior of the inter-group connections, ensures that the each group is anonymous.

The reminder of the paper is organized as follows. We start by reviewing the key graph theoretical concepts underpinning our work in section 2. In section 3 we propose our anonymization method based on the Szemerédi regularity lemma and in section 4 we evaluate it on three different networks abstracted from Facebook. Finally, section 5 concludes the paper.

2. SZEMERÉDI REGULARITY LEMMA

Let $G = (V, E)$ be an undirected graph with no self-loops, where V is the set of nodes and E is the set of edges. If X and Y are disjoint subsets of V, the *edge density* of this pair (X, Y) is defined as $d(X, Y) = \frac{|E(X,Y)|}{|X||Y|}$, where $E(X, Y)$ is the set of edges connecting nodes in X to nodes in Y. The edge density satisfies $0 \leq d(X, Y) \leq 1$.

Given a positive real $\varepsilon > 0$, a pair of node sets X and Y is called *ε-regular* if for all subsets $A \subseteq X$ and $B \subseteq Y$ satisfying $|A| \geq \varepsilon|X|$ and $|B| \geq \varepsilon|Y|$ we have $|d(X, Y) - d(A, B)| \leq \varepsilon$. Stated otherwise, the distribution of the edges between an ε-regular pair is almost uniform, i.e., the graph over $X \cup Y$ behaves like a random bipartite graph.

Let the node set V be divided into a partition $\mathcal{P}$ of l sets $V_1, \cdots, V_l$. $\mathcal{P}$ is an *ε-regular partition* if: (1) $||V_i| - |V_j|| \leq 1$, for $1 \leq i < j \leq l$ and (2) all except at most εl^2 pairs (V_i, V_j) $(1 \leq i < j \leq l)$, are ε-regular. With these definitions in hand, we can finally state the following.

Lemma 2.1 (Szemerédi regularity lemma)**.** *For every positive real $\varepsilon > 0$ and every positive integer m, there exist positive integers $N = N(\varepsilon, m)$ and $M = M(\varepsilon, m)$ such that, if $G = (V, E)$ is a graph with $|V| \geq N$ nodes, there is an ε-regular partition of V into l groups with sizes that differ at most by 1, where $m \leq l \leq M$.*

In other words, the Szemerédi regularity lemma states that a graph can be seen as a collection of groups of nodes such that the edges between these groups are almost uniformly distributed. More generally, as stated by Komlós and Simonovits (1996), the regularity lemma states that every graph can be approximated by generalized random graphs. Note that the lemma also states that there may be a number of ε-irregular pairs that do not behave like random bipartite graphs. However, for a sufficiently small ε, the number of such pairs will be low (i.e., smaller than εl^2).

Given a graph G and an ε-regular partition of its nodes, a reduced graph can be constructed by replacing each pair of ε-regular groups with two nodes connected by an edge. As shown

[1] Note that the lemma has been extended to sparse graphs as well (Gerke and Steger, 2005).

by the Key lemma (Komlós and Simonovits, 1996), the reduced graph inherits many of the fundamental structural properties of the original graph, to the point that the graph obtained by simply replacing each pair of connected nodes of the reduced graph with a complete bipartite graph over $2t$ nodes yields a new graph that can be used as a surrogate of the original one, where $t \geq 1$ is an integer.

Recall that the aim of this paper is to anonymize a graph $G = (V, E)$ by grouping V into sets of k-anonymous nodes. The Szemerédi regularity lemma states that the node set of each graph can be rearranged to reveal a random-like structure, where pairs of groups of k nodes are connected in an almost uniform (in other words, random) way. That is, for the purpose of graph de-anonymization, the edge information between the groups of nodes is unusable. Unfortunately, the intra-group connections can be still exploited to deanonymize the nodes. However, the Szemerédi regularity lemma and the fact that the reduced graph (where the intra-group connections are lost) preserves the fundamental structural properties of the original graph imply that these intra-group connections are small in number and structurally negligible.

3. ANONYMIZATION FRAMEWORK

In the previous section we introduced the Szemerédi regularity lemma and we showed how this can be seen as a first step toward obtaining a k-anonymous graph. To achieve full k-anonymity, however, we need to obfuscate the structural information contained in the intra-group connections of the ε-regular partition. Our solution involves rewiring these connections using the Erdös-Rényi model (Erdős, 1960), effectively replacing each subgraph (i.e., each group of the ε-regular partition) with an Erdös-Rényi graph over the same set of nodes. Crucially, for each subgraph, we set the parameter p, which governs the probability of adding/deleting an edge, equal to the density of the original subgraph. More specifically, our approach follows three steps: (1) we first find a regular partition using the regularity lemma; (2) then, we randomize the groups' intra-connections; and (3) finally, we randomize the edges connecting irregular pairs.

In the **first step** we apply the algorithm implemented by Fiorucci et al. (2019)[2]. This extends the previous algorithm of Fiorucci et al. (2017) by proposing a novel heuristic procedure where the node set is first partitioned into two groups of nodes and then these are recursively split into smaller groups until a desired cardinality is met and certain conditions that measure quality of the ε-regularity of the partition are satisfied (Pelillo et al., 2017). In particular Fiorucci et al. propose two different heuristics to split the groups, one called *degree based*, which groups together nodes with similar degrees (Fiorucci et al., 2017), and a second one called *indeg guided*, which splits a sparse (dense) partition into two sparse (dense) partitions. Note that using this method we can only get a number of ε-regular groups which is a power of 2.

[2]Code available at: https://github.com/MarcoFiorucci/graph-summarization-using-regular-partitions.

The **second step** involves randomly rewiring the connections within each group of vertices. To this end, we add or delete an edge with a probability p equal to the density of the subgraph H spanned by the group of nodes we are trying to anonymize. Note that we only change the internal connections of H, so we are not altering the ε-regularity relations. The resulting subgraph H' will have the same density of H, however its structural information will not be of any use when trying to deanonymize its nodes.

Recall that each ε-regular partition allows up to εl^2 irregular pairs, where l is the number of sets of the ε-regular partition. So far we ensured that the connections within and between ε-regular pairs are anonymous, however we have not yet dealt with irregular pairs. The **third step** addresses this and requires rewiring the connections between groups forming an ε-irregular pair. Let (V_i, V_j) be one such pair, with total number of nodes n. Consider the bipartite subgraph $H = (V_i \cup V_j, E_{ij})$ where we only consider the set of edges E_{ij} connecting nodes in V_i with nodes in V_j. In order to render the structural information contained in these edges unusable for deanonymization purposes, we randomly rewire each pair of nodes (u, v), with $u \in V_i$ and $v \in V_j$, by adding/deleting an edge to E_{ij} with probability p equal to $|E_{ij}|/(V_i \times V_j)$.

In this framework ε can be interpreted as a measure of the error made by the Szemerédi regularity lemma approximation, i.e., the smaller ε the better the anonymized graph approximates the original graph. In fact, the amount of structural information preserved is inversely proportional to the number of edges we need to rewire. The Szemerédi regularity lemma allows us to safely rewire intra-group connections, knowing that these are small in number and structurally negligible. So the key to preserving the structural information of the original graph is to minimize the number of ε-irregular pairs. This becomes particularly relevant when anonymizing real-world complex networks, which often display a scale-free structure (Barabási and Albert, 1999). In these networks a small number of nodes (i.e., hubs) has a very large degree. If an irregular pair contains a hub we will end up rewiring a large number of edges, potentially compromising the structural information for the sake of anonymity. Therefore, minimizing the number of ε-irregular pairs is of fundamental importance. Also, recall that the method of Fiorucci et al. is based on heuristics, and in general different runs of their algorithm can result in different ε-regular partitions. For this reason, we repeat the computation of the ε-regular partition `max_iter` times and we choose the partition with the minimum ε and number of ε-irregular pairs. Note that each iteration of the algorithm of Fiorucci et al. has computational cost $O(n^{2.376})$, and this cost dominates in the overall anonymization complexity.

4. EXPERIMENTAL RESULTS

We test the proposed method on three real-world networks abstracted from Facebook. Note that all the graphs are sparse, as shown in **Table 1**. *Facebook Combined* represents circles (or friend lists) from Facebook. It was introduced for the first time by Mcauley and Leskovec in Leskovec and Mcauley (2012). The

two remaining networks, *Tv Shows* and *Politicians* describe blue verified pages of different kinds, where edges represent mutual likes among them (Rozemberczki et al., 2018).

With these graphs in hand, we compute their anonymized versions and we measure the amount of structural information lost with respect to the original graphs. In particular, we track the changes in number of edges, degree distribution, average clustering coefficient (Watts and Strogatz, 1998), and page rank vector (Page et al., 1999). We compute these changes for different levels of k-anonymity, which in turn correspond to different choices of the partition cardinality l. Recall in fact that k and l are related by the fact that in a graph with n nodes an ε-regular partition groups the vertices into l sets of cardinality $k \approx \frac{n}{l}$.

Note also that larger values of l also imply larger values of εl^2, the maximum number of ε-irregular pairs we can find in the network. Irregular pairs force us to randomly rewire connections that are not guaranteed to be structurally negligible by the Szemerédi regularity lemma (like the intra-group connections), so in general for large values of l more effort has to go into finding an ε-regular partition with minimum value of ε (in these experiments we vary ε from 0.01 to 0.2, with steps of 0.025). This is also the reason why we were only able to compute the ε-regular partitions for a small range of values of l. In fact, for some combinations of dataset and l, the algorithm of Fiorucci et al. was unable to find an optimal partition within `max_iter` $= 100$ iterations. In our experiments, the runtime to compute an ε-regular partition varies between approximately 10 and 80 s, on a machine with an 8-core 3.6 GHz CPU and 16GB of RAM.

We start by comparing the degree distributions of the original graphs and the anonymized ones, using both the *degree based* and the *indeg guided* heuristics. **Figure 1** shows the log-log plots of the results. Note that larger values of l tend to correspond to more accurate approximations of the original degree distribution. This is confirmed by looking at the Jensen-Shannon (JS) divergence Lin (1991) between the degree distributions, which for the *degree guided* heuristic and the *Politicians* dataset goes from 0.062 (with $l = 4$) to 0.011

TABLE 1 | Summary of the main structural characteristics of the original graphs.

Dataset	Nodes	Density	Edges	Avg. clustering coefficient
Facebook Combined	4,039	0.011	88,234	0.606
Politicians	3,892	0.002	41,729	0.385
Tv shows	5,908	0.002	17,262	0.374

TABLE 2 | Average variation in the number of edges (average clustering coefficient) between the original graph G and the anonymized graph $\tilde{G}$, calculated as $|s_G - s_{\tilde{G}}|/s_G$, where s_G and $s_{\tilde{G}}$ are the statistics considered.

Dataset	$l = 4$	$l = 8$	$l = 16$	$l = 32$	$l = 64$
Facebook Combined	0.0012 (0.7162)	0.0012 (0.6310)	0.0010 (0.5696)	0.0010 (0.5302)	0.0010 (0.4822)
Politicians	0.0021 (0.6983)	0.0020 (0.6415)	0.0015 (0.5261)	0.09 (0.2395)	n.a.
Tv shows	0.0034 (0.6553)	0.0036 (0.5064)	0.0013 (0.3158)	n.a.	n.a.

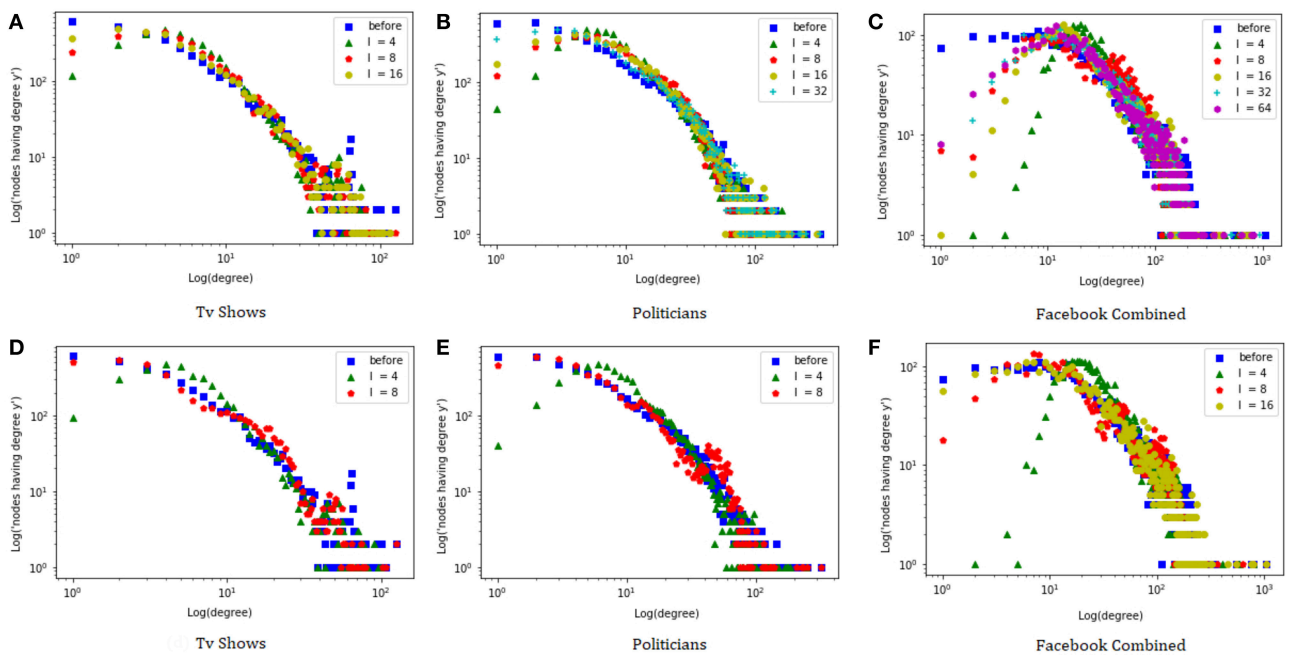

FIGURE 1 | Degree distribution of the graphs with *degree based* **(A–C)** and *indeg guided* **(D–F)** heuristics.

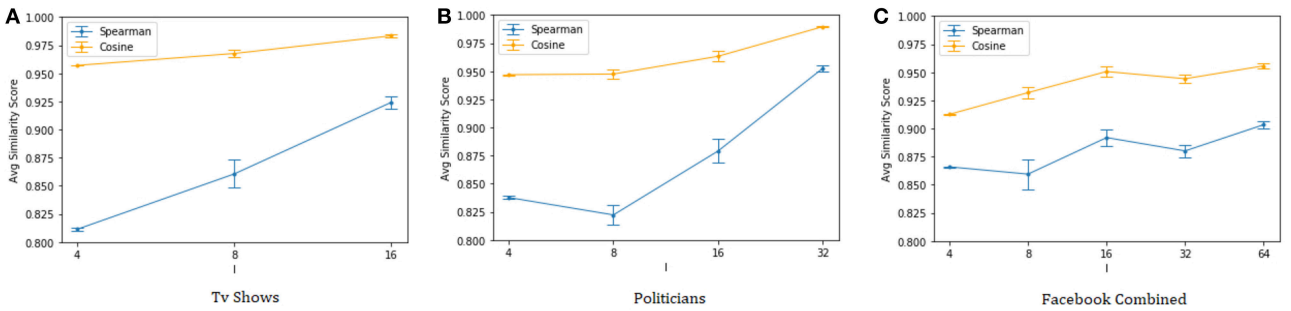

FIGURE 2 | Cosine similarity and Spearman's correlation of the page rank vectors (*indeg guided* heuristic). **(A)** Tv shows, **(B)** Politicians, and **(C)** Facebook Combined.

(with $l = 32$)[3]. Interestingly, the *indeg guided* heuristic seems to yield the best approximations. This could be because the degree-based heuristic struggles to create groups of nodes with similar degree when there are hubs among them. Indeed, for the *indeg guided* heuristic the JS divergence goes from 0.066 (with $l = 4$) to 0.016 ($l = 8$), whereas for $l = 8$ the *degree guided* heuristic achieves a JS divergence of 0.034[4]. In the remainder of the experiments we focus only on the *indeg guided* heuristic.

Table 2 shows the variation in the number of edges and average clustering coefficient with respect to the original graph. More precisely, we report $|s_G - s_{\tilde{G}}|/s_G$, where s_G and $s_{\tilde{G}}$ are statistics computed on the original and anonymized graphs, respectively (averaged over 10 anonymizations). We first note that the number of edges of the graphs changes only very slightly. Indeed, when we alter the structure of a group of vertices we do it by adding/deleting edges with a probability equal to the original edge density of the group. This in turn has the effect of keeping the number of edges approximately the same, regardless of the size k of the anonymity sets.

We then check the effect of the anonymization on the average clustering coefficient of the graph. **Table 2** shows that these statistics change significantly. Recall that the average clustering coefficient is proportional to the number of triangles in a network (Watts and Strogatz, 1998), however the Erdös-Rényi rewiring used to anonymize the vertex groups and the ε-irregular pairs is likely to break these triangles. While the Szemerédi regularity lemma ensures that the vertex groups are sufficiently sparse that we can ignore their inner structure, this clearly does not hold for ε-irregular pairs, which we also need to anonymize. This is particularly an issue when hubs fall within such an irregular pair. However, note that increasing l (i.e., reducing the size k of the anonymity sets) allows us to preserve the average clustering coefficient better. In general, a low value of l implies larger anonymity groups, but it also forces the heuristic procedure used to approximate the ε-regular partition to bring more edges (and triangles) inside the groups, which are then affected by the Erdös-Rényi rewiring. Indeed, high anonymity demands several more structural modifications. In practice it is common to look for smaller k-anonymity groups (i.e., larger l), and for these values we are better able to preserve the average clustering coefficient information.

Finally, **Figure 2** shows the cosine similarity and the Spearman's rank correlation between the page rank vectors (Page et al., 1999) of the original and anonymized graphs. The results confirm that the proposed anonymization procedure is able to preserve well the centrality information of the nodes, once again with the quality of the approximation generally improving as we reduce the size of the anonymity groups.

5. CONCLUSION

We considered the problem of protecting the identity of the nodes of a network from an attacker with background structural knowledge. We proposed to use the Szemerédi regularity lemma to compute an ε-regular partition of the original graph which is then anonymized by injecting Erdös-Rényi at selected locations. This creates a k-anonymous graph where the loss of structural information is minimized. We validated our method on three real-world networks abstracted from Facebook. Future work should perform a more extensive evaluation of the proposed method on larger graphs, with a wider range of values, and compare our method with alternative anonymization approaches.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: a https://snap.stanford.edu/data/index.html.

AUTHOR CONTRIBUTIONS

AT: conceptualization. LR and AT: methodology. DF: software. DF, LR, and AT: investigation, writing–review, and editing. LR: writing–original draft preparation.

[3] The JS divergence takes a value between 0 and 1, with 0 indicating identical distributions. Results on other datasets are omitted due to space constraints.

[4] Note, however, that the value of the JS divergence is biased by the fact that most of the probability mass is on low-degree nodes.

REFERENCES

Backstrom, L., Dwork, C., and Kleinberg, J. (2007). "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography," in *Proceedings of the 16th International Conference on World Wide Web (WWW '07)* (Banff, AB), 181–190. doi: 10.1145/1242572.1242598

Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509–512. doi: 10.1126/science.286.5439.509

Cheng, J., Fu, A. W.-C., and Liu, J. (2010). "K-isomorphism: privacy preserving network publication against structural attacks," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD '10)* (Indianapolis, IN), 459–470. doi: 10.1145/1807167.1807218

Chorley, M. J., Rossi, L., Tyson, G., and Williams, M. J. (2016). "Pub crawling at scale: tapping untappd to explore social drinking," in *Tenth International AAAI Conference on Web and Social Media* (Cologne). Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13048 (accessed May 20, 2019).

Diestel, R. (2012). *Graph Theory*. Graduate Texts in Mathematics, Vol. 173. Available online at: https://www.springer.com/gp/book/9783662536216

Erdős, P. (1960). Graphs with prescribed degrees of vertices (hungarian). *Mat. Lapok* 11, 264–274.

Fiorucci, M., Pelosin, F., and Pelillo, M. (2019). Separating structure from noise in large graphs using the regularity lemma. *CoRR* abs/1905.06917.

Fiorucci, M., Torcinovich, A., Curado, M., Escolano, F., and Pelillo, M. (2017). "On the interplay between strong regularity and graph densification," in *11th IAPR-TC-15 International Workshop, GbRPR 2017* (Anacapri), 165–174.

Fung, B., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: a survey of recent developments. *ACM Comput. Surveys* 42:14. doi: 10.1201/9781420091502

Gerke, S., and Steger, A. (2005). The sparse regularity lemma and its applications. *Surveys Combin.* 327, 227–258. doi: 10.1017/CBO9780511734885.010

Hay, M., Miklau, G., Jensen, D., Towsley, D., and Weis, P. (2008). Resisting structural re-identification in anonymized social networks. *Proc. VLDB Endow.* 1, 102–114. doi: 10.14778/1453856.1453873

Komlós, J., and Simonovits, M. (1996). Szemerédi's regularity lemma and its applications in graph theory. *Combinatorics* 2, 295–352.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). "What is twitter, a social network or a news media?," in *Proceedings of the 19th International Conference on World Wide Web (WWW '10)* (Raleigh, NC), 591–600. doi: 10.1145/1772690.1772751

Leskovec, J., and Mcauley, J. J. (2012). "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems*, 539–547.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Trans. Inform. Theor.* 37, 145–151. doi: 10.1109/18.61115

Liu, K., and Terzi, E. (2008). "Towards identity anonymization on graphs," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08)* (Vancouver, BC), 93–106.

Ma, J., Qiao, Y., Hu, G., Huang, Y., Sangaiah, A. K., Zhang, C., et al. (2018). De-anonymizing social networks with random forest classifier. *IEEE Access* 6, 10139–10150. doi: 10.1109/ACCESS.2017.2756904

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The Pagerank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66, Stanford InfoLab.

Pelillo, M., Elezi, I., and Fiorucci, M. (2017). Revealing structure in large graphs: Szemerédi's regularity lemma and its use in pattern recognition. *Pattern Recogn. Lett.* 87, 4–11. doi: 10.1016/j.patrec.2016.09.007

Qian, J., Li, X.-Y., Zhang, C., and Chen, L. (2016). "De-anonymizing social networks and inferring private attributes using knowledge graphs," in *IEEE INFOCOM 2016–The 35th Annual IEEE International Conference on Computer Communications* (San Francisco, CA), 1–9.

Rossi, L., and Musolesi, M. (2014). "It's the way you check-in: identifying users in location-based social networks," in *Proceedings of the Second ACM Conference on Online Social Networks (COSN '14)* (Dublin), 215–226. doi: 10.1145/2660460.2660485

Rossi, L., Musolesi, M., and Torsello, A. (2015a). "On the k-anonymization of time-varying and multi-layer social graphs," in *Ninth International AAAI Conference on Web and Social Media* (Oxford).

Rossi, L., Williams, M., Stich, C., and Musolesi, M. (2015b). "Privacy and the city: user identification and location semantics in location-based social networks," in *Ninth International AAAI Conference on Web and Social Media* (Oxford).

Rousseau, F., Casas-Roma, J., and Vazirgiannis, M. (2018). Community-preserving anonymization of graphs. *Knowl. Inform. Syst.* 54, 315–343. doi: 10.1007/s10115-017-1064-y

Rozemberczki, B., Davies, R., Sarkar, R., and Sutton, C. (2018). Gemsec: Graph embedding with self clustering. *arXiv preprint arXiv*:1802.03997.

Sperotto, A., and Pelillo, M. (2007). "Szemerédi's regularity lemma and its applications to pairwise clustering and segmentation," in *Proceedings of the 6th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR'07)* (Ezhou), 13–27.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature* 393, 440. doi: 10.1038/30918

Zhou, B., and Pei, J. (2011). The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowl. Inform. Syst.* 28, 47–77. doi: 10.1007/s10115-010-0311-2

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 04 June 2019
doi: 10.3389/fdata.2019.00008

Abusive Language Detection in Online Conversations by Combining Content- and Graph-Based Features

Noé Cécillon, Vincent Labatut, Richard Dufour and Georges Linarès*

LIA, Avignon University, Avignon, France

In recent years, online social networks have allowed world-wide users to meet and discuss. As guarantors of these communities, the administrators of these platforms must prevent users from adopting inappropriate behaviors. This verification task, mainly done by humans, is more and more difficult due to the ever growing amount of messages to check. Methods have been proposed to automatize this moderation process, mainly by providing approaches based on the textual content of the exchanged messages. Recent work has also shown that characteristics derived from the structure of conversations, in the form of conversational graphs, can help detecting these abusive messages. In this paper, we propose to take advantage of both sources of information by proposing fusion methods integrating content- and graph-based features. Our experiments on raw chat logs show not only that the content of the messages, but also their dynamics within a conversation contain partially complementary information, allowing performance improvements on an abusive message classification task with a final *F*-measure of 93.26%.

Keywords: automatic abuse detection, content analysis, conversational graph, online conversations, social networks

OPEN ACCESS

Edited by:
Sabrina Gaito, University of Milan, Italy

Reviewed by:
Roberto Interdonato, Territoires, Environnement, Télédètection et Information Spatiale (TETIS), France
Eric A. Leclercq, Université de Bourgogne, France

****Correspondence:***
Vincent Labatut
vincent.labatut@univ-avignon.fr

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *01 April 2019*
Accepted: *14 May 2019*
Published: *04 June 2019*

Citation:
Cécillon N, Labatut V, Dufour R and Linarès G (2019) Abusive Language Detection in Online Conversations by Combining Content- and Graph-Based Features. Front. Big Data 2:8. doi: 10.3389/fdata.2019.00008

1. INTRODUCTION

The internet has widely impacted the way we communicate. Online communities, in particular, have grown to become important places for interpersonal communications. They get more and more attention from companies to advertise their products or from governments interested in monitoring public discourse. Online communities come in various shapes and forms, but they are all exposed to abusive behavior. The definition of what exactly is considered as abuse depends on the community, but generally includes personal attacks, as well as discrimination based on race, religion, or sexual orientation.

Abusive behavior is a risk, as it is likely to make important community members leave, therefore endangering the community, and even trigger legal issues in some countries. Moderation consists in detecting users who act abusively, and in taking actions against them. Currently, this moderation work is mainly a manual process, and since it implies high human and financial costs, companies have a keen interest in its automation. One way of doing so is to consider this task as a classification problem consisting in automatically determining if a user message is abusive or not.

A number of works have tackled this problem, or related ones, in the literature. Most of them focus only on the content of the targeted message to detect abuse or similar properties. For instance (Spertus, 1997), applies this principle to detect hostility (Dinakar et al., 2011), for cyberbullying, and (Chen et al., 2012) for offensive language. These approaches rely on a mix of

standard NLP features and manually crafted application-specific resources (e.g., linguistic rules). We also proposed a content-based method (Papegnies et al., 2017a) using a wide array of language features (Bag-of-Words, *tf-idf* scores, sentiment scores). Other approaches are more machine learning intensive, but require larger amounts of data. Recently, Wulczyn et al. (2017) created three datasets containing individual messages collected from Wikipedia discussion pages, annotated for toxicity, personal attacks and aggression, respectively. They have been leveraged in recent works to train Recursive Neural Network operating on word embeddings and character *n*-gram features (Pavlopoulos et al., 2017; Mishra et al., 2018). However, the quality of these direct content-based approaches is very often related to the training data used to learn abuse detection models. In the case of online social networks, the great variety of users, including very different language registers, spelling mistakes, as well as intentional users obfuscation, makes it almost impossible to have models robust enough to be applied in all cases. (Hosseini et al., 2017) have then shown that it is very easy to bypass automatic toxic comment detection systems by making the abusive content difficult to detect (intentional spelling mistakes, uncommon negatives...).

Because the reactions of other users to an abuse case are completely beyond the abuser's control, some authors consider the content of messages occurring *around* the targeted message, instead of focusing only on the targeted message itself. For instance, (Yin et al., 2009) use features derived from the sentences neighboring a given message to detect harassment on the Web. (Balci and Salah, 2015) take advantage of user features such as the gender, the number of in-game friends or the number of daily logins to detect abuse in the community of an online game. In our previous work (Papegnies et al., 2019), we proposed a radically different method that completely ignores the textual content of the messages, and relies only on a graph-based modeling of the conversation. This is the only graph-based approach ignoring the linguistic content proposed in the context of abusive messages detection. Our conversational network extraction process is inspired from other works leveraging such graphs for other purposes: chat logs (Mutton, 2004) or online forums (Forestier et al., 2011) interaction modeling, user group detection (Camtepe et al., 2004). Additional references on abusive message detection and conversational network modeling can be found in Papegnies et al. (2019).

In this paper, based on the assumption that the interactions between users and the content of the exchanged messages convey different information, we propose a new method to perform abuse detection while leveraging both sources. For this purpose, we take advantage of the content-(Papegnies et al., 2017b) and graph-based (Papegnies et al., 2019) methods that we previously developed. We propose three different ways to combine them, and compare their performance on a corpus of chat logs originating from the community of a French multiplayer online game. We then perform a feature study, finding the most informative ones and discussing their role. Our contribution is twofold: the exploration of fusion methods, and more importantly the identification of discriminative features for this problem.

The rest of this article is organized as follows. In section 2, we describe the methods and strategies used in this work. In section 3 we present our dataset, the experimental setup we use for this classification task, and the performances we obtained. Finally, we summarize our contributions in section 4 and present some perspectives for this work.

2. METHODS

In this section, we summarize the content-based method from Papegnies et al. (2017b) (section 2.1) and the graph-based method from Papegnies et al. (2019) (section 2.2). We then present the fusion method proposed in this paper, aiming at taking advantage of both sources of information (section 2.3). **Figure 1** shows the whole process, and is discussed through this section.

2.1. Content-Based Method

This method corresponds to the bottom-left part of **Figure 1** (in green). It consists in extracting certain features from the content of each considered message, and to train a Support Vector Machine (SVM) classifier to distinguish abusive (*Abuse* class) and non-abusive (*Non-abuse* class) messages (Papegnies et al., 2017b). These features are quite standard in Natural Language Processing (NLP), so we only describe them briefly here.

We use a number of *morphological features*. We use the message length, average word length, and maximal word length, all expressed in number of characters. We count the number of unique characters in the message. We distinguish between six classes of characters (letters, digits, punctuation, spaces, and others) and compute two features for each one: number of occurrences, and proportion of characters in the message. We proceed similarly with capital letters. Abusive messages often contain a lot of copy/paste. To deal with such redundancy, we apply the Lempel–Ziv–Welch (LZW) compression algorithm (Batista and Meira, 2004) to the message and take the ratio of its raw to compress lengths, expressed in characters. Abusive messages also often contain extra-long words, which can be identified by collapsing the message: extra occurrences of letters repeated more than two times consecutively are removed. For instance, "looooooool" would be collapsed to "lool". We compute the difference between the raw and collapsed message lengths.

We also use *language features*. We count the number of words, unique words and bad words in the message. For the latter, we use a predefined list of insults and symbols considered as abusive, and we also count them in the collapsed message. We compute two overall *tf–idf* scores corresponding to the sums of the standard *tf–idf* scores of each individual word in the message. One is processed relatively to the *Abuse* class, and the other to the *Non-abuse* class. We proceed similarly with the collapsed message. Finally, we lower-case the text and strip punctuation, in order to represent the message as a basic Bag-of-Words (BoW). We then train a Naive Bayes classifier to detect abuse using this sparse binary vector (as represented in the very bottom part of **Figure 1**). The output of this simple classifier is then used as an input feature for the SVM classifier.

2.2. Graph-Based Method

This method corresponds to the top-left part of **Figure 1** (in red). It completely ignores the content of the messages, and only focuses on the dynamics of the conversation, based on the interactions between its participants (Papegnies et al., 2019). It is three-stepped: (1) extracting a conversational graph based on the considered message as well as the messages preceding and/or following it; (2) computing the topological measures of this graph to characterize its structure; and (3) using these values as features to train an SVM to distinguish between abusive and non-abusive messages. The vertices of the graph model the participants of the conversation, whereas its weighted edges represent how intensely they communicate.

The graph extraction is based on a number of concepts illustrated in **Figure 2**, in which each rectangle represents a message. The extraction process is restricted to a so-called *context period*, i.e., a sub-sequence of messages including the message of interest, itself called *targeted message* and represented in red in **Figure 2**. Each participant posting at least one message during this period is modeled by a vertex in the produced conversational graph. A mobile window is slid over the whole period, one message at a time. At each step, the network is updated either by creating new links, or by updating the weights of existing ones. This *sliding window* has a fixed length expressed in number of messages, which is derived from ergonomic constraints relative to the online conversation platform studied in section 3. It allows focusing on a smaller part of the context period. At a given time, the last message of the window (in blue in **Figure 2**) is called *current message* and its author *current author*. The weight update method assumes that the current message is aimed at the authors of the other messages present in the window, and therefore connects the current author to them (or strengthens their weights if the edge already exists). It also takes chronology into account by favoring the most recent authors in the window. Three different variants of the conversational network are extracted for one given targeted message: the *Before* network is based on the messages posted before the targeted message, the *After* network on those posted after, and the *Full* network on the whole context period.

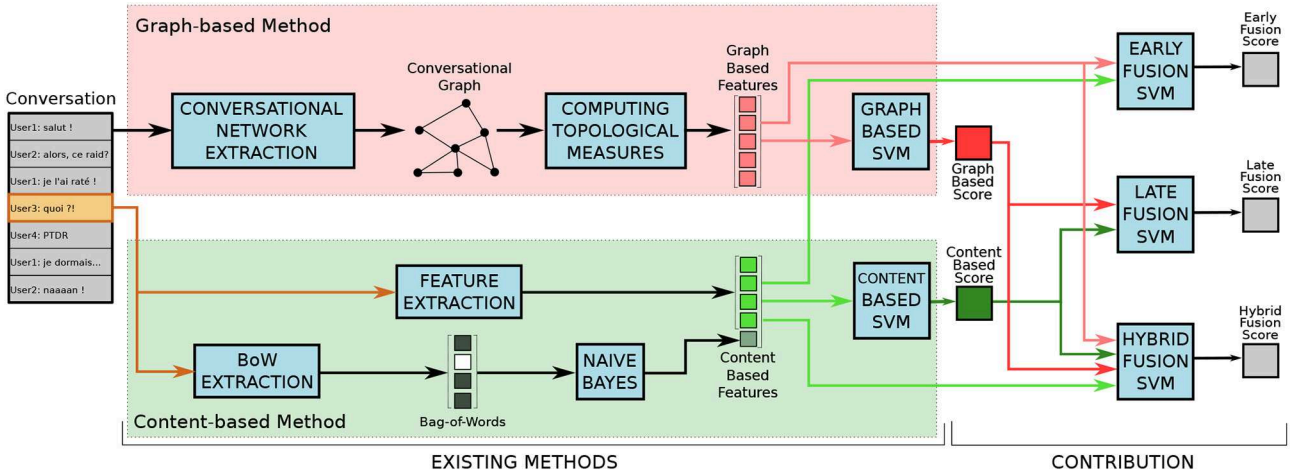

FIGURE 1 | Representation of our processing pipeline. *Existing methods* refers to our previous work described in Papegnies et al. (2017b) (content-based method) and Papegnies et al. (2019) (graph-based method), whereas the contribution presented in this article appears on the right side (fusion strategies). Figure available at 10.6084/m9.figshare.7442273 under CC-BY license.

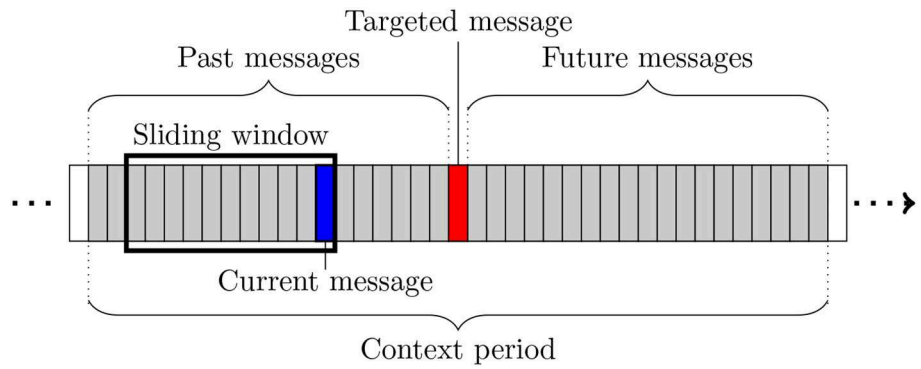

FIGURE 2 | Illustration of the main concepts used during network extraction (see text for details). Figure available at 10.6084/m9.figshare.7442273 under CC-BY license.

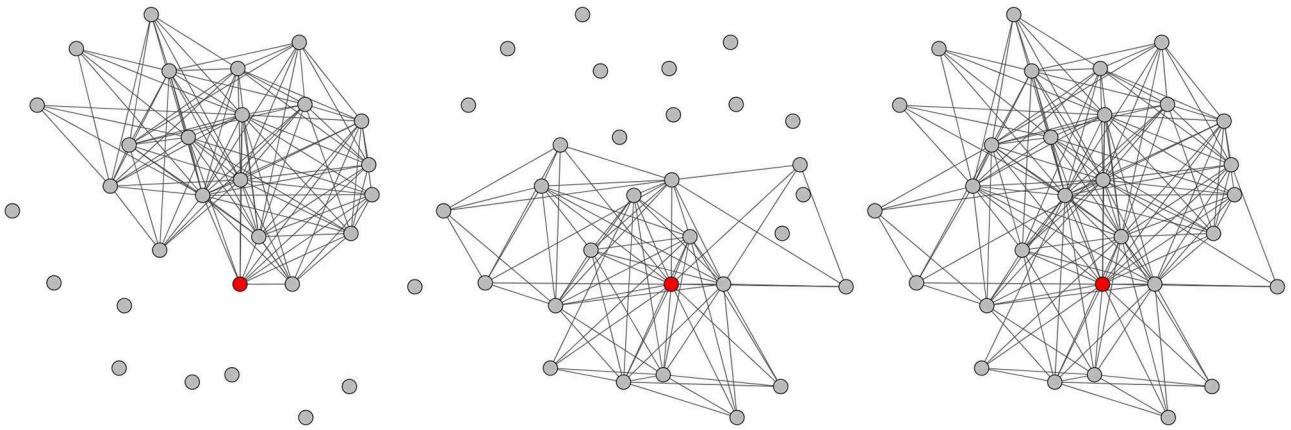

FIGURE 3 | Example of the three types of conversational networks extracted for a given context period: *Before* **(Left)**, *After* **(Center)**, and *Full* **(Right)**. The author of the targeted message is represented in red. Figure available at 10.6084/m9.figshare.7442273 under CC-BY license.

Figure 3 shows an example of such networks obtained for a message of the corpus described in section 3.1.

Once the conversational networks have been extracted, they must be described through numeric values in order to feed the SVM classifier. This is done through a selection of standard topological measures allowing to describe a graph in a number of distinct ways, focusing on different scales and scopes. The *scale* denotes the nature of the characterized entity. In this work, the individual vertex and the whole graph are considered. When considering a single vertex, the measure focuses on the *targeted author* (i.e., the author of the targeted message). The *scope* can be either micro-, meso-, or macroscopic: it corresponds to the amount of information considered by the measure. For instance, the graph density is microscopic, the modularity is mesoscopic, and the diameter is macroscopic. All these measures are computed for each graph, and allow describing the conversation surrounding the message of interest. The SVM is then trained using these values as features. In this work, we use exactly the same measures as in Papegnies et al. (2019).

2.3. Fusion

We now propose a new method seeking to take advantage of both previously described ones. It is based on the assumption that the content- and graph-based features convey different information. Therefore, they could be complementary, and their combination could improve the classification performance. We experiment with three different fusion strategies, which are represented in the right-hand part of **Figure 1**.

The first strategy follows the principle of *Early Fusion*. It consists in constituting a global feature set containing all content- and graph-based features from sections 2.1 and 2.2, then training a SVM directly using these features. The rationale here is that the classifier has access to the whole raw data, and must determine which part is relevant to the problem at hand.

The second strategy is *Late Fusion*, and we proceed in two steps. First, we apply separately both methods described in sections 2.1 and 2.2, in order to obtain two scores corresponding to the output probability of each message to be abusive given by the content- and graph-based methods, respectively. Second, we fetch these two scores to a third SVM, trained to determine if a message is abusive or not. This approach relies on the assumption that these scores contain all the information the final classifier needs, and not the noise present in the raw features.

Finally, the third fusion strategy can be considered as *Hybrid Fusion*, as it seeks to combine both previous proposed ones. We create a feature set containing the content- and graph-based features, like with *Early Fusion*, but also both scores used in *Late Fusion*. This whole set is used to train a new SVM. The idea is to check whether the scores do not convey certain useful information present in the raw features, in which case combining scores and features should lead to better results.

3. EXPERIMENTS

In this section, we first describe our dataset and the experimental protocol followed in our experiments (section 3.1). We then present and discuss our results, in terms of classification performance (sections 3.2) and feature selection (section 3.3).

3.1. Experimental Protocol

The dataset is the same as in our previous publications (Papegnies et al., 2017b, 2019). It is a proprietary database containing 4,029,343 messages in French, exchanged on the in-game chat of *SpaceOrigin*[1], a Massively Multiplayer Online Role-Playing Game (MMORPG). Among them, 779 have been flagged as being abusive by at least one user in the game, and confirmed as such by a human moderator. They constitute what we call the *Abuse* class. Some inconsistencies in the database prevent us from retrieving the context of certain messages, which we remove from the set. After this cleaning, the *Abuse* class contains 655 messages. In order to keep a balanced dataset, we further extract the same number of messages at random from the ones that have not been flagged as abusive. This constitutes our *Non-abuse* class.

[1]https://play.spaceorigin.fr/

Each message, whatever its class, is associated to its surrounding context (i.e., messages posted in the same thread).

The graph extraction method used to produce the graph-based features requires to set certain parameters. We use the values matching the best performance, obtained during the greedy search of the parameter space performed in Papegnies et al. (2019). In particular, regarding the two most important parameters (see section 2.2), we fix the *context period* size to 1,350 messages and the *sliding window* length to 10 messages. Implementation-wise, we use the iGraph library (Csardi and Nepusz, 2006) to extract the conversational networks and process the corresponding features. We use the Sklearn toolkit (Pedregosa et al., 2011) to get the text-based features. We use the SVM classifier implemented in Sklearn under the name SVC (C-Support Vector Classification). Because of the relatively small dataset, we set-up our experiments using a 10-fold cross-validation. Each fold is balanced between the *Abuse* and *Non-abuse* classes, 70% of the dataset being used for training and 30% for testing.

3.2. Classification Performance

Table 1 presents the Precision, Recall and *F*-measure scores obtained on the *Abuse* class, for both baselines [*Content-based* (Papegnies et al., 2017b) and *Graph-based* (Papegnies et al., 2019)] and all three proposed fusion strategies (*Early Fusion*, *Late Fusion* and *Hybrid Fusion*). It also shows the number of features used to perform the classification, the time required to compute the features and perform the cross validation (*Total Runtime*) and to compute one message in average (*Average Runtime*). Note that *Late Fusion* has only 2 direct inputs (content- and graph-based SVMs), but these in turn have their own inputs, which explains the values displayed in the table.

Our first observation is that we get higher *F*-measure values compared to both baselines when performing the fusion, independently from the fusion strategy. This confirms what we expected, i.e., that the information encoded in the interactions between the users differs from the information conveyed by the content of the messages they exchange. Moreover, this shows that both sources are at least partly complementary, since the performance increases when merging them. On a side note, the correlation between the score of the graph- and content-based classifiers is 0.56, which is consistent with these observations.

Next, when comparing the fusion strategies, it appears that *Late Fusion* performs better than the others, with an *F*-measure of 93.26. This is a little bit surprising: we were expecting to get superior results from the *Early Fusion*, which has direct access to a much larger number of *raw* features (488). By comparison, the *Late Fusion* only gets 2 features, which are themselves the outputs of two other classifiers. This means that the *Content-Based* and *Graph-Based* classifiers do a good work in summarizing their inputs, without loosing much of the information necessary to efficiently perform the classification task. Moreover, we assume that the *Early Fusion* classifier struggles to estimate an appropriate model when dealing with such a large number of features, whereas the *Late Fusion* one benefits from the pre-processing performed by its two predecessors, which act as if reducing the dimensionality of the data. This seems to be confirmed by the results of the *Hybrid Fusion*, which produces better results than the *Early Fusion*, but is still below the *Late Fusion*. This point could be explored by switching to classification algorithm less sensitive to the number of features. Alternatively, when considering the three SVMs used for the *Late Fusion*, one could see a simpler form of a very basic Multilayer Perceptron, in which each neuron has been trained separately (without system-wide backpropagation). This could indicate that using a regular Multilayer Perceptron directly on the raw features could lead to improved results, especially if enough training data is available.

Regarding runtime, the graph-based approach takes more than 8 h to run for the whole corpus, mainly because of the feature computation step. This is due to the number of features, and to the compute-intensive nature of some of them. The content-based approach is much faster, with a total runtime of $<$1 min, for the exact opposite reasons. Fusion methods require to compute both content- and graph-based features, so they have the longest runtime.

3.3. Feature Study

We now want to identify the most discriminative features for all three fusion strategies. We apply an iterative method based on the *Sklearn* toolkit, which allows us to fit a linear kernel SVM to the dataset and provide a ranking of the input features reflecting their

TABLE 1 | Comparison of the performances obtained with the methods (*Content-based*, *Graph-based*, *Fusion*) and their subsets of *Top Features* (TF).

Method	Number of features	Total runtime	Average runtime	Precision	Recall	*F*-measure
Content-Based	29	0:52	0.02s	78.59	83.61	81.02
Content-Based TF	3	0:21	0.01s	75.82	82.57	79.05
Graph-Based	459	8:19:10	7.56s	90.21	87.63	88.90
Graph-Based TF	10	14:22	0.03s	88.72	84.87	86.75
Early Fusion	488	8:26:41	7.68s	91.25	89.45	90.34
Early Fusion TF	4	11:29	0.17s	89.09	87.12	88.09
Late Fusion	488 (2)	8:23:57	7.64s	94.10	92.43	93.26
Late Fusion TF	13	15:42	0.24s	91.64	89.97	90.80
Hybrid Fusion	490	8:27:01	7.68s	91.96	90.48	91.22
Hybrid Fusion TF	4	16:57	0.26s	90.74	89.00	89.86

The total runtime is expressed as h:min:s. See text for details.

importance in the classification process. Using this ranking, we identify the least discriminant feature, remove it from the dataset, and train a new model with the remaining features. The impact of this deletion is measured by the performance difference, in terms of *F*-measure. We reiterate this process until only one feature remains. We call *Top Features* (TF) the minimal subset of features allowing to reach 97% of the original performance (when considering the complete feature set).

We apply this process to both baselines and all three fusion strategies. We then perform a classification using only their respective TF. The results are presented in **Table 1**. Note that the *Late Fusion TF* performance is obtained using the scores produced by the SVMs trained on *Content-based TF* and *Graph-based TF*. These are also used as features when computing the TF for *Hybrid Fusion TF* (together with the raw content- and graph-based features). In terms of classification performance, by construction, the methods are ranked exactly like when considering all available features.

The *Top Features* obtained for each method are listed in **Table 2**. The last 4 columns precise which variants of the graph-based features are concerned. Indeed, as explained in section 2.2, most of these topological measures can handle/ignore edge weights and/or edge directions, can be vertex- or graph-focused, and can be computed for each of the three types of networks (*Before*, *After*, and *Full*).

There are three *Content-Based TF*. The first is the *Naive Bayes* prediction, which is not surprising as it comes from a fully fledged classifier processing BoWs. The second is the *tf-idf score* computed over the *Abuse* class, which shows that considering term frequencies indeed improve the classification performance. The third is the *Capital Ratio* (proportion of capital letters in the comment), which is likely to be caused by abusive message tending to be shouted, and therefore written in capitals. The *Graph-Based TF* are discussed in depth in our previous article (Papegnies et al., 2019). To summarize, the most important features help detecting changes in the direct neighborhood of the targeted author (Coreness, Strength), in the average node centrality at the level of the whole graph in terms of distance (Closeness), and in the general reciprocity of exchanges between users (Reciprocity).

We obtain 4 features for *Early Fusion TF*. One is the *Naive Bayes* feature (content-based), and the other three are topological measures (graph-based features). Two of the latter correspond to the Coreness of the targeted author, computed for the *Before* and *After* graphs. The third topological measure is his/her Eccentricity. This reflects important changes in the interactions around the targeted author. It is likely caused by angry users piling up on the abusive user after he has posted some inflammatory remark. For *Hybrid Fusion TF*, we also get 4 features, but those include in first place both SVM outputs from the content- and graph-based classifiers. Those are completed by 2 graph-based features, including Strength (also found in the *Graph-based* and *Late Fusion TF*) and Coreness (also found in the *Graph-based*, *Early Fusion* and *Late Fusion TF*).

Besides a better understanding of the dataset and classification process, one interesting use of the TF is that they can allow decreasing the computational cost of the classification. In our case, this is true for all methods: we can retain 97% of the performance while using only a handful of features instead of

TABLE 2 | Top features obtained for our 5 methods.

Method	Top Features	Graph	Weights	Directions	Scale
Content-Based	Naive Bayes	–	–	–	–
	tf–idf Abuse Score	–	–	–	–
	Character Capital Ratio	–	–	–	–
Graph-Based	Coreness Score	F	–	I	G
	PageRank Centrality	A	U	D	N
	Strength Centrality	F	W	O	N
	Vertex Count	F	–	–	G
	Closeness Centrality	B	W	O	G
	Closeness Centrality	B	W	O	N
	Authority Score	B	W	D	G
	Hub Score	B	U	D	N
	Reciprocity	A	–	D	G
	Closeness Centrality	A	W	U	N
Early Fusion	Coreness Score	A	–	O	G
	Coreness Score	B	–	I	G
	Eccentricity	B	–	I	G
	Naive Bayes	–	–	–	–
Late Fusion	*Content-Based* TF ∪ *Graph-Based TF*	–	–	–	–
Hybrid Fusion	Graph-based output	–	–	–	–
	Content-based output	–	–	–	–
	Strength Centrality	A	W	O	N
	Coreness Score	B	–	I	G

The letters in the Graph column stand for Before (B), After (A), and Full (F). Those in the Weights and Directions columns stand for: Unweighted or Undirected (U), Weighted (W), Directed (D), Incoming (I), and Outgoing (O). Those in the Scale column mean Graph-scale (G) or Vertex-scale (N).

hundreds. For instance, with the *Late Fusion TF*, we need only 3% of the total *Late Fusion* runtime.

4. CONCLUSION AND PERSPECTIVES

In this article, we tackle the problem of automatic abuse detection in online communities. We take advantage of the methods that we previously developed to leverage message content (Papegnies et al., 2017a) and interactions between users (Papegnies et al., 2019), and create a new method using both types of information simultaneously. We show that the features extracted from our content- and graph-based approaches are complementary, and that combining them allows to sensibly improve the results up to 93.26 (*F*-measure). One limitation of our method is the computational time required to extract certain features. However, we show that using only a small subset of relevant features allows to dramatically reduce the processing time (down to 3%) while keeping more than 97% of the original performance.

Another limitation of our work is the small size of our dataset. We must find some other corpora to test our methods at a much higher scale. However, all the available datasets are composed of isolated messages, when we need threads to make the most of our approach. A solution could be to start from datasets such as the Wikipedia-based corpus proposed by Wulczyn et al. (2017), and complete them by reconstructing the original conversations containing the annotated messages. This could also be the opportunity to test our methods on an other language than French. Our content-based method may be impacted by this change, but this should not be the case for the graph-based method, as it is independent from the content (and therefore the language). Besides language, a different online community is likely to behave differently from the one we studied before. In particular, its members could react differently to abuse. The Wikipedia dataset would therefore allow assessing how such cultural differences affect our classifiers, and identifying which observations made for Space Origin still apply to Wikipedia.

DATA AVAILABILITY

The datasets for this manuscript are not publicly available because Private dataset. Requests to access the data should actually be addressed to the corresponding author, V. Labatut.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Balci, K., and Salah, A. A. (2015). Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. *Comput. Hum. Behav.* 53, 517–526. doi: 10.1016/j.chb.2014.10.025

Batista, L. V., and Meira, M. M. (2004). "Texture classification using the lempel-ziv-welch algorithm," in *Brazilian Symposium on Artificial Intelligence* (Berlin), 444–453. doi: 10.1007/978-3-540-28645-5-45

Camtepe, A., Krishnamoorthy, M. S., and Yener, B. (2004). "A tool for Internet chatroom surveillance," in *International Conference on Intelligence and Security Informatics*, Vol 3073 of *Lecture Notes in Computer Science* (Berlin: Springer), 252–265. doi: 10.1007/978-3-540-25952-7-19

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). "Detecting offensive language in social media to protect adolescent online safety," in *International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing* (Amsterdam: IEEE), 71–80.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Int. J.* 1695, 1–9.

Dinakar, K., Reichart, R., and Lieberman, H. (2011). "Modeling the detection of textual cyberbullying," in *5th International AAAI Conference on Weblogs and Social Media / Workshop on the Social Mobile Web* (Barcelona: AAAI), 11–17.

Forestier, M., Velcin, J., and Zighed, D. (2011). "Extracting social networks to understand interaction," in *International Conference on Advances in Social Networks Analysis and Mining* (Kaohsiung: IEEE), 213–219. doi: 10.1109/ASONAM.2011.64

Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. *arXiv* arXiv:1702.08138.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2018). "Neural character-based composition models for abuse detection," in *2nd Workshop on Abusive Language Online* (Brussels: Association for Computational Linguistics), 1–10. Available online at: https://www.aclweb.org/anthology/W18-5101

Mutton, P. (2004). "Inferring and visualizing social networks on Internet Relay Chat," in *8th International Conference on Information Visualisation* (London: IEEE) 35–43.

Papegnies, E., Labatut, V., Dufour, R., and Linares, G. (2017a). "Graph-based features for automatic online abuse detection," in *International Conference on Statistical Language and Speech Processing*, volume 10583 of *Lecture Notes in Computer Science* (Berlin: Springer), 70–81. doi: 10.1007/978-3-319-68456-7-6

Papegnies, E., Labatut, V., Dufour, R., and Linares, G. (2017b). "Impact of content features for automatic online abuse detection," in *International Conference on Computational Linguistics and Intelligent Text Processing*, volume 10762 of *Lecture Notes in Computer Science* (Berlin:Springer), 404–419. doi: 10.1007/978-3-319-77116-8-30

Papegnies, E., Labatut, V., Dufour, R., and Linares, G. (2019). Conversational networks for automatic online moderation. *IEEE Trans. Comput. Soc. Syst.* 6, 38–55. doi: 10.1109/TCSS.2018.2887240

Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. (2017). "Deep learning for user comment moderation," in *1st Workshop on Abusive Language Online* (Vancouver, BC: ACL), 25–35. doi: 10.18653/v1/W17-3004

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Spertus, E. (1997). "Smokey: automatic recognition of hostile messages," in *14th National Conference on Artificial Intelligence and 9th Conference on Innovative Applications of Artificial Intelligence* (Providence, RI: AAAI), 1058–1065.

Wulczyn, E., Thain, N., and Dixon, L. (2017). "Ex Machina: personal attacks seen at scale," in *26th International Conference on World Wide Web* (Geneva), 1391–1399. doi: 10.1145/3038912.3052591

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. (2009). Detection of harassment on Web 2.0. in *WWW Workshop: Content Analysis in the Web 2.0* (Madrid) 1–7.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fdata.2019.00011

A Digital Nudge to Counter Confirmation Bias

Calum Thornhill, Quentin Meeus, Jeroen Peperkamp and Bettina Berendt**

Department of Computer Science, KU Leuven, Leuven, Belgium

Fake news is increasingly an issue on social media platforms. In this work, rather than detect misinformation, we propose the use of nudges to help steer internet users into fact checking the news they read online. We discuss two types of nudging strategies, by presentation and by information. We present the tool BalancedView, a proof-of-concept that shows news stories relevant to a tweet. The method presents the user with a selection of articles from a range of reputable news sources providing alternative opinions from the whole political spectrum, with these alternative articles identified as matching the original one by a combination of natural language processing and search. The results of an initial user study of BalancedView suggest that nudging by information may change the behavior of users towards that of informed news readers.

Keywords: digital nudging, fake news, confirmation bias, NLP (natural language processing), Twitter

OPEN ACCESS

Edited by:
Sabrina Gaito,
University of Milan, Italy

Reviewed by:
Giuseppe Mangioni,
University of Catania, Italy
Shuhan Yuan,
University of Arkansas, United States

****Correspondence:***
Calum Thornhill
calum.thornhill@student.kuleuven.be
Bettina Berendt
bettina.berendt@cs.kuleuven.be

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *25 March 2019*
Accepted: *22 May 2019*
Published: *06 June 2019*

Citation:
Thornhill C, Meeus Q, Peperkamp J and Berendt B (2019) A Digital Nudge to Counter Confirmation Bias. Front. Big Data 2:11. doi: 10.3389/fdata.2019.00011

1. INTRODUCTION

Information disorder in current information ecosystems arises not only from the publication of "fake news," but also from individuals' subjective reading of news and from their propagating news to others.

Sometimes the difference between real and fake information is apparent. However, often a message is written to evoke certain emotions and opinions by taking partially true base stories and injecting false statements such that the information looks realistic. In addition, the perception of the trustworthiness of news is often influenced by confirmation bias. As a result, people often believe distorted or outright incorrect news and spread such misinformation further.

For example, it was shown that in the months preceding the 2016 American presidential election, organizations from both Russia and Iran ran organized efforts to create such stories and spread them on Twitter and Facebook (Cohen, 2018).

It is therefore important to raise internet users' awareness of such practices. Key to this is providing users with means to understand whether information should be trusted or not.

A solution put forward by social networks relies on users identifying suspicious articles shared on their platforms. Such articles are subsequently fact-checked by third-party volunteers. Then, when another user comes across such an article, they are given the chance to read an alternative article that has been deemed trustworthy.

However, this method is labor-intensive and requires highly skilled humans and therefore does not scale. In addition, important fact-checking organizations have become disillusioned by social networks' handling of the "fake news" problem and of their fact-checking efforts, and have withdrawn their support (Lee, 2019).

In this work, we propose BalancedView, a novel, low-cost, and scalable method for fighting the spread of misinformation without having to rely on users reporting or third parties checking news items.

The method presents the user with a selection of articles from a range of reputable news sources providing alternative opinions from the whole political spectrum, with these alternative articles identified as matching the original one by a combination of natural language processing and search. The strategy is a form of digital nudge, in which the user is presented with an original text together with articles showing wider context and alternative standpoints within close view.

Our main objective for such a tool is to educate people about sharing and believing information accessed online, which in turn can decrease the spread of fake news. We also hope to raise awareness of the different ways information can be presented and manipulated online.

In section 2, we briefly discuss the mechanisms of misinformation spreading online and how social networks are the perfect platforms to accelerate this process, and we give a brief overview of related research in the field of fake news and nudge design. section 3 gives a high level description of our approach. In section 4, we discuss the nudging strategies considered.

Technical design and the inner workings are covered in section 5, along with first evaluations of algorithm and user assessments in section 6. We conclude with an outlook on future research.

2. BACKGROUND AND RELATED WORK

2.1. Online Spreading of Misinformation

In this section, we discuss how social networks increase the spread of biased news and misinformation. We discuss confirmation bias, echo chambers and other factors that may subconsciously influence a person's opinion. We show how these processes can interact to form a vicious circle that favors the rise of untrustworthy sources.

Often, when an individual thinks they know something, they are satisfied by an explanation that confirms their belief, without necessarily considering all possible other explanations, and regardless of the veracity of this information. This is **confirmation bias** in action. Nickerson (1998) defined it as the tendency of people to both seek and interpret evidence that supports an already-held belief.

An **echo chamber** is a situation in which an individual can only hear echoes of things that have already been said (Garimella et al., 2018). Social networks such as Twitter and Facebook are environments that favor the creation of such chambers (Knobloch-Westerwick and Kleinman, 2012). People tend to mix with others who think like them and follow news sources that they favor. In so doing, they expose themselves to limited framing of events that obscures other perspectives for them.

Consider a user with a hard-line political belief on either side of the political spectrum. They may follow only people and news sources who share that belief. It is likely that upon publishing a tweet about a new policy or event, they would see similar tweets from their friends and receive feedback that favors their own opinion. The echo chamber around the user shelters them against conflicting opinions. The 2016 American presidential election illustrates this phenomenon very well. Donald Trump's victory came as a surprise to many people worldwide. One explanation of this surprise is that voters on either side of the political spectrum were enclosed in echo chambers.

Research and having a critical approach to information shared online can protect a user against biased views, but very few protections exist against the creation of echo chambers. People can learn to identify them, but to avoid them completely requires them to ensure that all opinions are represented within their social circle.

Social networks extensively use recommender systems algorithms for selecting the content that appears in the feeds of users (Chakraborty et al., 2016). The reason is simple: the amount of content being created is too large for any single person to keep track of. Also, social networks want to improve the user experience by displaying content that the user will appreciate. This only exacerbates the problems discussed as it implies that users are grouped into clusters of preferences and provided with filtered content.

These recommender systems rely mostly on artificial intelligence to decide which content is best for a particular user (Ricci et al., 2011). Whether they are based on content-based filtering, on collaborative filtering, or on hybrid models, they tend to provide users with more content similar to that already seen and deemed relevant by and for similar people—thus enabling confirmation bias and feeding echo chambers.

Indeed, this is a key part of the functionality of the platforms: users are provided with content that they will like by restricting material that may not encourage further interaction.

These phenomena together can create a vicious circle. Echo chambers arise from both the user's subconscious choice of surrounding themselves with like-minded people and the enticement by content presented by recommender systems. Viewing a limited framing of content further increases the confirmation bias that what they believe is right. Finally, when users respond to articles that they "like," they close the loop by feeding the recommender algorithms that provide them with content.

2.2. Approaches to Detecting and Fighting Fake News

Lazer et al. (2018) argue that a scientific approach is required to find a solution to fake news in social media. Homogeneous social networks allow polarization and closure to new information. Consequently, echo chambers can form because of the personalization of political information. An additional reason for their formation is linked to both human behavior and the technical foundations of the user experience.

Despite the intellectual high ground taken by fact checkers such as PolitiFact[1] and Snopes[2], they do not solve the issue that is the tendency of individuals not to question the veracity of sources unless their own values or beliefs are infringed. This suggests that it is unlikely that a user would actively engage in the fact checking process and use the services provided by these fact checkers. Instead, the authors argue that it is the responsibility of platforms to include signals as to the quality of a source or

[1]https://www.politifact.com/
[2]https://www.snopes.com/

article within their algorithm, for example the prioritization of reputable sources in the news feed. However, this does not solve misinformation or ensure that conflicting views are available to the user. Methods for addressing these problems are still lacking.

2.3. Digital Nudges

The day-to-day definition of nudge as defined by Thaler and Sunstein (2008) is "*to push mildly or poke gently in the ribs, especially with the elbow*" or, applied to an economical context, "*self-consciously [attempt to] move people in directions that will make their lives better*." In a digital world, the definition is no different: the idea is to influence someone's behavior into acting in such a way that will improve his or her user experience and/or choices.

Lazer et al. (2018) cite nudges as a reasonable solution to the problem laid out in section 2.2. If the reading of news on social media platforms without investigating alternatives is re-framed as a choice for belief without validation, it is possible to define an architecture around this choice. Thus, it is possible to adapt this architecture through implementation of a nudge.

Several researchers have put effort into understanding the impact of nudges in social media, including Acquisti et al. (2017) and Wang et al. (2014), who have considered nudges to encourage user awareness of privacy and the impact of posts on platforms. Acquisti et al. (2017) discuss nudging by means of information, presentation, defaults, incentives, reversibility, and timing. We summarize two of these strategies for nudges here: nudging with information and presentation.

Nudging with information involves providing information to raise awareness. For example, in the context of fake news, this may include giving a label or signal about the reputability or the political leaning of a source.

Nudging with presentation involves the framing and structure of a choice. In the context of reading news in social media, an example could be the placement of an article in relation to the story from across the political spectrum.

3. BALANCEDVIEW: AN APPROACH TO MITIGATE ONLINE BIAS AND MISINFORMATION

In the context of fighting confirmation bias and fake news in the Twitter news feed, several approaches can be imagined, e.g., removing all suspicious posts. Another example would be to not allow users to post political views that are judged to be too extreme. This second example reduces the platform's usability. Instead, a solution must be more subtle and not restrict a user from posting or reading any particular post. In the present section, we give an overview of our approach.

We propose the approach and tool *BalancedView*[3] that aims to encourage users to consider the wider view surrounding information. In a first proposal, we will focus on tweets from the well-known social platform Twitter. We aim to implement a tool that efficiently presents a full view on articles from relevant sources presenting opinions from everywhere in the political spectrum. Practically, a user would input a tweet and be shown articles from trustworthy sources reporting on the same topic but with different opinions.

By doing so, a user is given the opportunity to forge their own opinion by reading from multiple sources. They can then make an informed decision on whether to believe an article based on presented alternatives. The proposed nudge is equivalent to placing the healthier bananas at eye level alongside an unhealthier option. The aim of the nudge is to ensure that a reader of a post is not restricted to reading the original content and is instead given a balanced view of the information based on sound journalism. Rather than restrict content and usability, we place a balanced and reputable selection of news sources at eye level to a news item.

A user can input a tweet to the tool, which then extracts the relevant text to structure a query to an API of news sources. Afterwards, the user is presented with the alternative framing of the same information. In further work, this system will be embedded into the user experience within Twitter.

This choice architecture corresponds to the "nudging by presentation" strategy of Acquisti et al. (2017) (see the overview in section 2.3). We also compared this with the strategy of nudging by information, which is closer to the approach currently taken by Twitter itself. The design of these two nudges will be described in the next section. Section 5 will then detail the back-end processing that identifies the appropriate news articles with alternative framing.

4. TWO NUDGING STRATEGIES

The desired outcome for a user should be an increased awareness of the potential political bias in an article. Subsequently, this should bring about assessment of evidence and consideration for how bias may compromise the veracity of an article.

We focus on tweets posted on Twitter and discuss two approaches, nudging by presentation and nudging by information as proposed by Acquisti et al. (2017). In the former case, the user is directly presented with information that might affect their judgement. In the latter case, a visual cue is displayed that gives the user an idea about the veracity of information. Both approaches follow the development process for digital nudges proposed by Mirsch et al. (2017):

- **Define**: The context is defined as the news feed of a social media platform. In the environment, only one-sided opinions are visible in the personalized sources chosen by a user. The goal is to ensure that at all times, without restricting a user from viewing the original content, the user is encouraged to view a balanced representation of opinions on a subject.
- **Diagnose**: In understanding the decision process, a number of questions can be identified that would ideally be asked by any reader of news such that a reasonable investigation of reputability and veracity of source or story is made. That is, given the set of questions that a professional fact checker would ask, is there a change in the choice architecture that would encourage a non-fact checker to ask similar questions.

[3]https://fact-checker.herokuapp.com. The tool is currently hosted on a free server meaning a slight delay in the initial start up.

FIGURE 1 | Placement of the nudge relative to original text.

- **Select**: For the scope of this work, we selected a nudge by presentation and a nudge by information. These strategies are explained in detail below.
- **Implement**: The nudges should be embedded in the social media platform on which people read news, but the exact HCI choices should be designed on the basis of a formative evaluation. We therefore implemented a mock-up Twitter interface for the user test. In addition, an emulated environment with a web front end was created for testing the natural language processing required for the nudge. More information on back-end processing for the two nudging strategies is given in the following two subsections.
- **Measure**: The nudge was evaluated by means of an initial user study. In a survey, users were asked to rate an article based on perceived levels of truth and reputability, in the presence and absence of nudges. This is discussed in section 6.

4.1. Nudging by Presentation

The primary aim of the nudge is to present an unbiased view of a subject, without necessarily forcing a user to embrace it. The secondary aim, of equal importance to the first, is to ensure that the sources presented are of a sufficient level of reputability: even if occasionally headlines are sensationalized, the underlying article will not be entirely fictitious or propagandistic.

4.1.1. Select the Appropriate Nudge

The objective is to present a user with alternative information that should encourage judgement of the veracity of a news article.

4.1.2. Implementation

Natural language processing is used to extract meaning from a tweet, and an API of news sources, NewsAPI [4], is queried. These results are sorted by relevance and presented to the user. This is described from a high-level perspective in section 3 and discussed in more detail in section 5.

4.1.3. Presenting the Nudge

The alternative news sources are displayed directly below the original content. This does not restrict the user from reading the original content but achieves the purpose of placing the alternative view at eye level. This is shown in **Figure 1**.

[4]https://newsapi.org

TABLE 1 | A selection of trustworthy news providers.

Left	Center	Right
	Reuters	
The Guardian	The Financial Times	The Telegraph
Independent	BBC News	The Daily Mail
MSNBC	The Wall Street Journal	Fox News
Politico	CNN	
	Bloomberg	

4.1.4. Selecting News Sources

A key aspect to discuss is how the sources are selected. The perceived political affiliation of news sources is identified through reports from Pew Research (Mitchell et al., 2014) and YouGov (Smith, 2017). Based on their findings, we chose the news sources shown in **Table 1** for use in the tool.

4.2. Nudging by Information

The second approach aims to provide information to raise awareness. For example, in the context of fake news, this may include giving a label or signal about the reputability of a source or its political bias. Twitter has implemented this to some extent by classifying some accounts as "verified." The existing Twitter flag for verified accounts can be regarded as a nudge towards trusting a source. Building on this format familiar to Twitter users, we have designed a nudge to encourage users to question a source. This nudge consists of a small white cross surrounded by a red background. It does not necessarily suggest bias or lack of reputability but it is the antithesis of the current nudge. In the study reported in section 6, we tested only the more well-known Twitter flag for verified accounts.

5. ANALYZING AN ARTICLE AND IDENTIFYING ALTERNATIVES WITH DIFFERENT FRAMING

In this section, we describe the back-end behind BalancedView's nudging by presentation.

5.1. High-Level Description

When a user inputs a tweet, the system first extracts and summarizes into relevant keywords the information contained in the text using the TextRank algorithm (Mihalcea and Tarau, 2004). With the keywords, the system builds a query to search for articles using the NewsAPI. Articles from multiple sources are then displayed on the screen ranging from left-most to right-most view. The selection of news providers is discussed in section 4.

5.2. Overview

The system takes a text as input and displays a series of articles, sorted by relevance and by political affiliation. We have separated this process into three main steps: summarizing the input, querying the news providers, and displaying the results by categories.

5.2.1. Summarizing the Input

In order to be able to query news providers, it is necessary to summarize the input and extract only the keywords. Among the relevant algorithms, TextRank and its variants provide a simple method based on a strong theoretical ground (Mihalcea and Tarau, 2004; Barrios et al., 2016). This algorithm performs unsupervised identification of centrality of text, using pre-trained models for the low-level tasks like part-of-speech tagging and stemming, as well as graph-based models for the identification of relevant entities.

When the algorithm receives an input, it tokenises the text and removes stop words, numbers and punctuation as well as Twitter-specific keywords such as hash tags and user mentions. The remaining words go through a part-of-speech filter and only the nouns, adjectives and verbs are kept. Porter's stemmer (Porter, 1980) is then used to generalize the words further.

From there, the algorithm builds a graph where each token is a node and the edges represent the relations between them. An edge between two words denotes that these two words follow each other in the text. A scoring function assigns scores to each node based on the nodes that are reachable from the first word of the input text. In other words, any words for which a path can be found from the starting node will have a high score. Consequently, words that occur repeatedly or that occur after such repeated words are more likely to have a high score and words that occur only once at the end of the input will have a low score.

Next, the keywords are sorted by decreasing score and the three to five best keywords are kept for the next step. The selection is based on a minimum score of 10%. Both the optimal number of keywords and the minimum score were empirically selected based on the quality and quantity of results after querying the source providers. The whole process described above is depicted in **Figure 2**.

For the proof-of-concept, we deliberately chose this simple method for the initial testing of the approach set forward in this work. In the future it should be improved to increase robustness to the shorter text lengths used on Twitter and other platforms.

5.2.2. Querying the News Providers

Having identified the keywords, a query is built and sent to NewsAPI. This service allows us to query a plethora of sources at the same time and get results from a number of countries in multiple languages. However, the free version does not enable going back more than one month in the past, which limits the number of results. The sources selection is explained in section 4 and the sources are listed in **Table 1**.

5.2.3. Displaying the Results

As we have discussed in the previous sections, the nudge must be subtle and cannot overload the user with too much information. Consequently, the design must be clean: the two most relevant articles for each political affiliation are included and only an abstract of the articles is displayed, together with a photo when one is available.

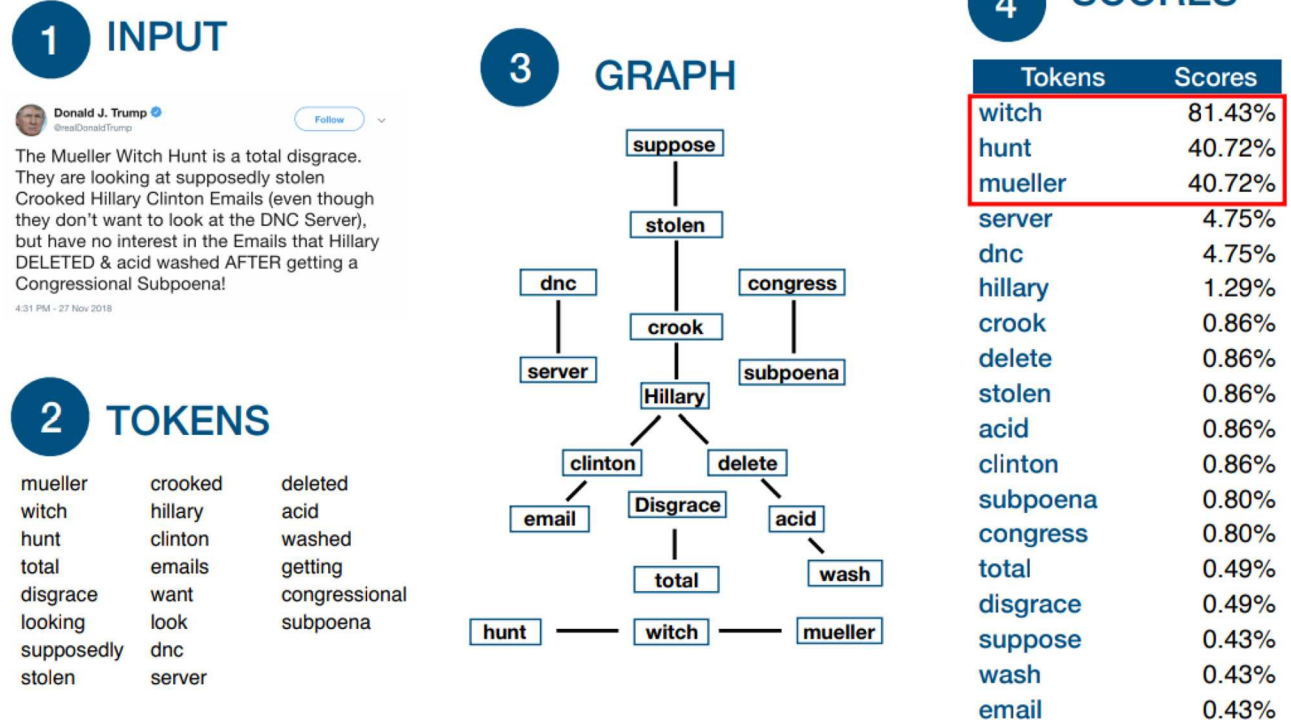

FIGURE 2 | Graph construction: (1) The input text from Twitter. (2) Token extraction based on the Porter stemmer (Porter, 1980). (3) A graph is constructed where nodes are words and edges denote whether two words are juxtaposed in the text. (4) Scores reflect whether a node is reachable from the start of the input text. If more than 5 words reach a score of 10%, only the best 5 keywords are selected. If less than 3 keywords reach this threshold, the 3 best keywords are selected.

6. EVALUATION

6.1. Relevance of the News Articles Presented

From a set of 35 tweets covering a number of stories in American and British politics, a query was built and evaluated. Such a test was deemed successful if at least two of the articles presented first in the results were considered relevant to the news surrounding the query.

Relevance was rated by the first two authors of the current paper. Their relevance ratings coincided in all 35 cases. Out of the 35 trials, all three articles were relevant eighteen times, two out of three were relevant eleven times, and in six cases, the system returned one or no relevant articles.

6.2. Effectiveness of the Nudge: User Study

We tested the usefulness of nudging by presentation and information in the context of perception of news. These experiments were made in survey form, in which participants were presented with tweets and asked to rate them on both impartiality and trustworthiness.

6.2.1. Method

We recruited twenty participants via an advertisement on our university's degree programme's Facebook page that contained a link to a survey. All participants were Master students of Artificial Intelligence, and they are regular users of social media including Twitter. No further demographic information was collected. Participation was voluntary and unpaid. Only aggregate results were retained.

We created a survey to test whether the nudge was effective in lowering trust in an intentionally selected politically biased tweet. Furthermore, the survey questioned whether the feature of a visual cue is useful in encouraging users to question reputability of a source.

The survey consisted of five individual web pages, each of which contained a screenshot of a tweet, enhanced (for questions 2, 3, and 5) by one of the two types of nudges tested, and a question regarding the trust in the news source or information.

For better readability, these questions are listed in third-person form as our research questions here; participants received a second-person "you" question. An example is shown in **Figure 1**. A PDF version of all survey questions is available as an online supplement to the current paper [5].

1. Do people trust obviously disreputable news sources in the absence of a nudge by information? Here a participant is presented with news from a disreputable news source on a news story that does not provoke an emotional response.

[5]https://people.cs.kuleuven.be/~bettina.berendt/BalancedView/BalancedView-Survey.pdf

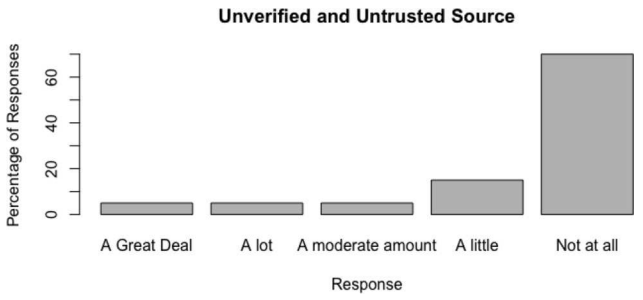

FIGURE 3 | Do people trust obviously untrustworthy news sources in the absence of a nudge by information?.

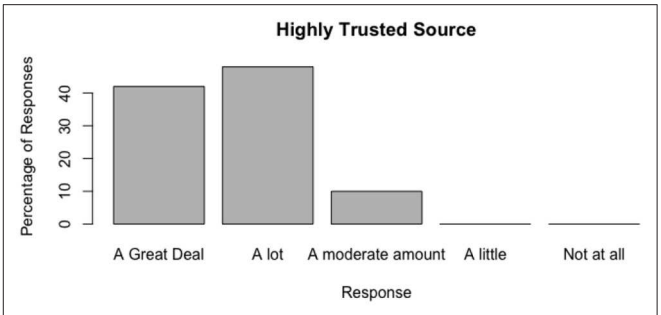

FIGURE 4 | Do people trust obviously reputable news sources?.

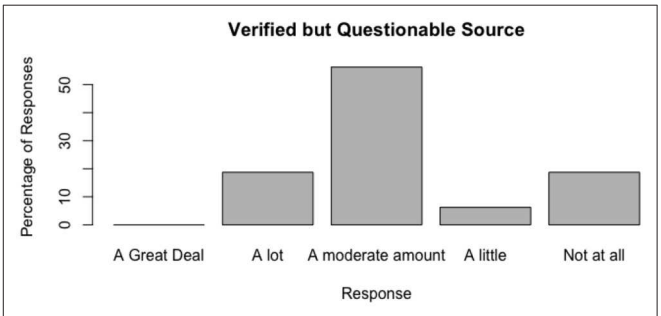

FIGURE 5 | Do people trust news sources of questionable reputability in the presence of a nudge by information?

2. Do people trust obviously reputable news sources in the presence of a nudge by information? A participant is presented with a story from a highly reputable news source, such as the BBC.
3. Do people trust news sources of questionable reputability in the presence of a nudge by information? A participant is presented with a story from a verified news source that is unlikely to be known as reputable or disreputable.
4. Do people consider politically biased information a fair representation of a view, in the total absence of nudges?
5. Do people consider politically biased information a fair representation of a view, given a nudge by presentation? The participant is presented with a politically biased statement and linked article, in the presence of the nudge designed in this work.

6.2.2. Results and Discussion

The distributions of responses are shown in **Figures 3–6**.

6.2.3. Do People Trust Obviously Untrustworthy News Sources in the Absence of a Nudge by Information?

Trust for the news source was generally low. Responses were concentrated in showing distrust or severe distrust of the news source. However, 15 percent of respondents placed moderate to high trust in the source despite no verification of the account.

6.2.4. Do People Trust Obviously Reputable News Sources?

There was a positive result for this test, people generally tended to trust or highly trust these sources. All respondents trusted the source moderately to highly.

6.2.5. Do People Trust News Sources of Questionable Reputability in the Presence of a Nudge by Information?

The results for this test were evenly spread between trusting and not trusting the source. The account was verified and the article featured was produced by a reputable news outlet. The spread of responses shows more trust than in the case of the obviously disreputable source, however, less trust is evident than in the case of the highly reputable source.

6.2.6. Do People Consider Politically Biased Information a Fair Representation of a View, Given a Nudge by Presentation? Do People Consider Politically Biased Information a Fair Representation of a View, in the Absence of a Nudge by Presentation?

The key test of the effectiveness of the balanced view nudge is the change in results for questions four and five. From the limited sample size, there is a visible shift in the responses to placing less trust in the singular view.

Initially, positive trust was placed in the fairness of the view being given. In the presence of the nudge, this opinion changed. In this case, results showed that people generally thought the view was unbalanced.

In sum, the results of this initial user study suggest that users generally recognisee obviously untrustworthy news sources, and that nudging by information may influence trust judgements less than a source's obvious reputability. There is evidence that the nudge by presentation, i.e., the central idea of BalancedView in which the user is offered a spectrum of diverse articles, helps participants question the trustworthiness of politically biased information.

Nonetheless, the survey questions need further development. The first questions in the current study were intended as a "sanity check" of intuitions about user trust in reputable and non-reputable sources, and about the basic workings of a nudge. The results support these intuitions and allow us to proceed to the more involved later questions. The results of the latter also show our user sample to be quite critical from the start, which may result in a ceiling effect in that nudges do not significantly change

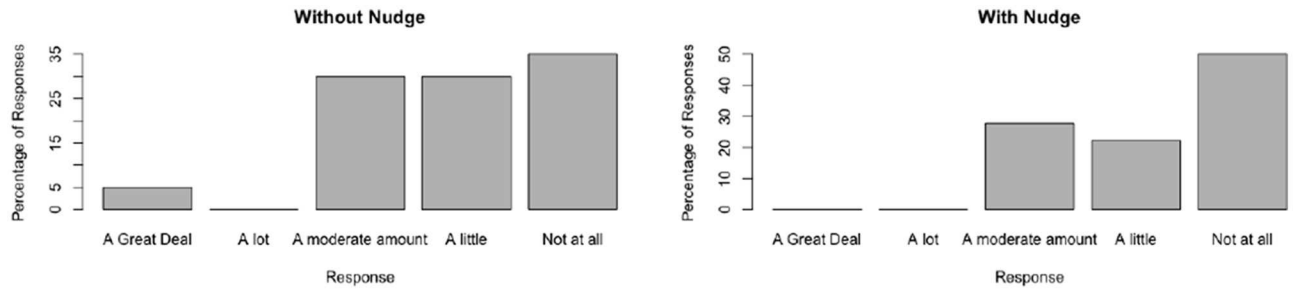

FIGURE 6 | Nudge comparison.

users' perceptions. In future work, more diverse groups of users should be drawn upon, such that the nudges' possible effects on their perceptions and actions become clearer.

7. SUMMARY, LIMITATIONS, AND FUTURE WORK

We have discussed nudges as a solution approach to the combined effects of confirmation bias and the algorithms of social media platforms that may create echo chambers and feedback loops of misinformation. This involves gently steering users towards adopting fact checking habits in their behavior online. Two nudging strategies were proposed: one that presents results in a way that pushes the user to look further and another that gives feedback on the quality of the posts that are shared online. The former option was implemented into an online tool that can be used to quickly browse articles relating to information expressed in a short text such as a tweet. The articles come from trustworthy news providers and are classified into political categories. In summary, the tool can be used to quickly and efficiently fact check any piece of information that one might read online.

In an initial user study, we investigated how questionable articles were perceived without any nudging strategy and with one of the two approaches discussed. The results suggest that the nudging strategies make people more aware of the trustworthiness of the sources. Furthermore, there is potential in presenting a balanced view of related news as a solution to lowering acceptance of a singular view.

These findings are encouraging. However, future work is needed to address a number of limitations:

- **Choice of methods and algorithms:** *BalancedView* in its current version uses relatively simple methods; our goal was to leverage the extensive toolbox of natural language processing and search algorithms for a new and timely purpose. We built this first version of our tool in order to establish a baseline from which to explore, in the future, different methods and algorithms with regard to their specific contributions to the task of countering confirmation bias.
- **Importance of the first words:** *BalancedView* gives more importance to words for which a path can be built starting at the first word of the graph. Consequently, the structure of the input tweet affects the relevance of the results.
- **Spelling and abbreviation:** The part-of-speech tagger used to identify relevant information is not robust to spelling errors and out-of-vocabulary words. This affects the relevance of the results as well.
- **Time-limited results:** The free version of the NewsAPI only returns results that are less than one month old. Consequently, texts referring to older events might not generate any results.
- **Limited number of sources:** The number of trusted sources should be increased, for example to reflect a wider range of political views. For this, there is a need for research in the field of political source trustworthiness.
- **Evaluation:** We have presented the results of a first relevance test and of an initial user study. Both evaluations were small-scale and need improvement along a number of dimensions. In particular, future studies should rest on larger sample sizes (both of article sets and of human participants) and experimental designs that allow for more fine-grained comparisons and contrasts between the choice architectures, and which take into account further factors such as demographics, as well as order effects.

In addition, extensions of the current approach are possible, including:

- **Multilingual Support:** Although this falls out of the scope of this project, we note that being able to not only search for tweets in any languages but also comparing information from different countries would be beneficial for the tool.
- **Deep Learning:** Recent developments in Deep Learning apply to text summarization as well as other of the limitations listed above and we think that using attention mechanisms and recurrent neural networks would help generate better results.
- **Fine-grained analysis of usage:** This can include recording interactions in the user experience, for example measuring how much time the users spend on the page and whether they still share and propagate unreliable news after having been in contact with *BalancedView*.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

CT: idea originator, first author and developer. QM: developer, second first author. CT and QM developed this as a project under the guidance of JP and BB, who worked as third and fourth authors to help refine the idea and write the paper in the later part of the project.

REFERENCES

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., et al. (2017). Nudges for privacy and security: understanding and assisting users' choices online. *ACM Comput. Surv.* 50:44. doi: 10.1145/3054926

Barrios, F., Lopez, F., Argerich, L., and Wachenchauzer, R. (2016). "Variations of the similarity function of TextRank for automated summarization," in *Argentine Symposium on Artificial Intelligence (ASAI 2015)*.

Chakraborty, A., Ghosh, S., Ganguly, N., and Gummadi, K. P. (2016). "Dissemination biases of social media channels: on the topical coverage of socially shared news," in *Tenth International AAAI Conference on Web and Social Media* (Cologne).

Cohen, M. (2018). *How Russian Trolls Manipulated Americanpolitics.* Available online at: https://edition.cnn.com/2018/10/19/politics/russian-troll-instructions/index.html (accessed October 20, 2018).

D'Angelo, P., and Kuypers, J. A. (2010). *Doing News Framing Analysis: Empirical and Theoretical Perspectives*. New York, NY: Routledge.

Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: echo chambers, gatekeepers, and the price of bipartisanship. *arXiv preprint* arXiv:1801.01665. doi: 10.1145/3178876.3186139

Knobloch-Westerwick, S., and Kleinman, S. B. (2012). Preelection selective exposure: confirmation bias versus informational utility. *Commun. Res.* 39, 170–193. doi: 10.1177/0093650211400597

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science* 359, 1094–1096. doi: 10.1126/science.aao2998

Lee, D. (2019), Feburary 2. Key fact-checkers top working with facebook. *BBC News*. Available online at: https://www.bbc.com/news/technology-47098021

Mihalcea, R., and Tarau, P. (2004). "TextRank: Bringing order into texts," in *Proceedings of EMNLP-04 and the 2004 Conference on Empirical Methods in Natural Language Processing* (Barcelona).

Mirsch, T., Lehrer, C., and Jung, R. (2017). "Digital nudging: altering user behavior in digital environments," in *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)* (St. Gallen), 634-648.

Mitchell, A., Gottfried, J., Kiley, J., and Matsa, K. E. (2014). *Political Polarization & Media Habits.* Washington, DC: Pew Research Center. Available online at: https://www.journalism.org/2014/10/21/political-polarization-media-habits/

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2:175.

Porter, M. (1980). An algorithm for suffix stripping. *Program* 14, 130-137.

Ricci, F., Rokach, L., and Shapira, B. (2011). *Introduction to Recommender Systems Handbook*. Boston, MA: Springer, 1–35.

Smith, M. (2017). *How Left or Right-Wing Are the UK's Newspapers?* Available online at: https://yougov.co.uk/topics/media/articles-reports/2017/03/07/how-left-or-right-wing-are-uks-newspapers

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness.* New Haven, CT; London: Yale University Press.

Wang, Y., Leon, P. G., Acquisti, A., Cranor, L. F., Forget, A., and Sadeh, N. (2014). "A field trial of privacy nudges for facebook," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (ACM)* (Toronto, ON), 2367-2376.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fdata.2019.00010

Discovering Topic-Oriented Highly Interactive Online Communities

*Swarna Das and Md Musfique Anwar**

Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh

Community detection is an interesting field of online social networks. Most existing approaches either consider common attributes of social network users or rely on only social connections among the users. However, not enough attention is paid to the degree of interactions among the community members in the retrieved communities, resulting in less interactive community members. This inactivity will create problems for many businesses as they require highly interactive users to efficiently advertise their marketing information. In this paper, we propose a model to detect topic-oriented densely-connected communities in which community members have active interactions among each other. We conduct experiments on a real dataset to demonstrate the effectiveness of our proposed approach.

Keywords: online social network, interaction strength, active community, query cohesiveness, structure cohesiveness

OPEN ACCESS

Edited by:
Roberto Interdonato, UMR9000 Territoires, Environnement, Télédétection et Information Spatiale (TETIS), France

Reviewed by:
Marinette Savonnet, Université de Bourgogne, France
Dino Ienco, National Research Institute of Science and Technology for Environment and Agriculture (IRSTEA), France

***Correspondence:**
Md Musfique Anwar
manwar@juniv.edu

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *01 April 2019*
Accepted: *20 May 2019*
Published: *06 June 2019*

Citation:
Das S and Anwar MM (2019) Discovering Topic-Oriented Highly Interactive Online Communities. Front. Big Data 2:10. doi: 10.3389/fdata.2019.00010

1. INTRODUCTION

Nowadays, Online Social Networks (OSN) are widely used by a large part of the general population. Similar interests, choices, and hobbies tend to form a group of users in a social network known as online community. There have been many attempts to detect these online communities for the purpose of business, marketing, recommendations, biological research, etc. Often the mere use of connection links does not provide an effective group of users. As a result, these groups do not bring efficient results.

There are two types of network topology. One is global, where information of a whole network is captured and another is local, i.e., a network that works with the similar nodes (Tang et al., 2017). There have been many approaches to detect communities and serve various other fields with it (Fortunato and Hric, 2016). An approach to detecting communities is Affinity Propagation, where the network is divided and a multiobjective evolutionary algorithm is introduced (Shang et al., 2016). For the purpose of local community formation, dynamic membership function can be used (Luo et al., 2018). Fuzzy relations can be used for non-overlapping community detection. The nearest node with each node's greater centrality and fuzzy relations are combined for the desired result (Luo et al., 2017).

Recent research works consider social users' topical interests in OSNs, e.g., (Yang et al., 2013), in order to find meaningful communities. However, these methods did not focus on the topical interactions among the community members. Therefore, such communities contain many members who have very inactive topical interactions among them which perform poorly in viral marketing. In order to avoid the inactivity problem authors (Lim and Datta, 2016) have proposed an approach where interaction pattern and frequency are considered rather than only counting the following/follower links.

Our observation is that social users have different degrees of topical intimacy among them. In this work, we propose an approach to discover topic-oriented *highly interactive* communities in OSNs, where the members in the community should have a certain degree of topical interactions with each other related to a given query. We also emphasize that the members in the retrieved communities should actively interact with at least k other members within the community. Below, we summarize our contributions:

- We propose a methodology to discover highly interactive online communities where community members have a high degree of interactions with each other on similar topics;
- We quantify the topical interaction strengths among the users;
- We perform experiments on a real dataset to demonstrate the effectiveness of the proposed method.

2. RELATED WORK

Earlier methods for community detection are based on structural information of the social graph such as modularity (Clauset et al., 2004), edge betweenness (Newman and Park, 2003), and neighborhood concepts (Cohen, 2008). Some approaches also considered the textual content published by the users along with social connections to detect like-minded users. For example, SA-Cluster applied random-walk to measure the closeness of a node in an augmented attributed graph (Zhou et al., 2009). A Topic-Link LDA model (Liu et al., 2009) is proposed which considers both the linkage structure and similarities of the contents of edges to detect communities. A probabilistic generative model named as CESNA is proposed by Yang et al. (2013) and combines community memberships, node attributes, and the network topology to find the communities.

More recently, some approaches have focused on the interaction strength between the users in order to find active communities. Dev et al. (2014) considered the impact of interaction between users as well as the impact of the group behavior without considering topical attributes of the nodes. Lim and Datta (2016) proposed the Highly Interactive Community Detection (HICD) method, which constructs a weighted network using the frequency of direct interactions between users. Correa et al. (2012) proposed the *iTop* algorithm, which constructs a weighted graph based on user interactions and maximizes the local modularity to detect topic-oriented communities based on a set of seed users. However, all these methods ignored topic-wise users' inter-activeness. Our goal is to discover communities where users have high interactions with others with regard to the given query consisting a set of topics.

3. METHODOLOGY

First we formally formulate the problem of discovering highly interactive topical communities in OSNs. Then we give an overview of our proposed approach.

Attributed Social Graph: An attributed social graph is denoted as $G = (U, E, \mathcal{A})$, where U represents the set of social users (nodes), E indicates the set of links (edges) between the users, and $\mathcal{A}=\{T_1, T_2, ..., T_m\}$ is the set of topics discussed by the social users in G.

In Twitter, users mention each other using "@." In order to construct a link (a, b) between users, @mention is used, i.e., $M_{a,b}$ denotes that user a has posted a tweet which contains @b.

***k*-Core:** Given an integer k ($k \geq 0$), the k-core of a graph G, denoted by C^k, is the maximal connected sub graph of G, such that $\forall u \in C^k$, $deg_{C^k}(u) \geq k$, where $deg_{C^k}(u)$ refers to the degree of a node u in C^k. A k-core component H_j^k is considered as a community from a structural point of view.

Node Core Number: The core number of a social user u in a k-core induced sub graph from G indicates the maximum k for which u belongs to that k-core sub graph.

Topic: A topic contains a set of related words that represents the topic. For example, the politics topic has words like election, vote, democracy, political party, etc.

Activity: Any action performed by a social user is referred to as an activity. For example, posting a new tweet or retweeting an existing tweet is considered as an activity. In our work, we consider only those actions that are performed between any two social users. For example, a user u in Twitter replies to a tweet posted by user v. This activity is recorded as an activity tuple $\langle u, v, \psi_{uv} \rangle$, where ψ_{uv} indicates the set of attributes (topics) exchanged between u and v (Anwar et al., 2018).

Query: An input query $Q=\{T_1, T_2..., T_n\}$ contains a set of query topics.

Active Interaction Edge: If any two social users u and v in G have a certain number of direct interactions ($\gamma(\geq 1)$) between them related to Q, then we consider the interaction link between those two users as an active interaction edge (e_{uv}). Factor w_{uv} indicates their involvement in direct interactions compared with the most active pair of users in the network.

$$w_{uv} = \frac{|ACTS(u, v, \psi_{uv})|}{max_{x,y \in U^Q} |ACTS(x, y, \psi_{xy})|} \quad (1)$$

where $ACTS(u, v, \psi_{uv})$ indicates the number of direct interactions between u and v containing $\psi_{uv} \subseteq Q$.

Active User: The users of an active interaction edge e_{uv} are considered as active users. The set of all the active users for a given query Q is denoted as U^Q.

3.1. Problem Definition

Given a graph $G = (U, E, \mathcal{A})$, an input query Q and an integer k, we first find the set of active edges between the social users by measuring interaction strength $w_{uv}(w_{uv} \in [0, 1])$. Then an induced sub graph H_j^k is considered as an active interactive community if it satisfies the following criteria.

1. ***Connectivity.*** $H_j^k \subset G$ is connected;
2. ***Structure cohesiveness.*** $\forall u \in H_j^k$ has interaction degree of at least k;
3. ***Active interaction.*** $\forall e_{uv} \in H_j^k$, the interaction strength of e_{uv} is $w_{uv} \geq \theta$ and $\theta \in [0, 1]$ is a threshold.

Figure 1 shows a social graph G with the core number for each node, e.g., the three-core nodes are {A,B,C,I}. **Table 1** represents

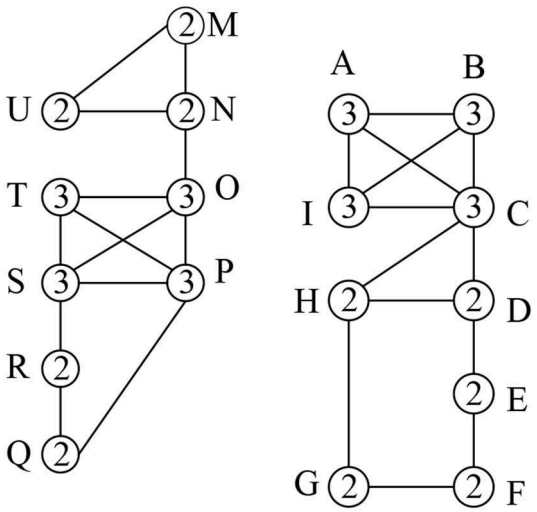

FIGURE 1 | Social Graph (the number denotes the node core number).

TABLE 1 | Interaction log.

(e_{ab})	T_1	T_2	(e_{ab})	T_1	T_2
(M,N)	6	18	(A,B)	20	7
(M,U)	3	14	(A,C)	18	6
(N,U)	5	15	(A,I)	14	6
(N,O)	6	8	(B,C)	10	13
(O,T)	20	13	(B,I)	13	8
(O,P)	19	7	(C,I)	15	9
(O,S)	11	6	(C,D)	7	12
(P,T)	14	8	(C,H)	6	17
(P,S)	18	9	(D,H)	12	9
(S,T)	16	9	(D,E)	7	18
(S,R)	12	4	(H,G)	5	8
(R,Q)	10	5	(E,F)	4	9
(P,Q)	20	9	(F,G)	6	20

the interaction frequencies among the users for topic T_1 and T_2. In **Table 2**, we show the interactive communities for a query $Q = \{T_1, T_2\}$. We get different community members for different values of Q, k, and θ. For example, when Q={T_1}, $k = \{2\}$, and $\theta = \{0.4\}$, we get H_1^2 = {A,B,C,I}, H_2^2 = {O,P,Q,R,S,T} while for the same values of Q and θ with an increase value of $k = \{3\}$, we get H_1^2 = {A,B,C,I}, H_2^2 = {O,P,S,T}. Again, for Q={T_1, T_2}, $k = \{2\}$ and $\theta = \{0.5\}$, we get H_1^2 = {A,B,C,D,H,I}, H_2^2 = {O,P,Q,R,S,T} and H_3^2 = {M,N,U}

3.2. Highly Interactive Community Detection Approach

In this work, we propose a method to detect highly interactive communities for a given a query Q in an online social attributed graph G. The desired communities from the graph G can be identified in the following three steps:

TABLE 2 | Community members for different values of Q, k, and θ.

Query	Community
Q={T_1}, k={2}, θ={0.4}	H_1^2={A,B,C,I}, H_2^2={O,P,Q,R,S,T}
Q={T_1}, k={3}, θ={0.4}	H_1^3={A,B,C,I},H_2^3={O,P,S,T}
Q={T_2}, k={2}, θ={0.4}	H_1^2={B,C,D,E,F,G,H,I},H_2^2={M,N,O,P,S,T,U}
Q={T_2}, k={2}, θ={0.5}	H_1^2={C,D,H},H_2^2={M,N,U}
Q={T_1,T_2}, k={2}, θ={0.4}	H_1^2={A,B,C,D,E,F,G,H,I},H_2^2={M,N,O,P,Q,R,S,T,U}
Q={T_1,T_2}, k={2}, θ={0.5}	H_1^2={A,B,C,D,H,I},H_2^2={O,P,Q,R,S,T},H_3^2={M,N,U}
Q={T_1,T_2}, k={3}, θ={0.4}	H_1^3={A,B,C,I},H_2^3={O,P,S,T}

Algorithm 1 `Query Algorithm`

Input: G=(U, E), Q, k, θ
Output: set of active interactive communities Φ_Q={H_1^k, H_2^k, ..., H_n^k}

1: **for** each $(u, v) \in E$ **do**
2: compute w_{uv}
3: **if** $w_{uv} > \theta$ **then**
4: $U^Q.add(u)$
5: $U^Q.add(v)$
6: compute the induced graph G^Q on U^Q
7: compute the maximal k-core $C^k(G^Q)$ of G^Q
8: Output the set of active connected components Φ_Q from $C^k(G^Q)$

1. Identify the set of active users based on their direct interaction with each other for a given query Q.
2. Refine the original social graph G by filtering the inactive social users.
3. Apply k-core technique on the refined social graph in order to detect the desired online communities.

The first step of our approach is measuring the interaction frequencies among the users for a given query Q in social graph G to filter the weakly connected topology links. For this purpose, we consider users who have direct communication with others via retweets or mentions and consider an interaction link between two users irrespective of whether they have a topology link or not.

After establishing the newly active interaction edges and filtering the inactive topology links from the social graph G, we apply k-core on the refined social graph to find the connected components in which every node has degree of at-least k.

We develop an algorithmic framework to detect highly interactive communities for a given Q.

Algorithm overview. The algorithm, called `Query Algorithm`, has three steps. First, it computes the interaction strength w_{uv} of each edge e_{uv} for a given query Q in order to find the set of active users (line 1-5). Next, we compute the induced sub graph G^Q from U^Q (line 6). Finally, we identify the maximal k-core of $C^k(G^Q)$ from the induced graph G^Q to find the set of active connected components (i.e., desired connected communities) (line 6-7) Φ^Q from $C^k(G^Q)$ (line 7-8).

4. EXPERIMENT AND RESULT

We conduct our experiment on an academic coauthor (DBLP) dataset (Jie et al., 2008) and choose research papers that were published within 2005 to 2011. This revised dataset is a

network of 15,516 authors with 48,862 co-author relationships between these authors and contains 193,512 research papers. The co-author information in DBLP is considered as interaction between the authors. We extract the authors' details, publication year, and abstract from each research paper. We apply latent dirichlet allocation (LDA) topic modeling (Blei et al., 2003) on the abstracts of the research papers in order to find the research topics.

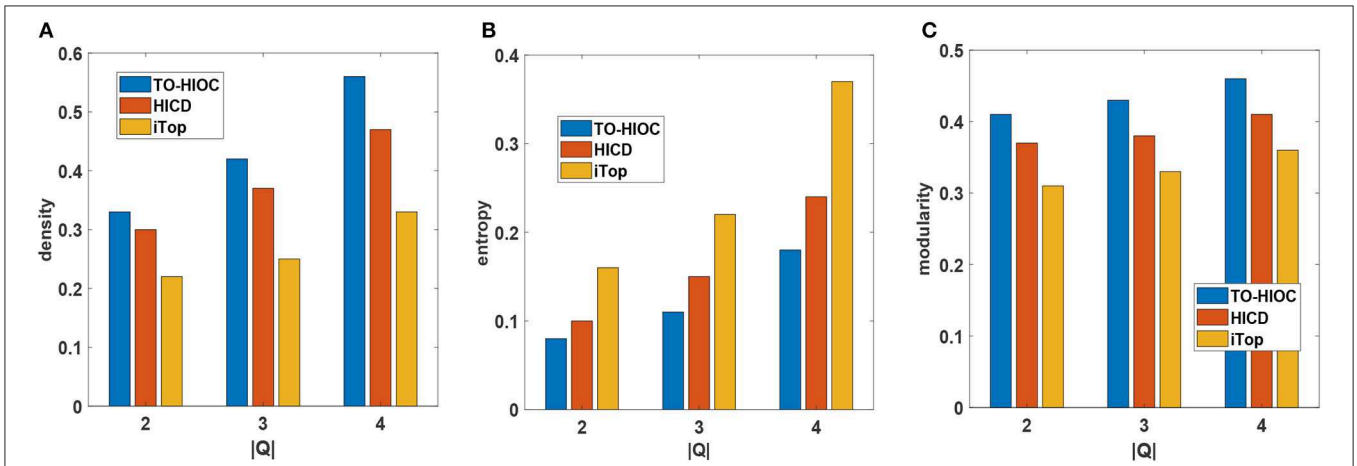

FIGURE 2 | Performance comparison on DBLP dataset **(A)** Density, **(B)** Entropy, **(C)** Modularity (in all cases, Q = {Semantic web, Data mining, Social network analysis}, k = {4}, θ = {0.5}, γ = {4}, the publications are chosen from the time period of 2005 to 2009).

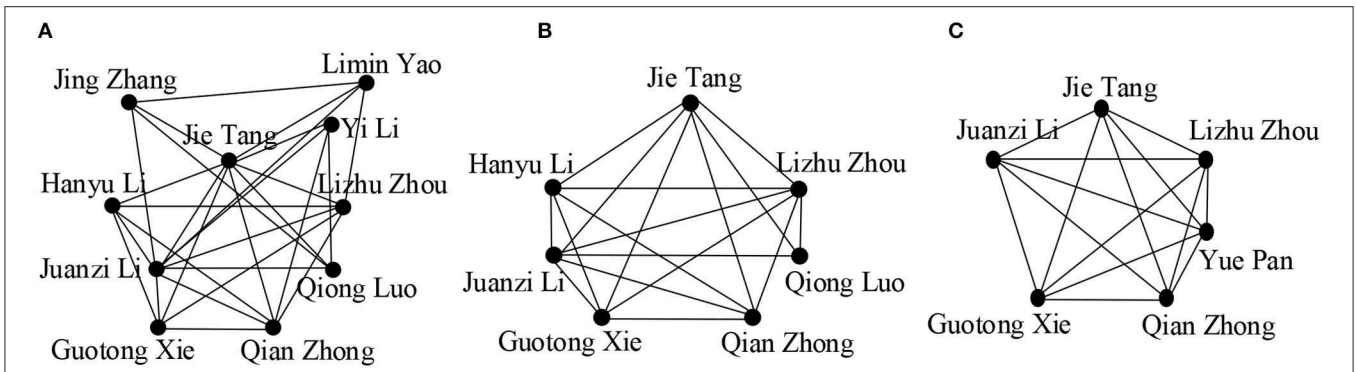

FIGURE 3 | (A) k = {3}; **(B)** k = {4}; **(C)** k = {5}; (in all cases, Q = {Semantic web}, θ = {0.5}, γ = {4}, the publications are chosen from the time period of 2005 to 2009).

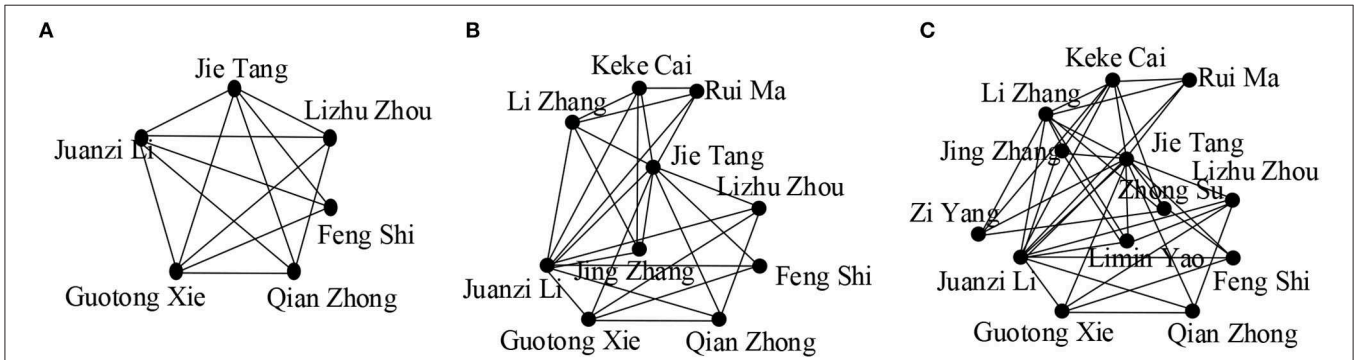

FIGURE 4 | (A) Q = {Semantic web}; **(B)** Q = {Semantic web, Data mining}; **(C)** Q = {Semantic web, Data mining, Social network analysis}; (in all cases, k = {4}, θ = {0.5}, γ = {4}, the publications are chosen from the time period of 2007 to 2011).

Comparison Methods. We compare our Algorithm 1 (`Query Algorithm`), denoted here as `TO-HIOC`, with two other existing methods: `HICD` method (Lim and Datta, 2016) and `iTop` algorithm (Correa et al., 2012).

Evaluation Measures. We vary the length of the Q to $|Q| = 2, 3, 4$ and use three measures of density, entropy and modularity to evaluate the quality of the detected online communities discovered by different methods. The definition of density, entropy, and modularity are as follows.

$$density(\{H_j^k\}_{j=1}^n) = \sum_j^n \frac{|\{(u,v)|u,v \in H_j^k, (u,v) \in E\}|}{|E|} \quad (2)$$

where n denotes the total number of detected communities. Density measures the compactness of the communities in structure.

$$entropy(\{H_j^k\}_{j=1}^n) = \sum_j^n \frac{|U(H_j^k)|}{|U|} entropy(H_j^k), \text{where}$$

$$entropy(H_j^k) = -\sum_{i=1}^n p_{ij} log_2 p_{ij} \quad (3)$$

and p_{ij} is the percentage of members in a community $\mathcal{G}_j$ who are active on the query topic T_i. $entropy(\{\mathcal{G}_j\}_{j=1}^n)$ measures the weighted entropy considering all the query topics over all the communities. Entropy indicates the randomness of the topics which are covered in the communities.

$$modularity(\{H_j^k\}_{j=1}^n) = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{d_i d_j}{2m}] \delta(s_i, s_j) \quad (4)$$

Here, m denotes the number of edges corresponding to an adjacency matrix A[1], d_i denotes the degree corresponding to node n_i, s_i denotes the community membership of node n_i and $\delta(s_i, s_j) = 1$ if s_i=s_j.

Generally, a good interactive community should have high density, high modularity, and low entropy.

Figure 2A shows the density comparison between all the methods on the DBLP dataset. We set $k = 4$ as there are usually many small-sized research groups existing in DBLP. We see that `TO-HIOC` achieves better performance compared to the other two methods because it considers query-oriented active interactions among the community members. The `HICD` method fails to achieve better density values as it requires interaction between users (authors) to the celebrities (i.e., very high profile researchers in DBLP), which is not very common. The `iTop` method ignores the interactions between the non-seed users, resulting in poor performance. We also observe that all the methods achieve better density values for higher values of $|Q|$. The reason is that the number of interactive connections of the users increases as $|Q|$ increases, which results in large and more densely connected communities.

Figure 2B shows the entropy comparison between the three methods. `TO-HIOC` achieves better performance in the aspect of the entropy as it considers the topical relevance (with regard to the query topics) during the interactions between the authors while forming a community. On the other hand, `HICD` achieves higher entropy value because not all the connected authors in a community have interests or active interactions in the common research topics. `iTop` also achieves a higher entropy value due to the lack of active topical interactions between the seed users and their followers. We see in **Figure 2C** that our proposed method `TO-HIOC` outperforms `HICD` and `iTop` in modularity comparison.

We examined a community in a co-author dataset which includes Jie Tang, who is one of the leading researchers in the data mining area, to see the differences in the community members for different values of k and Q ={semantic web, topic mining, social network analysis}.

We observe the effect of value k in **Figures 3A–C**. By varying the values of k, we get communities of different sizes. We see that the community size decreases for higher values of k as the cohesiveness constraint becomes more strict, resulting in the exclusion of some active community members, for example "Yi Li," "Jing Zhang," "Limin Yao" leave the group. We also see that more researchers joined the community when the length of Q is increased as higher values of $|Q|$ covered more interactive researchers (**Figures 4A–C**).

5. CONCLUSION

In this paper, a topic-oriented highly interactive community detection approach is proposed. This method detects global communities where users have active interaction with each other on common topics. We observed that users have different degrees of interactions for different topics. As future work, we will consider the temporal factor to measure the recency behavior of users' interactions.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

For this research paper, SD conducted the experiments (100%), designed the algorithm (75%), and wrote the paper (70%). MA designed the algorithm and experiments (25%), revised the paper (30%) as well as provided helpful insights and contribution in

[1] An adjacency matrix of a network is represented by A, where $A_{uv} = 0$ means there is no edge (no interaction) between nodes u and v and A_{uv}=1 means there is an edge between the two.

REFERENCES

Anwar, M. M., Liu, C., and Li, J. (2018). "Uncovering attribute-driven active intimate communities," in *Australasian Database Conference* (Gold Coast, QLD: Springer), 109–122.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E* 70:066111. doi: 10.1103/PhysRevE.70.066111

Cohen, J. (2008). Trusses: Cohesive subgraphs for social network analysis. *Natl. Secur. Agency Tech. Rep.* 16, 3–1. doi: 10.1.1.505.7006

Correa, D., Sureka, A., and Pundir, M. (2012). "itop: interaction based topic centric community discovery on twitter," in *Proceedings of the 5th Ph. D. Workshop on Information and Knowledge* (Maui, HI: ACM), 51–58.

Dev, H., Ali, M. E., and Hashem, T. (2014). "User interaction based community detection in online social networks," in *International Conference on Database Systems for Advanced Applications* (Bali: Springer), 296–310.

Fortunato, S., and Hric, D. (2016). Community detection in networks: a user guide. *Phys. Rep.* 659, 1–44. doi: 10.1016/j.physrep.2016.09.002

Jie, T., Jing, Z., Limin, Y., Juanzi, L., Li, Z., and Zhong, S. (2008). "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing: ACM), 990–998.

Lim, K. H., and Datta, A. (2016). An interaction-based approach to detecting highly interactive twitter communities using tweeting links. *Web Intell.* 14, 1–15. doi: 10.3233/WEB-160328

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). "Topic-link lda: joint models of topic and author community," In *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, QC: ACM) , 665–672.

Luo, W., Yan, Z., Bu, C., and Zhang, D. (2017). Community detection by fuzzy relations. *IEEE Trans. Emerg. Top. Comput.* 1–14. doi: 10.1109/TETC.2017.2751101

Luo, W., Zhang, D., Jiang, H., Ni, L., and Hu, Y. (2018). Local community detection with the dynamic membership function. *IEEE Trans. Fuzzy Syst.* 26, 3136–3150. doi: 10.1109/TFUZZ.2018.2812148

Newman, M. E., and Park, J. (2003). Why social networks are different from other types of networks. *Phys. Rev. E* 68:036122. doi: 10.1103/PhysRevE.68.036122

Shang, R., Luo, S., Zhang, W., Stolkin, R., and Jiao, L. (2016). A multiobjective evolutionary algorithm to find community structures based on affinity propagation. *Phys. A Stat. Mech. Appl.* 453, 203–227. doi: 10.1016/j.physa.2016.02.020

Tang, X., Xu, T., Feng, X., Yang, G., Wang, J., Li, Q., et al. (2017). Learning community structures: global and local perspectives. *Neurocomputing* 239, 249–256. doi: 10.1016/j.neucom.2017.02.026

Yang, J., McAuley, J., and Leskovec, J. (2013). "Community detection in networks with node attributes," in *2013 IEEE 13th International Conference on Data Mining* (Dallas, TX: IEEE), 1151–1156.

Zhou, Y., Cheng, H., and Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow.* 2, 718–729. doi: 10.14778/1687627.1687709

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fdata.2019.00009

An Innovative Way to Model Twitter Topic-Driven Interactions Using Multiplex Networks

Obaida Hanteer and Luca Rossi**

Data Science and Society Lab, IT University of Copenhagen, Copenhagen, Denmark

We propose a way to model topic-based implicit interactions among Twitter users. Our model relies on grouping Twitter hashtags, in a given context, into themes/topics and then using the multiplex network model to construct a thematic multiplex where each layer corresponds to a topic/theme, and users within a layer are connected if and only if they used the same hashtag. We show, by testing our model on a real-world Twitter dataset, that applying multiplex community detection on the thematic multiplex can reveal new types of communities that were not observed before using the traditional ways of modeling Twitter interactions.

Keywords: multiplex networks, multiplex community detection, thematic communities, thematic clusters, thematic multiplex, social network analysis, social media data analysis

OPEN ACCESS

Edited by:
Roberto Interdonato, Télédétection et Information Spatiale (TETIS), France

Reviewed by:
Sabrina Gaito, University of Milan, Italy
Antonio Caliò, University of Calabria, Italy

****Correspondence:***
Obaida Hanteer
obha@itu.dk
Luca Rossi
lucr@itu.dk

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *01 April 2019*
Accepted: *14 May 2019*
Published: *06 June 2019*

Citation:
Hanteer O and Rossi L (2019) An Innovative Way to Model Twitter Topic-Driven Interactions Using Multiplex Networks. Front. Big Data 2:9. doi: 10.3389/fdata.2019.00009

1. INTRODUCTION

The unprecedented amount of data that is produced, on a daily base, on social media has provided to researchers and practitioners a new opportunity to study, in depth, complex social dynamics at a large scale. Within this context, Twitter can easily claim the award for the most researched social media platform. Thanks to the large user-base and a relatively generous API policy, this micro-blogging platform has quickly evolved into the *de-facto* standard platform for multiple studies on social media dynamics.

The detection of cohesive subgroups in social networks, also called as community detection, has been perceived as one of the most valuable tools to better understand social networks (Papadopoulos et al., 2012). Given that members of the same community tend to share some properties, the community structure of a network can provide a better understanding of the overall functioning of this network. The application of this on social media data has provided useful insights about some of the dynamics and phenomena that take place in such systems (Silva et al., 2017).

A common approach to model Twitter interactions for community detection tasks is to build a network based on following/follower relations (Kwak et al., 2010), or networks based on either retweets (Conover et al., 2011) or explicit mentions indicated by the @ character (Yang and Counts, 2010). Advances on multiplex community detection have suggested that looking at more than one of these types of connections together can provide some insights that cannot be observed by looking at each of them separately. As to the content generated by Twitter users, it has been mostly used for topic detection tasks (Ibrahim et al., 2018) and sentiment analysis (Ceron et al., 2014). To the best of our knowledge, no previous work has addressed extracting network-like information from the content generated by users on social media platforms for community detection tasks.

Much of Twitter contemporary interactions happen in the form of conversations in many-to-many polyadic spaces defined by hashtags (Bruns and Burgess, 2011). In this type of

interactions, Twitter users are not necessarily retweeting, replying to, or mentioning each other but engaging directly with specific issues. This suggests that analyzing Twitter data by considering only the direct interactions among users (i.e., following/follower, retweet, and mention networks) is still far from providing a complete picture of Twitter-based interactions. In this paper, we address this gap by proposing an innovative way to model topic-driven interactions of Twitter users using the multiplex network model (Dickison et al., 2016). We test our model, the thematic multiplex, on a real-world dataset capturing the Twitter interactions of the Danish politicians during the parliamentary elections of 2015. We show that detecting communities on the thematic multiplex can reveal different dynamics than those observed by analyzing only explicit interactions. For example, we observed, using thematic multiplex community detection, that while some themes/topics were discussed by almost all the parties within the month leading to the election day, left and right-wing parties, at the same time, have also focused on themes that were politically closer to their traditional ideologies.

The rest of this paper is organized as follows. In section 2 we introduce the thematic multiplex and the thematic multiplex community detection. This is followed by our analysis of a real-world use case (section 3) which captures the Twitter interactions among Danish politicians during the parliamentary elections of 2015. We discuss our results in section 4 and conclude our findings in section 5

2. THE THEMATIC MULTIPLEX

On platforms like Twitter, when a user uses a specific hashtag in a tweet, he/she is not only increasing the visibility of that tweet, but also implicitly, even if not directly, communicating with other Twitter users who are using the same hashtag. This concept has been referred to as the *imagined audience* in the literature (Litt, 2012). Thus, we can assume a social tie (an edge) between two users who used the same hashtag and this is the main idea behind the thematic multiplex. The thematic multiplex, as the name suggests, is a multiplex network where each layer corresponds to a topic/theme and users within a layer are connected via a clique, if and only if, they used the same hashtag. An edge among two actors in the resulted thematic multiplex does not necessarily imply a direct interaction among them yet it suggests that they share a topical-interest. **Figure 1** illustrates a thematic multiplex where each layer represents a specific topic/theme (for example, refugees, education, etc..), and users who used the same hashtag within a topic are connected via a clique, which might result in multiple cliques within a layer (for example, the education theme). **Figure 2** illustrates a possible output for community detection on the thematic multiplex.

We claim that detecting communities on the thematic multiplex network using multiplex community detection can reveal different dynamics than those observed by analyzing the direct interactions among users. The reason is two folded: on one side, direct interactions are often driven by heterogeneous behavior from the users, e.g., Retweets can represent a form of endorsement or just a way to spread an information deemed to be relevant, Replies can equally be produced by amused conversations or endless fights between users. On the other side, direct interactions are just part of the whole Twitter data, thus any approach focusing solely on those will loose potentially relevant information. Thematic multiplex community detection, on the opposite, results in thematic communities were users are grouped together if they tend to discuss/be involved in the same topics/themes through direct or indirect interactions. More over, given that the qualitative analysis is added in the modeling phase, this intrinsically contributes to the qualitative power of community detection on the thematic multiplex network.

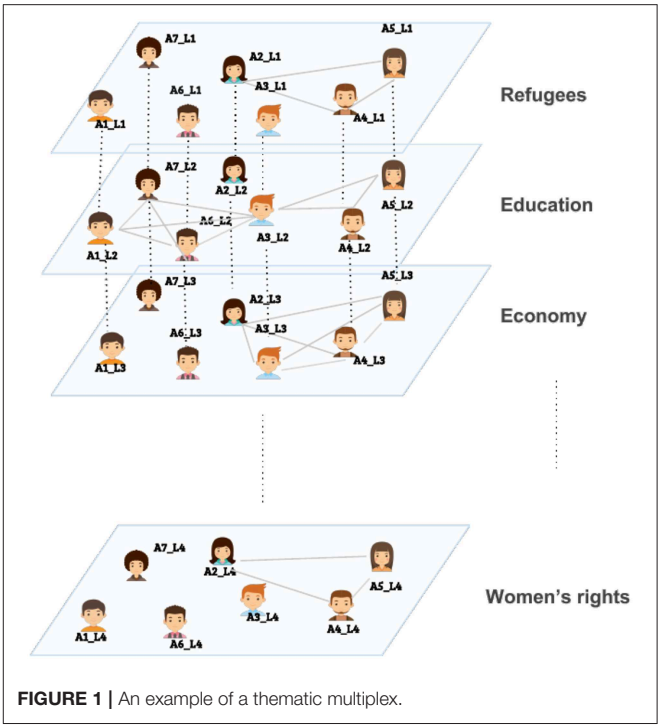

FIGURE 1 | An example of a thematic multiplex.

3. A CASE STUDY

We describe the dataset in section 3.1, then we discuss the construction of the correspondent thematic multiplex and some choices for our analysis tools in section 3.2. We report our observations on the results in section 3.3.

3.1. The DkPol Dataset

The data we use to test our model is collected during the month leading to the 2015 Danish parliamentary election. Starting from a list of all the Danish politicians running for the parliament who also had a Twitter account, we collected all the tweets written during the 30 days leading to the election. The initial dataset was formed by 490 politicians distributed across 10 parties, 5,985 original tweets, 633 replies, and 3,993 retweets. Together with their Twitter activity, we noted also the political affiliation of the 490 politicians. Given the complexity of the Danish multi-party system, the parties have also been

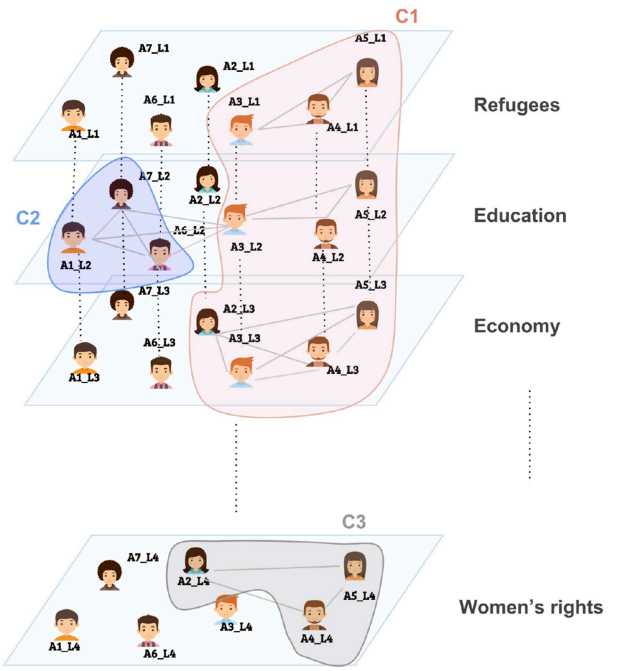

FIGURE 2 | A possible output for applying community detection on a thematic multiplex.

grouped into two main coalitions existing at the time: Red Block, currently at the opposition, and the Blue block, currently in government[1]. In order to use the hashtag contained in the tweets to build a thematic multiplex, some initial data cleaning was necessary. The hashtags were first qualitatively analyzed. We then excluded the hashtags that were just about the election campaign as such (like #dkpol) and those referring to political TV debates (like #tv2valg and #DRdinstemme). After this filtering we were left with only 23 hashtags used to refer to specific topics (12 topics). **Table 1** shows the grouping of these hashtags into topics. While our suggested grouping can be further discussed as hashtags can be grouped in many other ways, we chose to keep our focus on the correspondent thematic multiplex and the resulted communities for the sake of this paper.

3.2. Experimental Settings

Given the DkPol dataset, we constructed a twelve-layer thematic multiplex (layer per theme/topic). A topic/theme with k hashtags is interpreted as k cliques in the correspondent layer (a clique per hashtag) among all the users who used the same hashtag. We first show that detecting communities on the thematic multiplex reveals communities that are largely different from those detected using the traditional ways of modeling twitter

[1]The red block coalition groups the following parties: Alternativet, Radikale Venstre, Enhedslisten, Socialdemokratiet, and Socialistisk Folkeparti, while the blue block coalition groups: Dansk Folkeparti, KristenDemokraterne, Liberal Alliance, Venstre, and Det Konservative Folkeparti.

TABLE 1 | The main themes discussed on Twitter by the danish politicians during the parliamentary elections of 2015.

	Theme	Hashtag
1	Children	#dajegvar12
2	Climate	#dkgreen – #talklima – #verdensvildesteforskel
3	Economy	#talop – #dkain – #socialdumping – #nulv
4	Education	#skolechat – #uddpol
5	Election's Practices	#nypolitiskkultur
6	Europe	#eurdk
7	Government Interference	#frihed
8	Health	#sundpol – #sundhed
9	IT	#itpol – #itvalg
10	Refugees	#nuloverdeigen – #engangvarjegflygtning
11	Woman's Rights	#100aaret
12	Work	#arbejde – #dksocial – #dagpenge

interactions. **Figures 3**, **4** illustrate the communities detected on the multiplex constituted of the following/follower layer, the retweet layer and the reply layer (A), and those detected on the thematic multiplex (B). The two solutions are largely different in terms of the number of detected communities (8 in the first multiplex, and 3 in the second one), and the composition of each community in terms of the political coalition and the political affiliation of the members constituting each community.

As to the selection of the community detection method for our multiplex networks in this paper, we chose a modularity-maximization based community detection method, Generalized Louvain (Jutla et al., 2017) for this task. The reason is that we consider, by assumption, our networks to be undirected networks and our initial focus is on analyzing the communities resulted by the structural features of the network rather than the information flow. For that reason, we chose Generalized Louvain given that it is a well referenced method in the literature to detect this type of communities. The method define communities by optimizing the modularity of the network. In simple graphs, i.e., one layer networks, this translate to finding the best partitioning of nodes into groups, i.e., communities, that maximize the amount of edges within these groups and minimize the number of edges among them. As to the multi-layer extension of this method, it finds the best partitioning that maximize the multi-layer modularity function which is an extension of the simple modularity defined for simple networks. The extended version of modularity introduces a new parameter to the modularity function that is the coupling parameter ω among nodes that belong to the same actor (i.e., the same Twitter user in our case). When $\omega = 1$ (the default case), this means that the coupling among nodes that belong to the same actor is strong. As a result, a partitioning where multiple nodes that belong to the same user (a node represents the existence of a user in a specific layer) lie in the same community contributes intrinsically to the final score of the extended-modularity. In the rest of this paper, we will refer to the output of a community detection method (which is a set of communities) as a clustering.

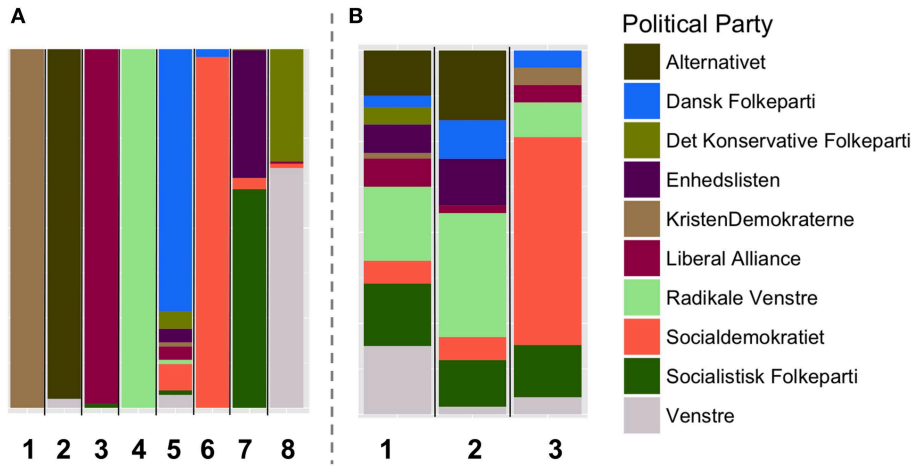

FIGURE 3 | The resulted communities by applying community detection using Generalized Louvain on two different multiplex networks over the DkPol dataset: **(A)** the multiplex constituted of the three layers (following/follower, retweet, and reply) and **(B)** the thematic multiplex. Each bar refers to a different community and the colors in each bar (i.e., community) refer to the composition of each community in terms of the political affiliation of the members constituting it.

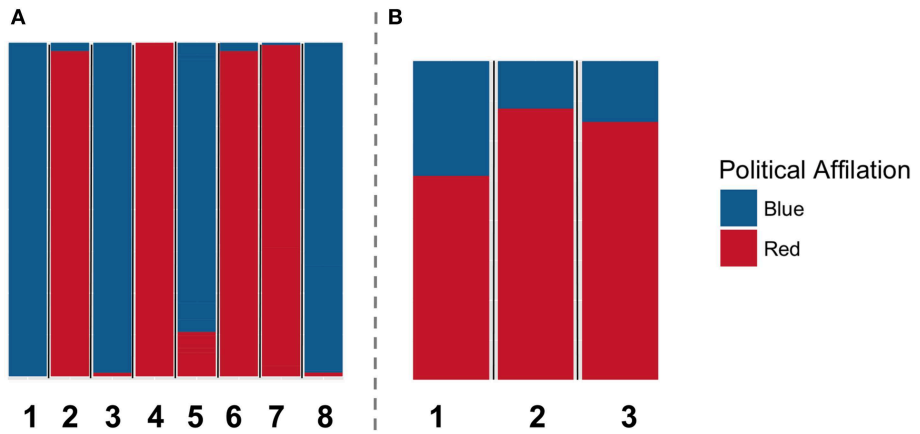

FIGURE 4 | The resulted communities by applying community detection using Generalized Louvain on two different multiplex networks over the DkPol dataset: **(A)** the multiplex constituted of the three layers (following/follower, retweet, and reply) and **(B)** the thematic multiplex. Each bar refers to a different community and the colors in each bar (i.e., community) refer to the composition of each community in terms of the political coalition (red, blue) of the members constituting it.

To better understand the topical dynamics during the month leading to the elections, we chose to create 4 thematic multiplex networks (one for each week content during the month leading the election day). The reason behind choosing "1 week" as a time-window based on which we split the data is that during the month leading to the elections, politicians had to debate on a public TV show once per week.

As illustrated in **Figure 2**, the resulted communities do not necessarily expand over all the layers, meaning that some topics can be absent in some communities. In addition, nodes may not be evenly distributed over layers (for example, community C_1 in **Figure 2** is constituted of 3 nodes in each of the Refugees layer and the Education layer and 4 nodes in the Economy layer). This suggests that topics have different weights, and as a result priorities, in each community which can be interpreted as: some communities, for example, discuss the topic Economy more intensely than they do with the topic Education. To clearly illustrate this, we construct a bipartite network from each clustering. The goal from these bipartite networks is to visualize the relationship between the communities of each clustering and the topics. The width of an edge in the bipartite network between a community and a topic reflects the extent to which that topic is prioritized in that community.

Figure 5 shows the resulted bipartite networks, one per week. We invite our reader to look at this figure together with **Figure 6** which reports , in the form of colored mini-tables , the composition of each community in terms of political coalitions. The existence of a party in a community is represented as a colored cell in the relevant column in that table. The color of that cell can either be red (if the party is from the red Block) or blue (if

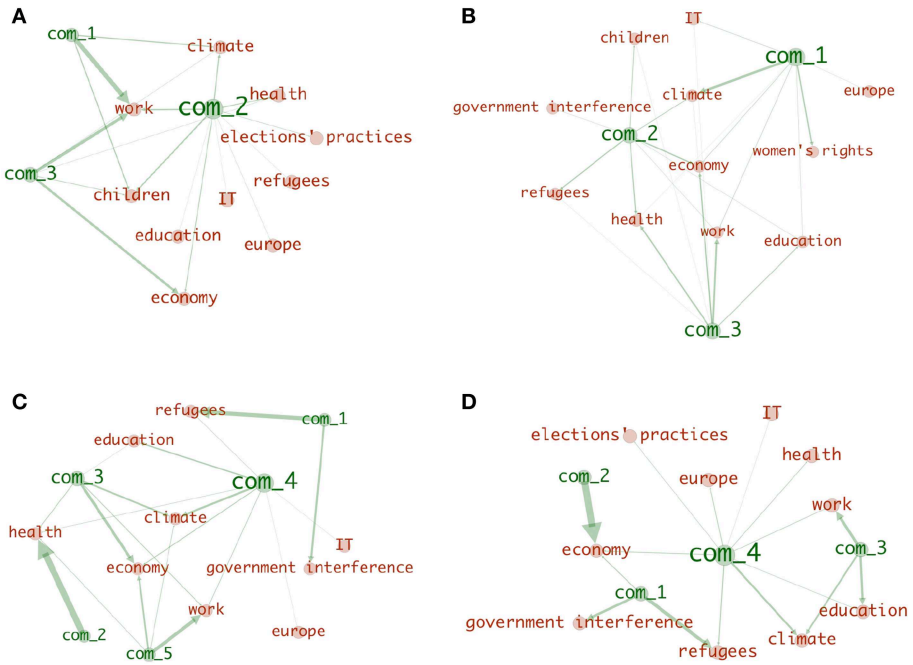

FIGURE 5 | The relationship between the topics and the thematic communities resulted by applying community detection on 4 thematic multiplex newtorks that captured the Twitter interactions of Danish politicians during the month leading the parliamentary lections of 2015 (one per week). **(A)** week 1, **(B)** week 2, **(C)** week 3, **(D)** week 4.

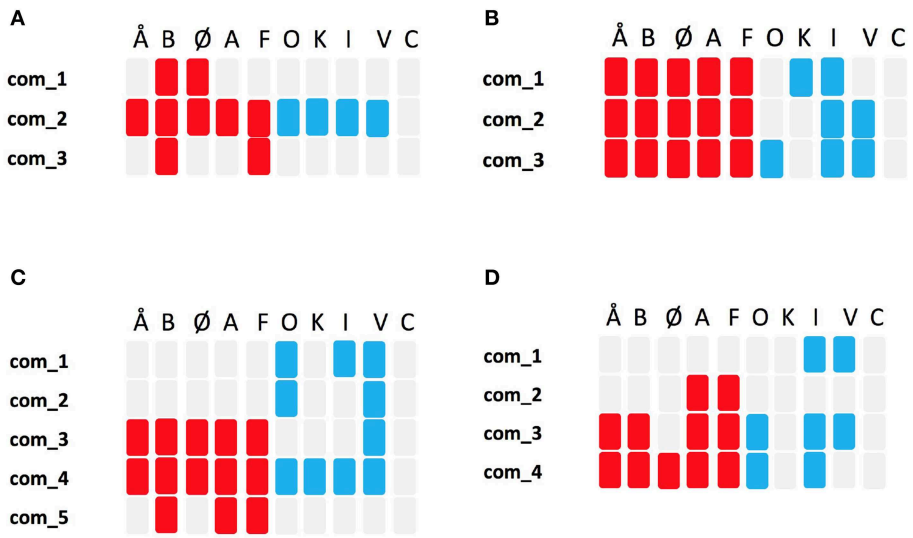

FIGURE 6 | The compositions of the communities reported in **Figure 5** in terms of their political affiliation/party and political coalition/block. **(A)** week 1, **(B)** week 2, **(C)** week 3, **(D)** week 4.

the party is from the blue block). A cell that is neither blue nor red implies the absence of that party (identified by the correspondent column) in the community identified by its row.

3.3. Observations

By looking at **Figure 5A** together with **Figure 6A**, we see that applying community detection on the thematic multiplex of the first week resulted in three communities. Two communities, com_1,com_3, that focus more on economic issues (economy theme and work theme) are composed solely of left-wing (red block) parties. In addition, one community, com_1, constituted of almost all the red block and the blue block parties, tackled all topics with more focus on children, climate, work, and economy themes. Only one of the 12 themes (woman's rights) is absent in all the online debates happened with the first week. The analysis of **Figures 5A**, **6A** shows how during the first week of the election

campaign there was a set of bipartisan topics, that were deemed to be central and worth debating, from both political blocks and other themes that were part of political messages of only one of the two blocks.

This scenario seems to change during the second week as **Figure 5B** together with **Figure 6B** report the absence of single-coalition communities. However, the differences among the communities can be observed on the level of their topical interests. For example, com_1 has more focus on woman's rights and climate issues, com_2 equally prioritized refugees, health and economy issues, while com_3 had focused on work, education, work, and economy. It is also interesting to observe how some of the topics that were, during the previous week, part of a single coalition community (e.g., "economic issues" in com_3 during the first week but part of a bipartisan community - com_2 - in the second week), are now part of the bipartisan conversation. While the detailed study of this dynamic process is outside the goal of this paper, this seems to suggest that opposite coalition might follow each others' themes in order to be present in the topical debate.

During the third week can observe a new polarization of the picture. **Figures 5C**, **6C** show, com_1, com_2, constituted of only blue block parities with interests in refugees, government interference, and health issues. One community, com_5, is constituted of only red block parties with interests on economical issue (work and economy themes). One community, com_3, constituted of almost mostly red block parties (with only one blue block party) with interests in both climate and economy. A debate among almost all parties is still present in the third week represented by com_3 with more focus on climate. These topical division seems very much aligned with the core political values of the two blocks at the time of the election.

This topical difference is largely maintained into the fourth week, the week of election, where we can see—**Figures 5D**, **6D**—four thematic communities. Com_1 which is constituted of only right-wing parties (blue block) with interests in refugees and government interference issues, com_2 which is constituted of only left wing parties (red block) with interests only in economy, and both com_3, com_4 which are mixed in terms of the coalitions, and with main interests in (work/education/climate) and climate, respectively.

4. DISCUSSION

A clear difference has been shown when analyzing the communities on the thematic multiplex versus those detected on a multiplex constituted of the following/follower, retweet, and reply layers. This strongly suggests that community detection on the thematic multiplex reveals different dynamics than those observed using traditional ways of modeling twitter interactions. This is not to say that the thematic multiplex can substitute the traditional ways of modeling Twitter activities, but just to shed a light on different dynamics that can be observed using this way of modeling.

Applying longitudinal community detection on the thematic multiplex network obtained from Twitter data allowed us to observe several interesting dynamics. Given that the dataset captured the interactions among Danish politicians during the month leading the parliamentary elections of 2015, we were able to capture the interest of a political party (or coalition) in specific issues, regardless of the fact that the issue produced an explicit interaction with other users through retweets or replies. During a political campaign, when much of the communication is aimed at promoting the party's agenda to the potential voters, which does not necessarily involve retweeting or replying actions , this type of implicit communication is of key importance. Nevertheless, the thematic multiplex network approach was also able to observe the topics that were more contentious between the parties as well as the topics highly polarized. Moreover, the combination between multiplex thematic analysis and longitudinal data allowed us to show how the political debate, and resulting political communities, are highly dynamic and driven by the ongoing events or campaign themes.

While there might exist other ways to model topic driven implicit interactions on Twitter for clustering tasks, we still think that using multiplex network model offers clear advantages. First, the multiplex network model is a well-developed and widely used model for modeling complex systems (Cardillo et al., 2013; De Domenico et al., 2015) and therefore, provides a powerful, and at the same time flexible, modeling tool that allows for translating properties and variables of complex systems into multi-layer graph proprieties. Second, the plethora of community detection methods developed to detect communities in multiplex networks provides practitioners with more power to choose what works the best for the context of their data.

The idea of moving the qualitative analysis to the modeling phase in the thematic multiplex adds lots of power to the interpretability of the output of a community detection task on this multiplex network. While a fully automated approach to group hashtag into themes/topics could seem a tempting idea, the real complexity behind social media hashtagging is still far from being fully understandable by natural language processing tools and text mining technologies currently at hand. An example are two of the hashtags in our collection: *#engangvarjegflygtning* (translated: *one day I was a migrant*) and *#dajegvar12* (translate: *when I was twelve*). In both cases an *emotional* hashtag is used to discuss specific issues, the refugee crisis with the first and children policies with the latter. The connection between the topic and the hashtag is not explicit, and while both hashtags are clearly topical hashtags (thus referring to a specific topic or event and suggesting the desire of the user to participate to an ongoing larger conversation Bruns and Moe, 2014) they also contain an emotional layer that, as well as the specific topic, is hard to understand if taken out of the specific cultural and societal context.

A future iteration on this work should consider testing the thematic multiplex on other datasets. An important extension should also consider the scalability problem with large scale datasets. The main complexity of this model comes from the greedy approach of connecting the user with his imagined audience via a clique. This means that by using a hashtag for only one time, a user is adding to the model a number of edges equals to the number of all other users who used the same hashtag. While a naive approach to minimize the size of these cliques could be to apply a threshold on the number times a user should use a hashtag before being part of the clique, we still think that

further research should be carried out to find other alternatives for the clique concept in the thematic multiplex without any loose in the information.

Even though the idea of using hashtags to gather communications of users that are not otherwise connected (e.g., not following each other) was originally introduced by Twitter, many other platforms such as Facebook and Instagram have adopted this idea in various ways. Thus, we suggest that this model should not be limited to Twitter data as it could be easily applied to other hashtag-based communicative contexts (e.g., Instagram) as well as to other conceptually similar digital contexts (e.g., participation in Facebook pages).

On a separate note, we would like to mention the fact the resulted communities may largely depend on the chosen community detection method. Indeed, weather or not the thematic communities will be significantly different among different community detection methods can be a research question on its own and we think that answering this question is out of the scope of this paper.

5. CONCLUSION

In this paper we propose an innovative model, the thematic multiplex, to model topic-driven interactions on Twitter. The thematic multiplex is a multi-layer network where each layer corresponds to a different topic, and users (nodes) within a layer will be connected via a clique if and only if they used the same hashtag. We explain the motivation behind the thematic multiplex which is the fact that it considers implicit interactions among users on Twitter that are usually neglected in other models. We construct the thematic multiplex of a real-world Twitter dataset describing the Twitter interactions among the danish politicians during the parliamentary elections of 2015. We show that applying multiplex community detection on the thematic multiplex allows us to observe different dynamics than those we would observe on other models.

DATA AVAILABILITY

The datasets analyzed for this study can be found in the dkpol GitHub repository on the following link [https://github.com/obaidaITU/dkpol].

AUTHOR CONTRIBUTIONS

OH and LR conceived of the presented idea and developed the theory, discussed the results, and contributed to the final manuscript. OH performed the experiments. LR supervised the findings of this work.

FUNDING

This research is supported by the VIRT-EU project funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 727040.

REFERENCES

Bruns, A., and Burgess, J. E. (2011). "The use of twitter hashtags in the formation of *ad hoc* publics," in *Proceedings of 6th European Consortium for Political Research General Conference* (Reykjavik: University of Iceland), 1–9.

Bruns, A., and Moe, H. (2014). "Structural layers of communication on twitter," in *Twitter and Society, Vol. 89*, eds K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann (New York, NY: Peter Lang), 15–28.

Cardillo, A., Gómez-Gardeñes, J., Zanin, M., Romance, M., Papo, D., del Pozo, F., et al. (2013). Emergence of network features from multiplexity. *Sci. Rep*. 3:1344. doi: 10.1038/srep01344

Ceron, A., Curini, L., Iacus, S. M., and Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media Soc.* 16, 340–358. doi: 10.1177/1461444813480466

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., and Flammini, A. (2011). "Political polarization on twitter," in *Proceedings of the Fifth International Conference on Web and Social Media* (AAAI Press), 89–96.

De Domenico, M., Nicosia, V., Arenas, A., and Latora, V. (2015). Structural reducibility of multilayer networks. *Nat. Commun.* 6:6864. doi: 10.1038/ncomms7864

Dickison, M. E., Magnani, M., and Rossi, L. (2016). *Multilayer Social Networks*. Cambridge: Cambridge University Press.

Ibrahim, R., Elbagoury, A., Kamel, M. S., and Karray, F. (2018). Tools and approaches for topic detection from twitter streams: survey. *Knowl. Inf. Syst.* 54, 511–539. doi: 10.1007/s10115-017-1081-x

Jutla, I. S., Jeub, L. G. S., and Mucha, P. J. (2017). *A Generalized Louvain Method for Community Detection Implemented in Matlab*. Technical report.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). "What is twitter, a social network or a news media?," in *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, NC: ACM), 591–600.

Litt, E. (2012). Knock, knock. Who's there? The imagined audience. *J. Broadcast. Electron. Media* 56, 330–345. doi: 10.1080/08838151.2012.705195

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. (2012). Community detection in social media. *Data Min. Knowl. Discov.* 24, 515–554. doi: 10.1007/s10618-011-0224-z

Silva, W., Santana, A., Lobato, F., and Pinheiro, M. (2017). "A methodology for community detection in twitter," in *Proceedings of the International Conference on Web Intelligence* (New York, NY: ACM), 1006–1009.

Yang, J., and Counts, S. (2010). "Predicting the speed, scale, and range of information diffusion in twitter," in *Proceedings of the Fourth International Conference on Weblogs and Social Media* (Washington, DC), 355–358.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor and reviewer (SG) declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration

frontiers in Big Data

PERSPECTIVE
published: 12 June 2019
doi: 10.3389/fdata.2019.00012

Identifying Travel Regions Using Location-Based Social Network Check-in Data

*Avradip Sen and Linus W. Dietz**

Department of Informatics, Technical University of Munich, Munich, Germany

Travel regions are not necessarily defined by political or administrative boundaries. For example, in the Schengen region of Europe, tourists can travel freely across borders irrespective of national borders. Identifying transboundary travel regions is an interesting problem which we aim to solve using mobility analysis of Twitter users. Our proposed solution comprises collecting geotagged tweets, combining them into trajectories and, thus, mining thousands of trips undertaken by twitter users. After aggregating these trips into a mobility graph, we apply a community detection algorithm to find coherent regions throughout the world. The discovered regions provide insights into international travel and can reveal both domestic and transnational travel regions.

Keywords: data-mining, human mobility modeling, spatial clustering, region detection, visualization

OPEN ACCESS

Edited by:
Roberto Interdonato, Télédétection et Information Spatiale (TETIS), France

Reviewed by:
Cristian Molinaro, University of Calabria, Italy
Sabrina Gaito, University of Milan, Italy

***Correspondence:**
Linus W. Dietz
linus.dietz@tum.de

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *24 March 2019*
Accepted: *27 May 2019*
Published: *12 June 2019*

Citation:
Sen A and Dietz LW (2019) Identifying Travel Regions Using Location-Based Social Network Check-in Data. Front. Big Data 2:12. doi: 10.3389/fdata.2019.00012

1. INTRODUCTION

The destinations visited within a trip may overarch existing administrative divisions of provinces, federal states, and countries. For example, visiting the Alps of Europe, one is not restricted in travel by country borders as all adjacent countries are members of the Schengen Area. When developing a travel region recommender system for composite trips this is a challenge, because one needs a region model to choose the recommendations from Dietz (2018). To come up with such a model, we propose to observe traveler mobility behavior, aggregate it using spatial clustering methods, thereby re-drawing the boundaries of the world's travel regions using a data-driven approach.

Data collected from location-based social networks has previously been used as a proxy for human mobility, however, such data sets are either not readily available, are focused on small areas, such as cities, or have too sparse check-ins of the users. Hence, we use public Twitter APIs to collect traveler data in the form of geotagged tweets. From the series of tweets, we determine the home location of the user and then extract the trips (Dietz et al., 2018). These trips are then aggregated into a weighted graph of tourist flows with nodes being cities and edges being the number of trips from one city to another. This graph is then fed into a community detection algorithm (Bohlin et al., 2014), whose results constitute the world's travel regions irrespective of established political boundaries.

In this position paper, we want to motivate this approach, describe our ideas to implement and evaluate such a method. Furthermore, we outline the implications and benefits of a data-driven region model in other domains, such as recommender systems.

2. METHOD

Twitter allows algorithmic access to a stream of public tweets through their APIs, which can be queried to build a data set of geotagged tweets. By querying timelines of users who have enabled sharing the geolocation of their tweets, we can follow their movement patterns. To reduce noise, the individual geolocations are matched to the nearest city. Thus, each tweet in the timeline constitutes a check-in to a city. After the home city of the user has been determined by the highest number of check-ins, consecutive check-ins outside of the home city can then be combined to a trip. To focus on travelers, we exclude all trips shorter than 7 days. Furthermore, we require at least one check-in within 5 days, to ensure sufficient data quality. For more details on the trip mining, we refer to our previous paper (Dietz et al., 2018).

The trips are then transformed into an undirected graph, where each city is a node, and the edges represent the flows divided by the distance between the two cities. The flows are computed by summing up the co-occurrences of the two nodes in a clique formed by all cities in a trip. For example, if somebody traveled from Munich to Berlin via Nuremberg in one trip, we would also count the flow from Munich to Berlin as one. Including the distance into the edge weight was useful to reduce noise in the flow graph introduced by distant traffic hubs, such as airports. With this graph-based representation, we can run the Infomap multi-level community detection algorithm to see which cities form coherent clusters (Rosvall et al., 2009).

3. PRELIMINARY RESULTS

Running this approach with trips from Twitter reveals four major clusters on the highest hierarchy:

1. North and Central America,
2. South America,
3. Europe, Russia, Arabia, Western and South Africa, and
4. Eastern Africa, Asia, and Oceania.

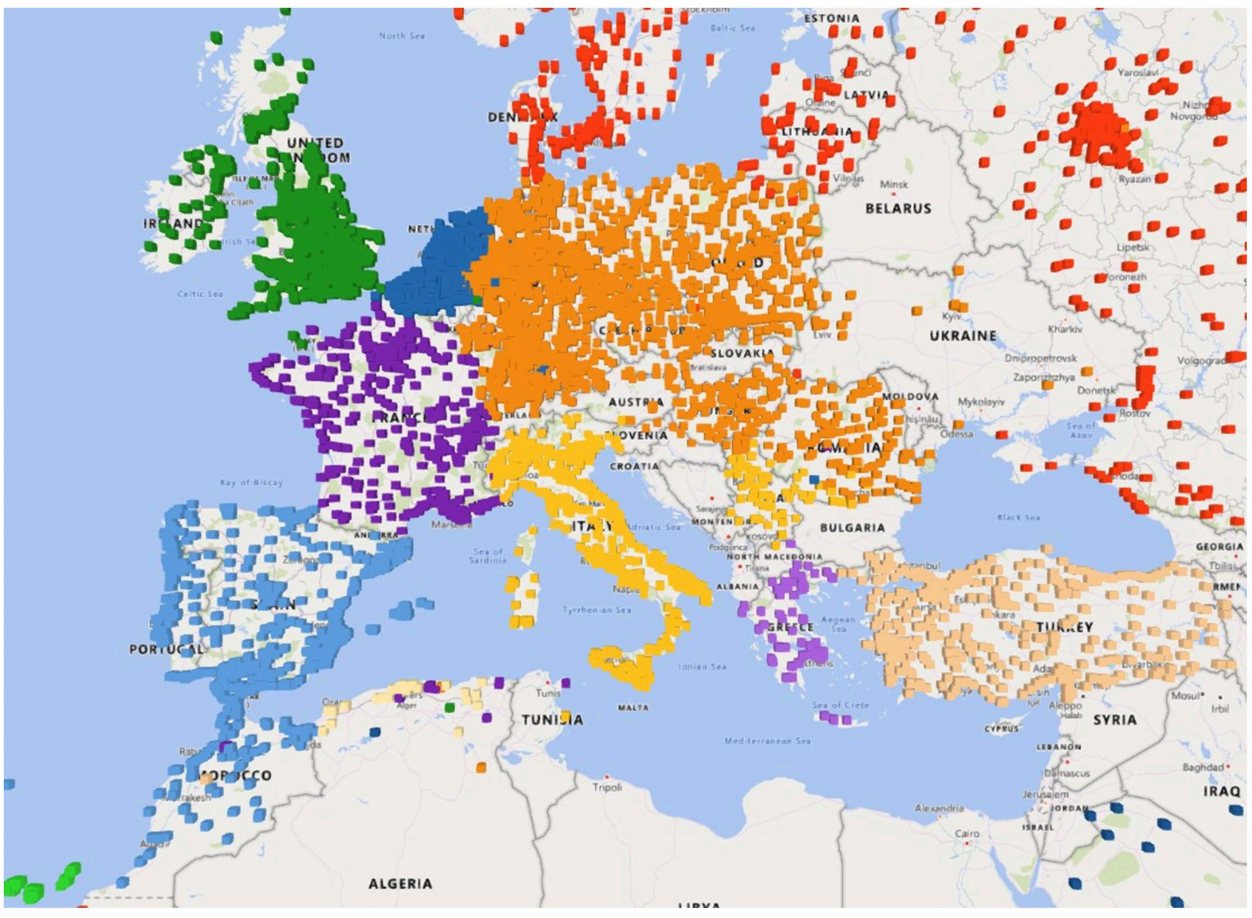

FIGURE 1 | The second-level community structure of Europe.

The level two clusters of Europe, depicted in **Figure 1**, correspond to groups of similar countries. The British Isles, the Iberian Peninsula, and much of Central and Eastern Europe are merged into respective clusters, while countries like France, Italy, and Turkey roughly retain their own clusters. This is already an interesting result, as it shows that political boundaries have a strong influence on the travel behavior. Subdividing these clusters reveals further regions, however the results become more fuzzy and subject to thorough evaluation. One major challenge is to find a termination criterion to decide whether to continue splitting these clusters. In our opinion, this cannot be decided with the current data, but requires further analysis of the regions, such as the number of cities and the area covered. An evaluation of the quality of the discovered region will also prove to be challenging. However, comparing our third-level clusters of the United Kingdom with those of Ratti et al. (2010) revealed high similarities.

4. RELATED WORK

Human mobility analysis has helped us to improve our understanding of traffic forecasting (Kitamura et al., 2000), the spread of diseases (Eubank et al., 2004), and also computer viruses (Kleinberg, 2007). Researchers have already attempted to define regions based on human mobility data for various purposes such as administrative region discovery (del Prado and Alatrista-Salas, 2016), topical region discovery (Taniguchi et al., 2015), and political redistricting (Joshi et al., 2009). Closest to our approach is the work of Hawelka et al. (2014), who aim to find larger regions of mobility, by combining several countries. We aim to find touristic regions that are smaller and potentially independent of countries.

There are various algorithms to perform spatial clustering and community detection, such as the Louvain method (Blondel et al., 2008), GDBSCAN (Ester et al., 1996), and Infomap (Rosvall et al., 2009). They are comparable in runtime complexity, however (Fortunato and Hric, 2016) finds that the Infomap algorithm outperforms the Louvain method in the quality of the communities. GDBSCAN uses the distance between points explicitly to form clusters that are geographically contiguous. Thus, we use Infomap, as it allows to use self-computed weights for the graph and can detect hierarchies. This resolves the resolution limit problem, where the size of communities depend on the size of the graph, which can result in recognized communities being merged together in large networks.

5. CONCLUSIONS

This position paper introduces an approach for spatial clustering of touristic regions from trips mined from Twitter. To the best of our knowledge, this is the first application of geo-located tweets to find travel regions, with data spanning the whole world. The analysis of results finds a coherent hierarchy of clusters. This confirms that the use of tweets to find traveler mobility patterns and define regions based on the patterns is a feasible approach.

In future, we plan to make a thorough evaluation of the resulting regions using numeric method, but also to visually compare them to findings of other region discovery approaches.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

AS: Prototype implementation, experimentation, literature analysis. LD: Main author of manuscript, developed the trip mining library.

FUNDING

This work was supported by the German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program.

REFERENCES

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.* 2008, 1–12. doi: 10.1088/1742-5468/2008/10/P10008

Bohlin, L., Edler, D., Lancichinetti, A., and Rosvall, M. (2014). "Community detection and visualization of networks with the map equation framework," in *Measuring Scholarly Impact: Methods and Practice*, eds Y. Ding, R. Rousseau, and D. Wolfram (Cham: Springer), 3–34.

del Prado, M. N., and Alatrista-Salas, H. (2016). "Administrative regions discovery based on human mobility patterns and spatio-temporal clustering," in *Proceedings of 13th International Conferences on Mobile Ad Hoc and Sensor Systems, MASS'16* (Brasilia: IEEE), 65–74.

Dietz, L. W. (2018). "Data-driven destination recommender systems," in *Proceedings of 26th Conferences User Modeling, Adaptation and Personalization, UMAP '18* (New York, NY: ACM).

Dietz, L. W., Herzog, D., and Wörndl, W. (2018). "Deriving tourist mobility patterns from check-in data," in *Proceedings of the WSDM WS on Learning from User Interactions* (Los Angeles, CA).

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). "Density-Based Clustering Methods," in *KDD-96 Proceedings* (Portland, OR).

Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., et al. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature* 429, 180–184. doi: 10.1038/nature02541

Fortunato, S., and Hric, D. (2016). Community detection in networks. *Phys. Rep.* 659, 1–44. doi: 10.1016/j.physrep.2016.09.002

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inform. Sci.* 41, 260–271. doi: 10.1080/15230406.2014.890072

Joshi, D., Soh, L.-K., and Samal, A. (2009). "Redistricting using heuristic-based polygonal clustering," in *Proceedings of 9th International Conference on Data Mining* (Miami, FL: IEEE), 830–835.

Kitamura, R., Chen, C., Pendyala, R. M., and Narayanan, R. (2000). Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation* 27, 25–51. doi: 10.1023/A:1005259 324588

Kleinberg, J. (2007). Computing: the wireless epidemic. *Nature* 449, 287–288. doi: 10.1038/449287a

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., et al. (2010). Redrawing the map of Great Britain from a network of human interactions. *PLoS ONE* 5:e14248. doi: 10.1371/journal.pone.0014248

Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. *Eur. Phys. J. Spec. Top.* 178, 13–23. doi: 10.1140/epjst/e2010-01179-1

Taniguchi, Y., Monzen, D., Ariestien, L. S., and Ikeda, D. (2015). "Discover overlapping topical regions by geo-semantic clustering of tweets," in *29th International Conference Advanced Information Networking and Applications Workshops* (Gwangiu: IEEE), 552–557.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor and reviewer SG declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

frontiers in Big Data

ORIGINAL RESEARCH
published: 20 June 2019
doi: 10.3389/fdata.2019.00016

Choosing Optimal Seed Nodes in Competitive Contagion

Prem Kumar[1]*, *Puneet Verma*[1]*, *Anurag Singh*[1] and *Hocine Cherifi*[2]

[1] Department of Computer Science and Engineering, National Institute of Technology Delhi, New Delhi, India, [2] LIB EA 7534, University of Burgundy, Dijon, France

In recent years there has been a growing interest in simulating competitive markets to find out the efficient ways to advertise a product or spread an ideology. Along this line, we consider a binary competitive contagion process where two infections, A and B, interact with each other and diffuse simultaneously in a network. We investigate which is the best centrality measure to find out the seed nodes a company should adopt in the presence of rivals so that it can maximize its influence. These nodes can be used as the initial spreaders or advertisers by firms when two firms compete with each other. Each node is assigned a price tag to become an initial advertiser which varies according to their importance in the network. Considering their fixed budgets, they initially determine the payoff of their products and the number of their initial seeds in the network. Under this setting, we study the question of whether to choose a small number of influential nodes or a larger number of less influential nodes.

Keywords: competitive contagion, complex networks, game theory, seed nodes, competitive marketing, centrality measures

OPEN ACCESS

Edited by:
Andrea Tagarelli,
University of Calabria, Italy

Reviewed by:
Francesco Gullo,
UniCredit, Italy
Martin Atzmueller,
Tilburg University, Netherlands

***Correspondence:**
Prem Kumar
kprem193@gmail.com
Puneet Verma
puneetverma866@gmail.com

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *25 March 2019*
Accepted: *28 May 2019*
Published: *20 June 2019*

Citation:
Kumar P, Verma P, Singh A and Cherifi H (2019) Choosing Optimal Seed Nodes in Competitive Contagion. Front. Big Data 2:16. doi: 10.3389/fdata.2019.00016

1. INTRODUCTION

Contagion in general life means the communication of disease from one person or organism to another by close contact. This definition can be extended by replacing the disease with a product or an ideology. Competitive Contagion is a type of contagion which deals with conflict and race of multiple firms who want to influence or infect more people than others. There are a lot of situations that can be described in such a way, for example: Two political parties trying to influence the citizens by giving incentives to some influential people in the country and directing them to advertise their ideology, Two mobile phone manufactures competing to advertise their mobile phones of same segment by hiring celebrities or tech reviewers and giving them incentives. So, it is important to simulate such an environment and provide algorithms and properties for optimal seed selection for the competitive contagion process. While doing such competition this work can be used by firms to select the initial spreaders or advertisers by analyzing their network topological properties.

Diffusion on networks is a fundamental process which involves spreading of an ideology (or infection) in a population, e.g., epidemic disease contagion, spread of innovation by word-of-mouth. Considering a network diffusion model, the influence maximization problem consists of finding a set of initial seed nodes so that the expected size of the resulting cascade is maximized. Supposing that there is a limit k on the number of nodes to target (e.g., due to advertising budgets), the goal is to efficiently find an appropriate set of k nodes with which to "seed" a diffusion process. Classical works by Kempe et al. (2003, 2005), on this subject are competitive unaware. They focused on designing models of spreading of a single influence (or idea) and algorithms to find out the optimal seed nodes for maximal adoption of a product of a single firm only. However in

real life scenarios, several firms compete in the same market and multiple infections can occur simultaneously in the same network. This has led to the increase in effort toward finding seeds for more realistic settings. Recent works by Bharathi et al. (2007) and Goyal et al. (2014) focused on the modeling the competitive contagion of multiple firms using game theory. They proposed an algorithm to select the seed nodes in a network and discussed the Nash equilibrium when multiple firms compete against each other. Despite the considerable progress made toward finding the seeds in the social network in competitive settings, some very basic questions remain unanswered. Indeed, one approach used to make the influence maximization is to reduce the problem into the ranking of the nodes according to the centrality measures. In other words, the seed nodes are selected by mentioning its ranks using various centrality measure. This raises an important question as to which of the centrality measures should firms use to rank the nodes while selecting them as seed nodes. To answer this question, we compare various centrality methods for finding the rank of nodes. In order to do so, for all the centralities under investigation, we assign a seed strategy based on a centrality to one firm and another to the other one and we compare their spreading efficiency. We also address the question whether a firm should select a small number of highly influential nodes or a larger number of less influential nodes. To answer this question we assign each node a price tag using the best centrality measure found during our analysis and give a fixed and same amount of budget to both firms. Then one firm stakes his funds in buying large number of cheap seed nodes and other in buying small number of expensive seed nodes and then we compare their influence or number of nodes infected when stability is achieved.

The rest of paper is organized as follows: In section 2 basic terminologies and definitions mentioned in the paper are recalled. In section 3, the diffusion model is presented. Section 4 deals with comparing the efficiency of classical centrality measures in the competitive contagion in order to choose the seed nodes. Section 5 compares the strategies of using few highly influential seeds rather than a higher number of less influential seeds with the same budget. Section 6 concludes the paper.

2. BACKGROUND

We can classify the contagion processes (Dassios and Zhao, 2011) into three categories based on the dependency of one disease A to another disease B:

1. Competitive: This type of process occurs only if there are multiple diseases (or information about products or ideologies) to propagate. Here, if a node is already infected by a disease A it resists the infection by another disease B, e.g., diffusion of ideology of two political parties.
2. Cooperative: It is just the opposite of competitive contagion. Here, if a node is already infected by a disease A, and another infection B is trying to infect it, disease A helps disease B to infect the node. e.g., : Diffusion of two diseases (Tuberculosis and common flu).
3. Independent: As the name suggests in this type of contagion no infection (or information about product) interacts with each other and are independent.

We need to calculate the importance of a node in the network to assign its price. Higher rank nodes will be considered costlier in comparison with lower rank nodes. Centrality is a measure for calculating the importance of a node based on its topological properties in the network. There are numerous centrality measures based on various topological properties of nodes which are used in order to assign a score of importance to every node (Gupta et al., 2015, 2016). In this work we restrain our attention to the most influential measures. Their definitions are given below:

1. Degree Centrality: It considers that the node centrality is linked to the size of its neighborhood. It is simply the number of nodes at a distance of one edge.
2. Closeness Centrality: It considers nodes having smaller distance with all other nodes to be more central.

$$Closeness(v) = \frac{1}{\sum_{i \neq v} d_{vi}}$$

   - where d_{vi} is distance between node v to i.

3. Betweenness Centrality: It works on the concept that the more often a node acts as a bridge along the shortest path between any two nodes, the more central it is.

$$Betweeness(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

   - where σ_{st} is total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

4. EigenVector Centrality: It works on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. For a given graph $G := (V, E)$, Let A be the adjacency matrix.

$$Ax = \lambda x$$

   - where λ is the eigenvalue and x is the resulting eigenvector which contains the centrality measure of *ith* node at *ith* row.

   There will be multiple eigenvalues λ for which non-zero solution exists. However, (by the Perron, 1907; Frobenius, 1912 theorem) only the greatest eigenvalue results in desired centrality measure.
5. Page Rank Centrality (Page et al., 1999): It is a variant of the EigenVector Centrality. It works on the assumption that more important nodes are likely to receive more links from other nodes.

$$PR(u) \propto \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

i.e., the PageRank value for a page u is dependent on the PageRank values for each page v contained in the set B_u (the set containing all pages linking to page u), divided by the number $L(v)$ of links from page v. The algorithm involves a damping factor for the calculation of the pagerank.

3. DIFFUSION PROCESS

We study a competitive process of adoption of multiple products made by multiple firms who use their respective monetary resources for advertisement of their product to the consumers located in a network. Each firm has a fixed budget to advertise their products to the users in a social network. Therefore, each firm needs to optimally choose a set of seed nodes using the assigned budget for maximum adoption of their product. We use the generic game theoretic model (Osborne and Rubinstein, 1994) for the study of competition between firms. In view of game theoretic scenario in competitive market, we propose a diffusion algorithm for the spreading of any information about a product.

The proposed game theoretic model may be represented as:

Players: The firms (A and B).
Actions: Each firm's set of actions is to choose the initial seed nodes or their advertisers.
Preferences: Each firm's preference is to maximize the adoption of their product in the network or to infect the maximum number of nodes possible.

Multiple firms may try to spread information about their products in the underlying social network. Here, in this work two firms are considered for spreading information about their two products, respectively. As the only action provided to firms is to choose the seed nodes at the beginning, so the result of the entire game depends on the strategy to choose the set of starting spreader (or seed) nodes. Each of the firms (A and B) comes into the open market to advertise their product with limited budget C_A and C_B. They have their node ranking algorithm using which they rank the nodes present in the network and then select the nodes whose price is less than their remaining budget, starting from highest rank (numerically lowest: Rank 1) till the funds remained are not enough to hire any node. The proposed diffusion Algorithm 1 used to simulate the dynamics is an extension of a previous work on simulating epidemic and rumor spreading (Kumar et al., 2018) in which we proposed a simple cascade algorithm for diffusion, discussed various characteristics of epidemic and rumor spreading and the relation among various attributes of epidemic and rumor spreading. This algorithm is a cascade based algorithm in which at each timestamp all the infected nodes try to transmit their disease (or ideology) to their direct connections and the probability that infection will transmit depends upon λ_A, λ_B, c_{iA}, and c_{iB}. **Figure 1** shows the conversion model of nodes based on Algorithm 1.

λ_A and λ_B are the infection rates of infection A and infection B which are constant. c_{iA} and c_{iB} are the competitive measures of node i. It is 1 at time $t = 0$ for every node and changes when the node is infected by any infection. When ith node is infected by A, c_{iA} gets multiplied by α_A and if jth node is infected by B, c_{jB} gets multiplied by α_B. α_A and α_B are the competitive constants, larger the α_X more is the resistance of a node infected by X for another infection.

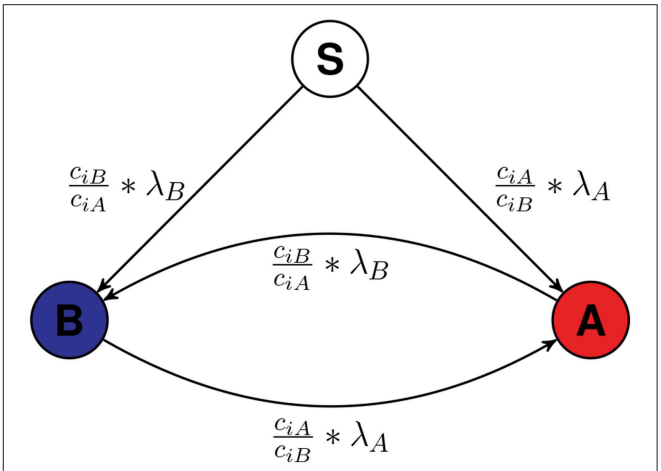

FIGURE 1 | The model of spreading. The transparent node labeled as S is a Susceptible node which is not yet influenced by either firms. Initially all the nodes except seed nodes are Susceptible and they can be influenced by A or B. Once infected by either A or B, they will stay infected. As the scenario is competitive, a node infected by A can be converted to a node infected by B and vice versa, if the required condition is satisfied. The value along the arrow is the probability of infection. The node is assumed as the ith labeled node.

The population is divided into two compartments: Susceptible and Infected and infected is further divided into two compartments: Infected by A and Infected by B. Let N be the total population, S be the number of Susceptible nodes, X_A and X_B be the number of seed nodes of A and B, I_A and I_B be the number of nodes infected by A and B respectively. At the start of simulation $I_A = X_A$ and $I_B = X_B$.

$$\begin{aligned} &\textit{The law of conservation will be}: N = S + I_A + I_B \\ &\textit{where } X_A \subset I_A \ \& \ X_B \subset I_B \end{aligned} \tag{1}$$

4. FINDING THE OPTIMAL CENTRALITY MEASURE

The agents present in the network take a fixed amount to advertise or spread the product of the firms and that value is decided in accordance to centrality value of the agents. To find out which centrality measure is more effective for finding the most influential nodes in a competitive contagion scenario, we compare the following five centrality measures: Page Rank, Degree, Betweenness, Closeness, EigenVector. To do so, we consider each method as a node ranking algorithm of a firm trying to advertise its product. Therefore, there are total $\binom{5}{2}$ matches (*Match of every centrality against every other centrality*. Each firm ranks the top 10 nodes according to their node ranking algorithm and put them in their seed nodes set. As there are cases in which both competing centralities have common nodes in their

Algorithm 1 Diffusion algorithm.

G: Population (Graph).
S_A, S_B: The set of seed nodes of firms A and B respectively.
λ_A, λ_B: Probability of spreading information about the product of A and B firms respectively.
c_{iA}, c_{iB}: competitive measures for Firms A and B respectively.
α_A, α_B: competitive constant for respectively of A and B.
For each node, i other than the initial seeds $S_A \cup S_B$:
procedure ($G, S_A, S_B, \lambda_A, \lambda_B, c_{iA}, c_{iB}, \alpha_A, \alpha_B$)
    Count the number of neighbors infected by A (n_A) and B (n_B) and respectively
    $x \leftarrow n_A * \lambda_A * c_{iA}/c_{iB}$
    $y \leftarrow n_B * \lambda_B * c_{iB}/c_{iA}$
    if $x > y$ **then**
        Generate a random number (r_0) between 1 and 100.
        if $r_0 < \lambda_A * c_{iA}/c_{iB}$ **then**
            $node_i$ gets infected by A
            $c_{iA} \leftarrow c_{iA} * \alpha_A$
        end if
    else
        generate another random number (r_1) between 1 and 100.
        if $r_1 < \lambda_B * c_{iB}/c_{iA}$ **then**
            $node_i$ gets infected by B
            $c_{iB} \leftarrow c_{iB} * \alpha_B$
        end if
    end if
    if $x < y$ **then**
        Generate a random number (r_2) between 1 and 100.
        if $r_2 < \lambda_B * c_{iB}/c_{iA}$ **then**
            $node_i$ gets infected by B
            $c_{iB} \leftarrow c_{iB} * \alpha_B$
        end if
    else
        generate another random number (r_3) between 1 and 100.
        if $r_3 < \lambda_A * c_{iA}/c_{iB}$ **then**
            $node_i$ gets infected by A
            $c_{iA} \leftarrow c_{iA} * \alpha_A$
        end if
    end if
end procedure

top 10 list, we assign only unique nodes to each. An example of this distribution is given in **Table S2**. **Table S2** contains the list of seed nodes for competitions of various centrality measures. After making the set of seeds for each match, we run the simulation for dynamics of infection using the Algorithm 1.

5. CHOOSING THE TYPE AND NUMBER OF SEEDS

A general confusion among the firms is whether to choose small number of highly influential advertisers or large number of less or average influential advertisers. To solve this problem, we simulate a competition between a large group of less (or average) influential nodes and a small group of highly influential nodes both needing nearly same amount of budget.

For ranking the nodes while simulating the competition between group of small number of highly influential nodes and group of large number of less influential nodes we will use the most optimal centrality method found during the simulation discussed in section 4. We select the two sets such that both of them cost nearly same.

$$Cost \propto Centrality\ score \tag{2}$$

To investigate the *high-less* (highly influential nodes in small numbers) vs. *low-more* (low influential nodes in large number) competition. We took a set of less influential nodes mostly from different clusters in the *low-more* set and most influential nodes in the *high-less* set such that the cost of both is same.

TABLE 1 | Properties of data-sets used.

Network	Nodes	Edges	Av. Clustering Coe.	Diameter
Wikipedia vote	7,115	103,689	0.1409	7
Chess interaction	7,301	65,053	0.126	13
Human interaction	410	2,765	0.436	9

TABLE 2 | Results of various competitions on various network datasets.

Player 1	Player 2	Winner-Wiki	Winner-Chess	Winner-Human
Pagerank	EigenVector	**EigenVector**	**Eigenvector**	**Pagerank**
Closeness	EigenVector	**Closeness**	**Closeness**	**Closeness**
Betweenness	EigenVector	**EigenVector**	**Eigenvector**	**Betweenness**
Degree	EigenVector	**EigenVector**	**Degree**	**Degree**
Pagerank	Betweenness	**Betweenness**	**Pagerank**	**Pagerank**
Betweenness	Closeness	**Closeness**	**Closesness**	**Tie**
Degree	Betweenness	**Degree**	**Degree**	**Degree**
Degree	Closeness	**Closeness**	**Degree**	**Degree**
Pagerank	Degree	**Degree**	**Degree**	**Pagerank**
Pagerank	Closeness	**Closeness**	**Closeness**	**Pagerank**

Bold column values show the winners of respective matches.

6. RESULTS AND ANALYSIS

We use three empirical network data-sets to perform the experiments [the wikipedia vote network which is available on: SNAP Stanford[1], the chess interaction network which is available on KONECT[2]], and the human contact network available on KONECT[3]. We have used the Wikipedia Vote Network as our primary data-set and others for verification.

[1](accessed February 3, 2019). *SNAP: Network datasets: Social circles.*
[2](accessed February 25, 2019). *KONECT Networks.*
[3](accessed May 13, 2019). *KONECT Networks.*

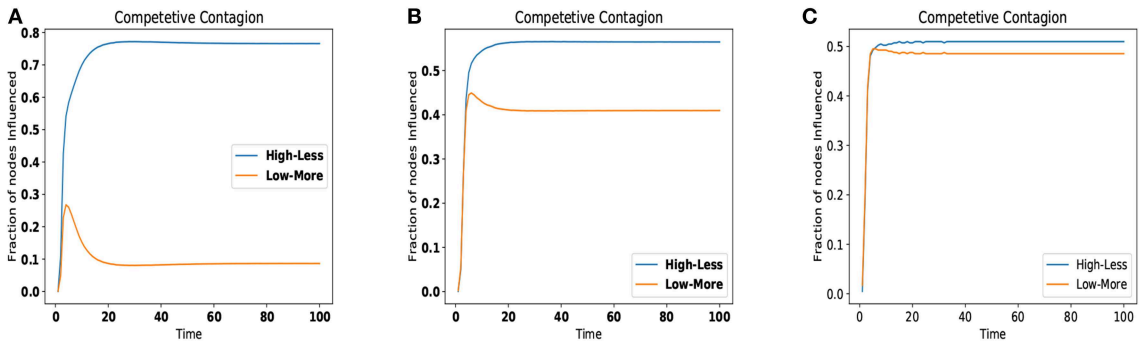

FIGURE 2 | Competition of High-Less vs. Low-More for various datasets. **(A)** Wikipedia Vote Network. **(B)** Chess interaction network. **(C)** Human interaction network.

The results for networks other than chess interaction is added in the **Supplementary Material**. Details of the basic topological properties of all networks is given below in **Table 1**.

6.1. Finding the Optimal Centrality Measure

The simulation proposed in section 4 is run on a fixed rate of spreading ($\lambda_A = \lambda_B = 0.6$) and fixed competitive constant ($\alpha_A = \alpha_B = 1.1$) for both firms for 100 timestamps and for each timestamp, average of 50 iterations are considered. Output of the simulation is the ratio of nodes infected by each firm after each timestamp. **Table 2** shows the results of matches among various centrality methods when simulated with various data sets. The individual curves for the matches (Fraction of nodes infected by each firm vs. time) is provided in **Supplementary Material**.

The simulation results (**Table 2**) shows that no centrality performs best in competitive setting of contagion and it is data dependent.

6.2. Choosing the Type and Number of Seeds

As proposed in section 5, we simulated the competition between two firms, one having higher number of less influential node and one having small number of highly influential node using all the three datasets. As we have seen that none of centrality is best for all datasets but it is data dependent so, we will use the centrality method which performed best for that particular dataset. So for Wikipedia Vote Network it will be Closeness, for Chess Interaction it will be Degree, and for Human Interaction network it will be Pagerank.

As the **Figure 2** depicts for all three datasets, the number of infected (or influenced) nodes remains same for both the sets up to few timestamps, but after that *High-less* set takes over then, stabilization is achieved. Overall winner is *High-less* (less number of highly influential nodes) if we use the better performing centrality measure as per datasets to assign the costs of nodes. For further verification we used a synthetic dataset, but in the case of synthetic dataset all the centralities demand the nearly the same nodes at each rank to it is not possible to allocate the nodes to any centrality and simulate the competition of centralities.

7. CONCLUSION

In this paper, we investigate empirically two linked issue in a competitive contagion setting. First, we simulate a set of competitions between strategies based on ranking according to various centrality measures. The goal is to choose the best centrality method to rank the nodes for initial adoption with a motive to maximize the adoption of the product or ideology. Results show that no centrality is universally best but it depends on the network properties of the network dataset used. The second part deals with solving the general dilemma of whether to choose group of small number of highly influential nodes or a group of large number of less influential nodes. We conclude that it is better to select a small number of highly influential nodes than a higher number of less influential nodes. We can extend this work by taking variable rate of spreading, cooperativity and competitive constant of the diffusion model. Future works could also be done by considering more sophisticated alternative network properties for selecting the seed nodes.

DATA AVAILABILITY

The datasets analyzed for this study can be found http://snap.stanford.edu/data/ego-Facebook.html.

AUTHOR CONTRIBUTIONS

PK and PV discussed the idea with AS. PK, PV, AS, and HC designed the methodology for the experiments. PK did the simulation and data pre-processing. PV did the data representation and image editing. AS and HC guided and helped with bugs and theoretical understanding. All authors wrote the article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2019.00016/full#supplementary-material

REFERENCES

Bharathi, S., Kempe, D., and Salek, M. (2007). "Competitive influence maximization in social networks," in *International Workshop on Web and Internet Economics* (Berlin; Heidelberg: Springer), 306–311.

Dassios, A., and Zhao, H. (2011). A dynamic contagion process. *Adv. Appl. Probabil.* 43, 814–846. doi: 10.1239/aap/1316792671

Frobenius, G. (1912). "Über Matrizen aus nicht negativen Elementen," in *Preussische Akademie der Wissenschaften Berlin: Sitzungsberichte der Preußischen Akademie der Wissenschaften zu Berlin* (Reichsdr).

Goyal, S., Heidari, H., and Kearns, M. (2014). Competitive contagion in networks. *Games Econ. Behav.* 113, 58–79. doi: 10.1016/j.geb.2014.09.002

Gupta, N., Singh, A., and Cherifi, H. (2015). "Community-based immunization strategies for epidemic control," in *7th International Conference on Communication Systems and Networks*. Available online at: https://ieeexplore.ieee.org/abstract/document/7098709

Gupta, N., Singh, A., and Cherifi, H. (2016). Centrality measures for networks with community structure. *Phys. A Statist. Mech. Appl.* 452, 46–59. doi: 10.1016/j.physa.2016.01.066

Kempe, D., Kleinberg, J., and Tardos, É. (2003). "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC: ACM), 137–146.

Kempe, D., Kleinberg, J., and Tardos, É. (2005). "Influential nodes in a diffusion model for social networks," in *International Colloquium on Automata, Languages, and Programming* (Berlin; Heidelberg: Springer), 1127–1138.

Kumar, P., Verma, P., and Singh, A. (2018). "A study of epidemic spreading and rumor spreading over complex networks," in *Towards Extensible and Adaptable Methods in Computing* (Singapore: Springer), 131–143.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. Cambridge, MA: MIT Press.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The Pagerank Citation Ranking: Bringing Order to the Web.* Technical report, Stanford InfoLab.

Perron, O. (1907). Zur theorie der matrices. *Math. Annalen* 64, 248–263.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 25 June 2019
doi: 10.3389/fdata.2019.00017

Twitter Response to Munich July 2016 Attack: Network Analysis of Influence

Ivan Bermudez[1], *Daniel Cleven*[1], *Ralucca Gera*[1]*, *Erik T. Kiser*[1], *Timothy Newlin*[1] and *Akrati Saxena*[2]

[1] Naval Postgraduate School, Monterey, CA, United States, [2] Department of Computer Science, National University of Singapore, Singapore, Singapore

Social Media platforms in Cyberspace provide communication channels for individuals, businesses, as well as state and non-state actors (i.e., individuals and groups) to conduct messaging campaigns. What are the spheres of influence that arose around the keyword #*Munich* on Twitter following an active shooter event at a Munich shopping mall in July 2016? To answer that question in this work, we capture tweets utilizing #*Munich* beginning 1 h after the shooting was reported, and the data collection ends approximately 1 month later[1]. We construct both daily networks and a cumulative network from this data. We analyze community evolution using the standard Louvain algorithm, and how the communities change over time to study how they both encourage and discourage the effectiveness of an information messaging campaign. We conclude that the large communities observed in the early stage of the data disappear from the #*Munich* conversation within 7 days. The politically charged nature of many of these communities suggests their activity is migrated to other Twitter hashtags (i.e., conversation topics). Future analysis of Twitter activity might focus on tracking communities across topics and time.

Keywords: Twitter data analysis, Munich July 2016 attack, social network analysis, meme propagation, influence spread

OPEN ACCESS

Edited by:
Sabrina Gaito,
University of Milan, Italy

Reviewed by:
Vincent Labatut,
Laboratoire Informatique d'Avignon, France
Marinette Savonnet,
Université de Bourgogne, France

***Correspondence:**
Ralucca Gera
rgera@nps.edu

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *24 March 2019*
Accepted: *03 June 2019*
Published: *25 June 2019*

Citation:
Bermudez I, Cleven D, Gera R, Kiser ET, Newlin T and Saxena A (2019) Twitter Response to Munich July 2016 Attack: Network Analysis of Influence. Front. Big Data 2:17. doi: 10.3389/fdata.2019.00017

1. INTRODUCTION

1.1. Event Background

On July 22, 2016, a mass shooting occurred in a shopping mall in Munich, Germany. The attacker was quickly identified by local police as an 18 year old German-Iranian dual national resident of Munich (Harrison, 2016). As is often the case after high impact incidents like mass shootings, there was a high volume of conversation in social media associated with this shooting. Conversations range from official government accounts providing instructions to affected people, speculation regarding the identity and motivation of the attacker(s), and individuals or news organizations providing reports (accurate or otherwise) of the event. As micro-blogging services like Twitter become more popular, it becomes interesting to analyze the data generated by the service in an attempt to extract topologies or trends that may provide insight into the event in question. A timeline of this event is displayed in **Figure 1**.

[1]The collected dataset will be posted online for public use once the research work is published.

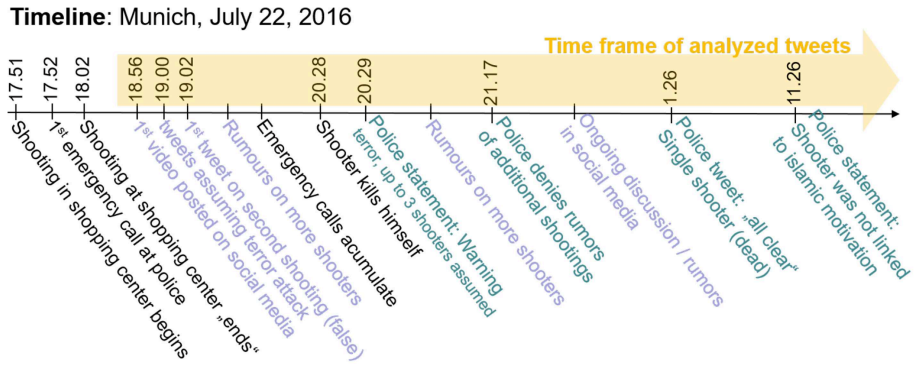

FIGURE 1 | Timeline of attack (Zeitung, 2016).

1.2. Motivation: Twitter Connection to Information Warfare

The Internet and Cyberspace foster many types of activities that involve different aspects of human social interaction. We can visualize and analyze the relationships that convey these interactions using network science techniques. Twitter is undoubtedly a common channel by which significant online social interaction occurs. Individuals, organizations, and nation states all use this tool as a medium of communication and many interested listeners, then retweet statements they believe deserve the attention of others. Both the real world and contrived activity generate conversations around a particular hashtag. Regardless of the authenticity of an event, the social interactions that occur during and after its occurrence have a real effect on the way humans perceive the world and can influence their future actions both in the world and in Cyberspace. To improve our appreciation for how message information both spreads and decays, we increasingly study and understand how information campaigns develop and change.

Information Warfare has been occurring for as long as parties have been trying to deceive their opponents. While the information itself may not be physical, it is considered by social scientists and the Department of Defense in Joint Publication 1 (Department of Defense, 2013) as one of the instruments of national power that the nation states utilize in order to pursue their ends. The other instruments include Diplomatic, Military, and Economic power (DIME) (Department of Defense, 2013). At first glance, Twitter seems to offer the empowerment of free speech to any user, and yet our analysis of the retweeting that occurs helps demonstrate how little many users are interested in genuine original thought. Rather, the majority of traffic tends to gravitate toward sharing the thoughts of a few accounts. We believe that such influence, while not forced by any entity, still offers tremendous power for parties engaged in Information Warfare to increase their power within the domain of Cyberspace. This power is not limited to national security and a nation's foreign policy but extends into the realms of domestic politics, sports, business, and many other areas.

Does Twitter offer the empowerment of free speech to any user? And does that make a difference? To understand this, we analyze the Twitter data we collected on the Munich attack using Netlytic (Gruzd, 2016), a software that captures data and can perform social network analysis as well. We collected dataset focused on the surge in Twitter activity using *#Munich* linked to the July 22 shootings which garnered international attention across social media and traditional reporting channels. We analyze both temporal slices of the data and the cumulative dataset to better understand how information and messages propagate across Twitter. In particular, we are interested in the community structure, its evolution, and the role of top influential leaders within these communities.

The main contributions of the paper are: (1) The collection of the hashtaged *#Munich* dataset from Twitter for an active shooter event at a Munich shopping mall in July 2016. (2) The general analysis of the cumulative network of retweets for this incident. (3) The evolution of the influence flow-based communities in temporal network of timeslices by day. In section 2 we discuss related work. Section 3 covers the problem definition and the details of the collected dataset. In sections 4 and 5, we discuss the methodology and results, respectively. The paper concludes with several future directions.

2. RELATED WORK

In the current era of social networking, information sharing has been easier by posting microblogs (Kempe et al., 2003; Leskovec et al., 2007). The influence spreads very fast over the network and impacts the opinion of the users or maybe groups of users, i.e., communities (Lin et al., 2008). Researchers have studied the influence propagation on Social networking platform and their impact on network structure (Sadikov and Martinez, 2009; Chen et al., 2013; Saxena et al., 2015). Hong et al. (2011) proposed a method that successfully predicts popular tweets using the content of the message, temporal information, metadata of messages and users, structural properties of the users' social network.

Of more specific interest to us is the study of spreading behavior of tweets in case of attacks, hazards, natural calamities, etc, and how it affects the opinion of the users. Nadamoto et al. (2013) observed that the spreading of rumor during the disaster

situation is different from the normal situation. In a disaster situation, the rumor goes through two or three hierarchy, but in the normal situation, it passes through many hierarchies. In the case of news spreading during disasters, Jin et al. (2014) showed that lies, half-truths, and rumors spread in the same way as true news using tweets during the Ebola crisis. On the other hand, Mendoza et al. (2010) showed that the propagation of rumor differs from the true news, and this information can be used to detect rumors using aggregated analysis on tweet dataset collected on the 2010 earthquake in Chille. Spiro et al. (2012) proposed a model for the waiting time of retweets and showed that the hazard related tweets have a shorter waiting time. For a non-disaster situations, Vosoughi et al. (2018) observed that the false news spread faster, farther and deeper, and are more prominent in the case of political news than financial, disaster, terrorism or science-related news. This research is based on Twitter data spanning 11 years comprising around 126,000 stories tweeted by around 3 million people. A brief survey on influence propagation on online social networks can be seen at Bonchi (2011).

How do communities emerge while influence spread? Gupta et al. (2016) studied the role of core-periphery structure in the information propagation to multiple communities. Complementing the spreading behavior, we are also interested in identifying influential user or users on Twitter, the emergence of influential leaders in different communities, how they shift from one community to another and how they die out (Tsur and Rappoport, 2012; Riquelme and González-Cantergiani, 2016). More specific, understanding this phenomenon based on dominant language per hashtag to trace which users overlap between the thematic and linguistic communities delineated by different information streams (Bastos et al., 2013). Our research examines several language communities that intermix with political leanings of conversations, Spanish, French, and English all use *#Munich* although it is important to remember the German discussion mostly emerged under #München. By studying the dependencies between global features such as graph topology and content features emergence helps in explaining how long members might remain in the community and the importance of repeated messaging to maintain the community of influence over time. Successful prediction of the spread of memes can improve marketing efforts whether the target is a commercial product or an idea being promoted.

Influence propagation has also been studied using the multilayered structure of online social networks. The layers depict either different type of relationship, allowing the researchers to perform studies at different granularity (Li et al., 2012; Zhuang and Yağan, 2016) or the layers representing followers, mentions or retweeting (Borondo et al., 2015). We also include the multilayer aspect in our research in a different way, namely temporally. Wang et al. (2008) stated that most nodes lack stability in the evolution of the network between time steps, and the manner in which time is partitioned will determine how communities are detected. This inspires our analysis to examine if and how accounts migrate between communities over time. Yet, in terms of stability, Romero et al. (2011) highlighted that hashtags on politically controversial topics are particularly persistent, with repeated exposures continuing to have unusually large marginal effects on adoption. In this research, we do not specifically examine how long certain messages persist, but the observation about political messages lasting longer is related to how long individuals choose to continue retweeting the same leader accounts over multiple days. That information is captured in the multipartite temporal network, and it is shown in section 5.

Smith et al. (2014) from the Pew Research Center found six different network structures (Polarized Crowds, Tight Crowd, Brand Clusters, Community Clusters, Broadcast Network, Support Network) that emerge in social media networks. They study how the structures differ based on the content of the issues driving the discussion, highlighting the importance that most real social networks are usually a hybrid of multiple structures. The research shows that Broadcast Network, and Support Network have large size groups, Tight Crowd has medium size groups, and Brand clusters and Community Clusters have many small sized groups. The structures of interest to the *#Munich* dataset were the Community Clusters, and Broadcast Network. Both of these structures appeared within the context of the retweets in the month following the July 2016 attack in Munich. The Pew researchers give voice to the idea that mapping the social landscape using networks helps interpret trends, topics, and implications of the technologies being used. Our analysis of the *#Munich* data regarding the polarized crowd supports the Pew team's statement that if a topic is political, it is common to see two separate groups take shape and they form two distinct discussion groups that mostly do not interact with each other. The groups are recognizably liberal or conservative (Smith et al., 2014). Each group links to a different set of influential people or organizations that can be found at the center of each conversation cluster.

3. PROBLEM DEFINITION AND DATASET

The broader problem examined in this research is how social media spheres of influence in Cyberspace can be employed to conduct information operations campaigns. We analyze the communities of influence in the considered dataset and their evolution over time.

The approach to solving this problem uses efforts similar to Pew Research (Smith et al., 2014) work on Social Media, personalized for the *#Munich* Dataset. Our background research leads to the understanding that the structure of the network we create affects how communities emerge. In this work, we focus on retweets only, because they convey the aspect of influence, as individuals choose to associate with particular leader's thoughts. The network's nodes are thus the Twitter accounts that have retweets at least once, and directed edges connect retweeting accounts to the account of origin for that message.

The original data captured in Netlytics consists of 13 files of total 655 MB (Gruzd, 2016). It conveys all Tweets captured from July 22, 2016 to August 22, 2016 labeled with *#Munich*. This discussion topic involved 147,116 Twitter accounts that either tweeted or re-tweeted *#Munich* messages during those 32 days. Each row of the dataset containes several categories including the text of the Tweet, date, time, author, type of device it was

posted from when user/account was created, and the Twitter profile location.

Our research focuses on the Tweets that contain the retweet indicator, "RT@", in the text or body of the message. Of the total 925,019 Tweets, 79.8% were retweets, and 72% of all retweets occurred between July 22 and July 25 which corresponds to the first 3 days after the shooting. The Tweets cover several languages including English, French, and Spanish, all of which use the spelling Munich for the city. However, very few German language Tweets are captured because German Twitter users use the German spelling of "München" instead of "Munich."

4. METHODOLOGY

The raw data was used to build a directed graph G_0 of the *#Munich* data using the following methodology.

(1) Every unique Twitter account that occurred in our data is represented by a node,
(2) A directed edge is placed from node u to node v if user u retweets the tweet that was initially posted by user v,
(3) Edge weights represent the number of times user u retweeted user v's tweets, and
($*$) Edges in G_0 did not contain any temporal information.

The resulted graph G_0 has 147, 116 nodes, 191, 002 edges. To this graph we apply a standard community detection algorithm called Louvain (Blondel et al., 2008). The algorithm assigns nodes randomly to communities, measures the strength of the community partition using modularity (Newman, 2006), and shuffles neighbors from one community to another while maximizing modularity. The result of the Louvain algorithm is 5, 807 communities, which will become part of our cumulative analysis of this network.

Our temporal analysis of the raw data reveals that over 72% of all retweets occurred between July 22, 2016 and July 25, 2016 as shown in **Figure 2**. For temporal analysis, we thus focus the analysis on these 4 days, for which we build sub-graphs $G_{22}, G_{23}, G_{24}, G_{25}$ using the same methodology described above, but only capturing retweets of the top twenty leaders for each day.

Building upon Smith's observations (Smith et al., 2014), we propose and compute the leader score for every node. This score shows which user accounts are influential within a community, and provide a relative scale of their influence. The leader score for every node u in G is computed as follows.

$$leaderscore(u) = \frac{deg_{in}(u)}{deg_{out}(u) + 1}. \quad (1)$$

We further propose a metric to compare leader-centric communities across time, computed in two steps: (1) run Louvain community detection on sub-graphs $G_{22}, G_{23}, G_{24}, G_{25}$, and then (2) add an edge between communities if they have a shared user. Let U_t be the set of users comprising community U on day (t), and V_{t+1} be the set of users comprising community V on day $(t + 1)$. We then compute the similarity between two communities as,

$$similarity(U_t, V_{t+1}) = average\left(\frac{U_t \cap V_{t+1}}{U_t}, \frac{U_t \cap V_{t+1}}{V_{t+1}}\right). \quad (2)$$

Similarity metric assigns the value while considering the sizes of the communities, as community sizes may vary a lot due to their sphere of influence. Next, we present the analysis results using the discussed metrics.

5. RESULTS AND ANALYSIS

In exploring the data set of all the Tweets with *#Munich*, we notice that about 80% of the Tweets were individuals retweeting other users. This dynamically captures the influence of a very small portion of the overall accounts, because these tweets include content that a large number of other users identify with as they get retweeted.

The distribution of retweets vs. day is shown in **Figure 2**. Observe that the distribution of all retweets for 32 days has a strong positive skew with the majority of retweets occurring the day after the attack. Notice that within a week, the activity returns to a level similar to before the attack.

We begin our study with the community structure of the cumulative dataset using Louvain algorithm, identifying 5, 807 communities. For better visualization, we create a graph G^* from G_0 by selecting the 20 largest communities in G_0. G^* contains less than 1% of the communities, but it still accounts for over 70% of the nodes and 75% of the edges in G_0. **Figure 3** shows a plot of the G^* using the ForceAtlas visualization from Gephi (Bastian, 2009). A large number of edges or high edge weights between two communities corresponds to greater proximity on the visualization; whereas communities which share few or no edges will be spaced further apart on the visualization.

The information box for each community in **Figure 3** conveys the following information:

1. The percent of nodes in G_0 that the community comprises.
2. The predominant language of the community, as,
   - EN: English
   - ES: Spanish
   - FR: French
   - HI: Hindi
3. A brief characterization of the community based on the profiles of its leaders using commonly accepted definitions of conservative and liberal social views. The term social commentary is used to emphasize the proffering of opinions rather than the objective conveyance of information.

Figure 3 reveals a partitioning of the communities along linguistic and political lines. We observe that a community built around a common language and/or shared political views is more likely to have a higher edges density. One can visually interpret the data in **Figure 3** as follows:

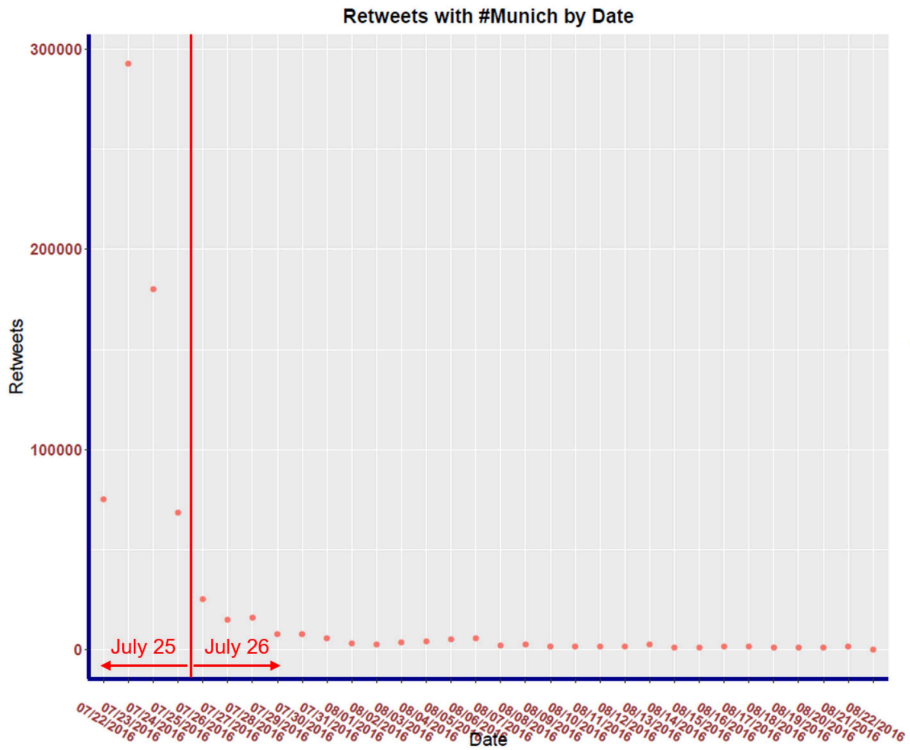

FIGURE 2 | *#Munich* retweets by date.

- Horizontally (left to right): socially conservative communities, politically neutral communities/news sources, socially liberal communities
- Vertically (top to bottom): English language communities, French language communities, Spanish language communities.

Note that the FC Bayern Munich community might be outside the scope of the study of the July 22nd attack, rather tweets on football using the same hashtag(#Munich). Since the data captured it anyway, we have shown it in the analysis.

The partitioning of communities along language and political views reinforces the findings of the Pew study (Smith et al., 2014). The relatively small size of the communities represents news sources given a large number of Twitter followers many of these news outlets have. This is likely a result of how the network is built since it only captures the accounts who actively retweet others yet fail to capture passive users who consume Tweets but do not actively retweet.

Terrorist events such as the Munich attack create a unique circumstance where we assume that leaders within preexisting communities (fundamental communities) attach themselves to a particular hashtag (e.g., topic) and form topic communities. This creates the following cases for followers and leaders.

1. A user exists in the fundamental community but not in the topic community
2. A user exists in both the fundamental community and the topic community
3. A user exists in a topic-specific community but not in the underlying fundamental community.

For example, user u agrees with the sentiment of leader v's Tweet on the topic of the *#Munich* and retweets v's message. Users u and v are in the same topic community, but not necessarily in the same fundamental communities if in general their views do not coincide.

To understand the influential hierarchy of the network, we first apply the core-periphery analysis of the network using K-shell decomposition method (Seidman, 1983). The k-shell decomposition method assigns a k-shell value to each node, and it works in the following way. The k-shell method first removes all nodes of degree one until there is no node of degree one or less, and assigns them k-shell value 1. Iteratively, it will remove nodes of degree 2, 3, 4... and will assign them k-shell value 2, 3, 4... respectively. While removing the nodes of degree k, if any node is ended up having degree k or less, will also be removed in the same iteration. The method is stopped once each node has been assigned a k-shell value. k-core of a network contains all the nodes having k-shell value equal to or higher than k. **Figure 4** presents the split of the 1-core, 3-core and the 5-core between the communities for a better understanding of the core-periphery structure. We observe that the cores of different communities are connected with each other, thus leaders communicate or influence each others. We also observe that the smaller communities do not have higher influential nodes having a higher k-shell value, which could be the reason why they did not become larger communities

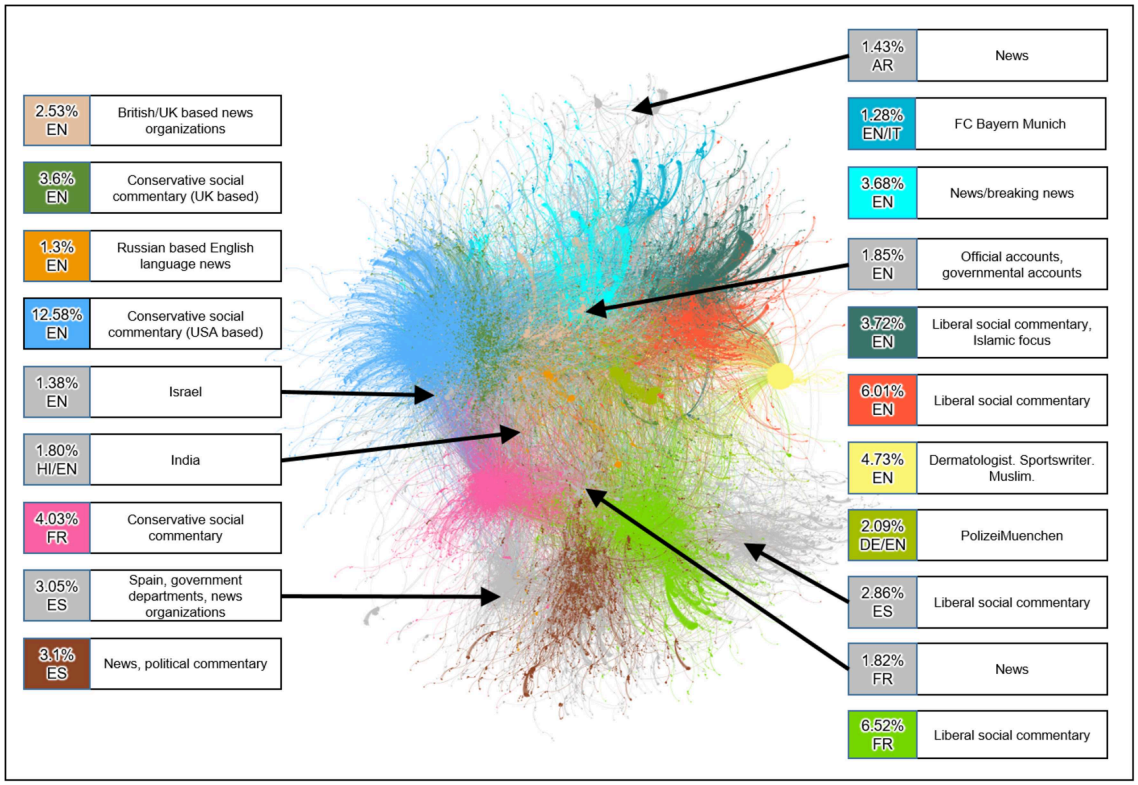

FIGURE 3 | Visualization of G^*.

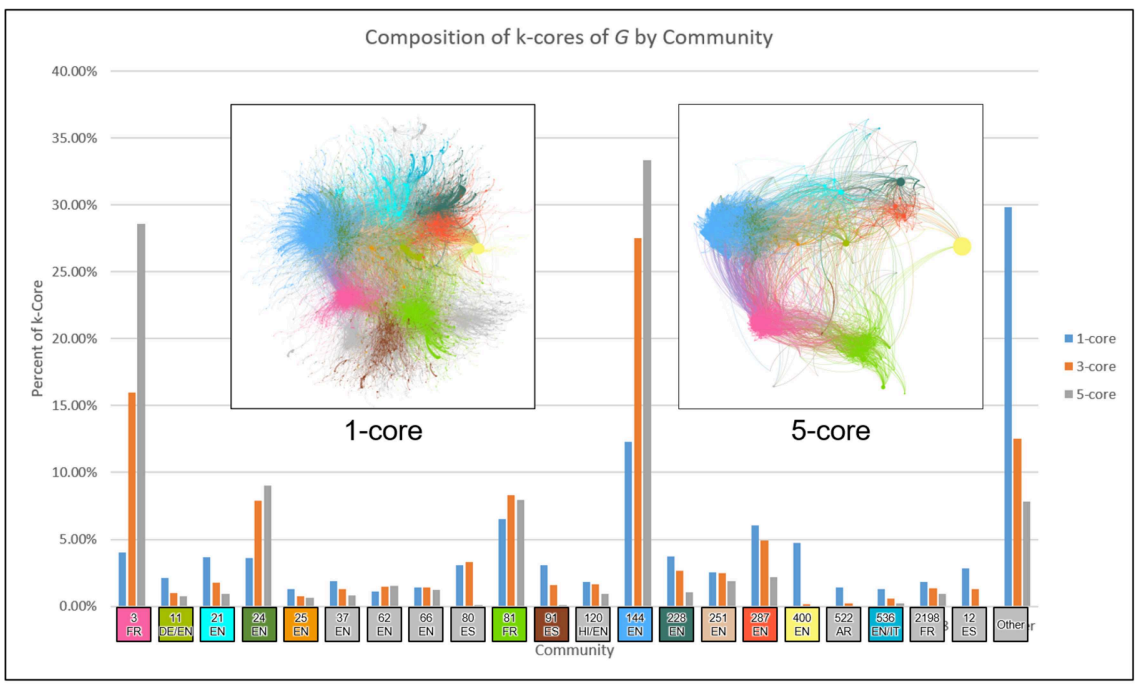

FIGURE 4 | Community partition within different k-cores.

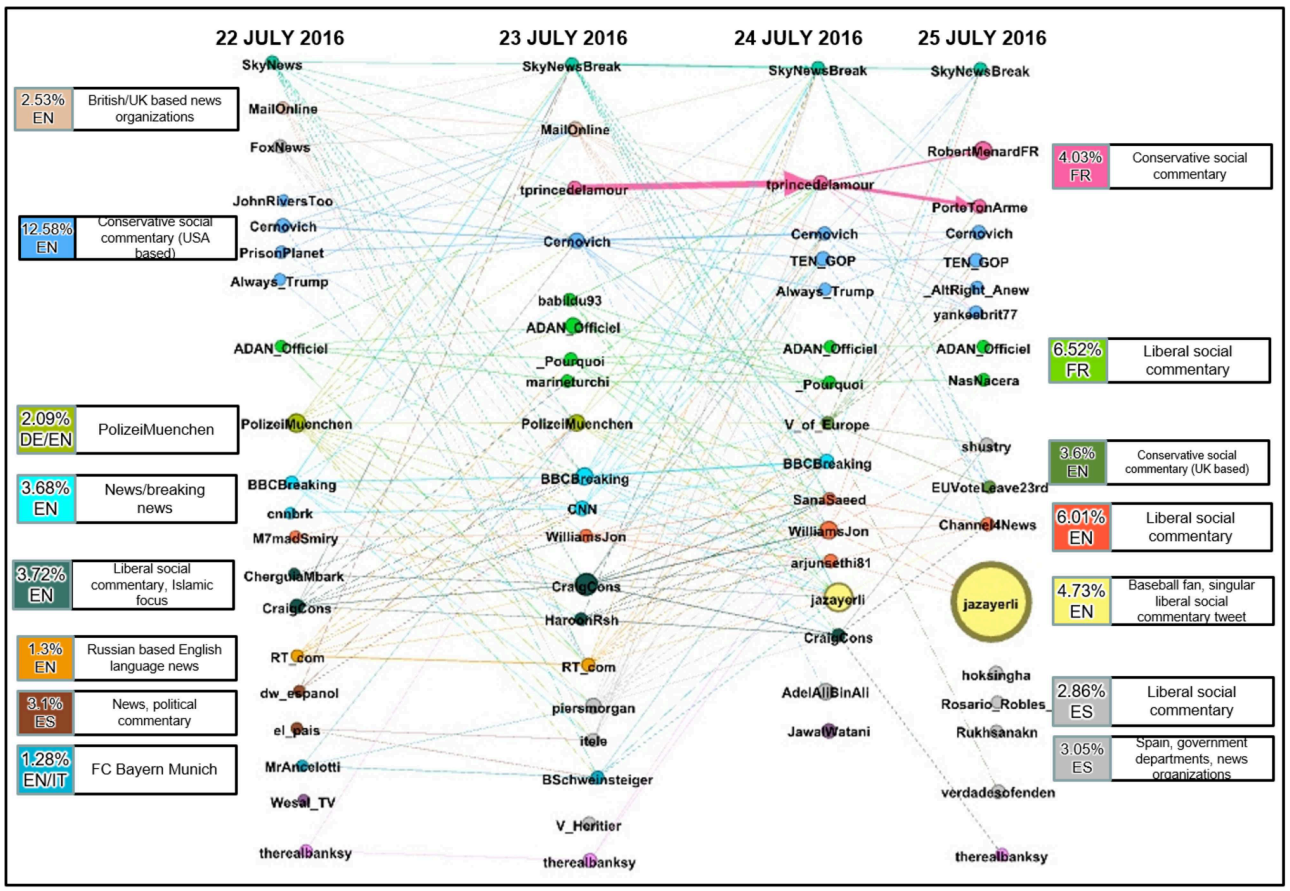

FIGURE 5 | Multipartite temporal community daily evolution.

overtime. Next, we do the temporal analysis of retweet networks for better understanding the role and evolution of the communities.

5.1. Temporal Leader Networks

In this section, we study the evolution of the communities found in the cumulative graph for the first 4 days succeeding the attack. This subset is deemed adequate by examining the frequency of retweets in each day for the whole period. As seen in **Figure 2**, the majority of traffic occurs from the 22 to 25 of July 2016. This subset of data is used to create the multilayer network seen in **Figure 5**.

In this multilayer network, the nodes represent each of the top 20 communities, and a single layer is created for each day. The nodes are sized by the number of users in that community and colored by the communities they belong to in the cumulative graph. The label of each node corresponds to the leader in that community as defined by Equation (1). We then add an edge from a node in one layer to a node in another layer if there are any shared users between the two communities. This captures the continuity of community membership. The weight of the edge is then computed using Equation (2) as described in the methodology.

The resulting network provides significant insight into the evolution of communities over time. **Figure 5** provides visualization for the migration of users and leaders between different communities. We also observe that community leaders appear or disappear each day depending on whether they generate a tweet message and the volume of retweets. For example, the Russian based English language news (RTcom) community dies out after July 23.

We further observe that a significant amount of users retweet from the same community. Although this observation is prevalent in the data, it is most evident in the community labeled as French conservative social commentary. We observe that a high amount of users that retweeted from @tprincedelamour on July 23, 2016, did so again on the next day.

Following communities from left to right we see how they can merge from several nodes to one or split from one over each day, as is the case with the English conservative social commentary (USA based) community. Lastly, the @jazayerli community is seen to grow from the 24 to 25 of July with no connecting edge. The lack of an edge between these two nodes is because the community consists of just one Twitter message generated by @jazayerli that is retweeted several times over the 2 days.

The cumulative and temporal graphs of the Twitter data complement each other by providing overlapping insight into the communities described. **Figure 5** provides insight into the nature of each community; how did leaders and followers' activities for a given community change across time. **Figure 3** provides an overview of each community and its relative importance across time and the degree to which communities and leaders are connected. Examining the yellow colored community lead by @jazayerli, it becomes clear from **Figure 3** that this community is a peninsula (a very small community) and **Figure 5** illustrates that this particular community dominated the *#Munich* retweets on July 25.

6. CONCLUSIONS AND FURTHER DIRECTIONS

In this work, we collect and analyze the Twitter data of *#Munich* July 2016 attack corresponding to a month-long period after the July 22 shootings in Munich. We study the community structure in the cumulative dataset as well as daily partitions and classify each community based on the nature of its leaders and their tweets. This study provides insight on how information spreads on Twitter in case of an event, and we observe how the important leaders disappear from the network of *#Munich* retweets after a week of the attack. The leaders in the first week tended to be news organizations or social leaders with strong or extreme views. Communities expressing strong opinions were the most active; however, as mentioned, the collected data is unable to account for passive users (e.g., users who may read a Tweet and internalize the information or message but do not retweet the message). One can further study the impact of the event on other users who are not directly involved in tweeting and retweeting, however, have been affected by the event. The analysis can also be extended to different social media platforms for better understanding.

This research opens up several questions to be studied for a better understanding of the evolution of the network in case of terrorist attacks. One can identify the leaders and follow them across multiple hashtags to determine topic communities of leaders for each hashtag. By comparing a leader's topic networks and identifying users that retweet the leader across multiple different topics, we can understand the development of the fundamental and topic-based communities represented by that leader. These communities can be further classified based on different parameters, such as is the community passive where the leader has many followers, but few retweets; or is it active where the majority of users following a leader actively retweet the leader across many different topics.

All these approaches will be fruitful in a deeper understanding of how communities generate influence in social media networks. Given the increasing use of online social media, the implications for how corporations, organizations, and nation states conduct influence campaigns will continue to grow as part of future information operations.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

IB, DC, EK, and TN collected the dataset, discussed the research methodology, implemented the project, and contributed to the writing of the manuscript. RG and AS contributed to the research methodology and its implementation, analysis of the results, and writing of the manuscript.

ACKNOWLEDGMENTS

We would like to thank the DoD for partially sponsoring the current research.

REFERENCES

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM* 8, 361–362.

Bastos, M. T., Puschmann, C., and Travitzki, R. (2013). "Tweeting across hashtags: overlapping users and the importance of language, topics, and politics," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT'13* (New York, NY: ACM), 164–168.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

Bonchi, F. (2011). Influence propagation in social networks: a data mining perspective. *IEEE Intell. Inform. Bull.* 12, 8–16. doi: 10.1109/WI-IAT.2011.292

Borondo, J., Morales, A., Benito, R., and Losada, J. (2015). Multiple leaders on a multilayer social media. *Chaos Solit. Fract.* 72, 90–98. doi: 10.1016/j.chaos.2014.12.023

Chen, W., Lakshmanan, L. V., and Castillo, C. (2013). Information and influence propagation in social networks. *Synt. Lect. Data Manag.* 5, 1–177. doi: 10.2200/S00527ED1V01Y201308DTM037

Department of Defense (2013). *JP 1, Doctrine for the Armed Forces of the United States*. Available online at: https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp1_ch1.pdf:174

Gruzd, A. (2016). *Netlytic: Software for Automated Text and Social Network Analysis*. Computer Software. Available online at: http://Netlytic.org (accessed May 25, 2017).

Gupta, Y., Saxena, A., Das, D., and Iyengar, S. (2016). "Modeling memetics using edge diversity," in *Complex Networks VII* eds H. Cherifi, B. Gonçalves, R. Menezes, and R. Sinatra (Springer), 187–198.

Harrison, J. (2016). *German-Iranian teen behind Munich shootings*. BBC. Available online at: http://www.bbc.com/news/live/world-europe-36870986

Hong, L., Dan, O., and Davison, B. D. (2011). "Predicting popular messages in twitter," in *Proceedings of the 20th International Conference Companion on World Wide Web* (Hyderabad: ACM), 57–58.

Jin, F., Wang, W., Zhao, L., Dougherty, E. R., Cao, Y., Lu, C.-T., et al. (2014). Misinformation propagation in the age of twitter. *IEEE Comput.* 47, 90–94. doi: 10.1109/MC.2014.361

Kempe, D., Kleinberg, J., and Tardos, É. (2003). "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC: ACM), 137–146.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Jose, CA: ACM), 420–429.

Li, C., Luo, J., Huang, J. Z., and Fan, J. (2012). "Multi-layer network for influence propagation over microblog," in *Pacific-Asia Workshop on Intelligence and Security Informatics* (Springer), 60–72.

Lin, Y.-R., Chi, Y., Zhu, S., Sundaram, H., and Tseng, B. L. (2008). "Facetnet: a framework for analyzing communities and their evolutions in dynamic networks," in *Proceedings of the 17th International Conference on World Wide Web* (Beijing: ACM), 685–694.

Mendoza, M., Poblete, B., and Castillo, C. (2010). "Twitter under crisis: can we trust what we rt?," in *Proceedings of the First Workshop on Social Media Analytics*, (Washington, DC: ACM), 71–79.

Nadamoto, A., Miyabe, M., and Aramaki, E. (2013). "Analysis of microblog rumors and correction texts for disaster situations," in *Proceedings of International Conference on Information Integration and Web-based Applications & Services* (Vienna: ACM), 44.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Riquelme, F., and González-Cantergiani, P. (2016). Measuring user influence on twitter: a survey. *Inform. Process. Manag.* 52, 949–975. doi: 10.1016/j.ipm.2016.04.003

Romero, D. M., Meeder, B., and Kleinberg, J. (2011). "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *Proceedings of the 20th International Conference on World Wide Web, WWW '11* (New York, NY: ACM), 695–704.

Sadikov, E., and Martinez, M. M. M. (2009). *Information Propagation on Twitter*. CS322 Project Report.

Saxena, A., Iyengar, S., and Gupta, Y. (2015). "Understanding spreading patterns on social networks based on network topology," in *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on* (Paris: IEEE), 1616–1617.

Seidman, S. B. (1983). Network structure and minimum degree. *Soc. Netw.* 5, 269–287. doi: 10.1016/0378-8733(83)90028-X

Smith, M. A., Rainie, L., Shneiderman, B., and Himelboim, I. (2014). Mapping twitter topic networks: from polarized crowds to community clusters. *Pew Res. Center* 20, 1–56. Available online at: https://scholar.googleusercontent.com/scholar?q=cache:eryksMy4YjgJ:scholar.google.com/&hl=en&as_sdt=0,5

Spiro, E., Irvine, C., DuBois, C., and Butts, C. (2012). "Waiting for a retweet: modeling waiting times in information propagation," in *2012 NIPS Workshop of Social Networks and Social Media Conference*, Vol. 12. Available online at: http://snap.stanford.edu/social2012/papers/spiro-dubois-butts.pdf

Tsur, O., and Rappoport, A. (2012). "What's in a hashtag? content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12* (New York, NY: ACM), 643–652.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science* 359, 1146–1151. doi: 10.1126/science.aap9559

Wang, Y., Wu, B., and Du, N. (2008). *Community Evolution of Social Network: Feature, Algorithm and Model.* arXiv:0804.4356.

Zeitung, S. (2016). Timeline der panik. Available online at: http://gfx.sueddeutsche.de/apps/57eba578910a46f716ca829d/www/

Zhuang, Y., and Yağan, O. (2016). Information propagation in clustered multilayer networks. *IEEE Trans. Netw. Sci. Eng.* 3, 211–224. doi: 10.1109/TNSE.2016.2600059

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 25 June 2019
doi: 10.3389/fdata.2019.00021

Temporal Mobility Networks in Online Gaming

Essa Alhazmi[1,2]*, *Nazim Choudhury*[1], *Sameera Horawalavithana*[1] and *Adriana Iamnitchi*[1]*

[1] Computer Science and Engineering, University of South Florida, Tampa, FL, United States, [2] Computer Science and Information Technology, Jazan University, Jizan, Saudi Arabia

This data-driven study focuses on characterizing and predicting mobility of players between gaming servers in two popular online games, Team Fortress 2 and Counter Strike: Global Offensive. Understanding these patterns of mobility between gaming servers is important for addressing challenges related to scaling popular online platforms, such as server provisioning, traffic redirection in case of server failure, and game promotion. In this study, we build predictive models for the growth and the pace of player mobility between gaming servers. We show that the most influential factors in predicting the pace and growth of migration are related to the number of in-game interactions. Declared friendship relationships in the online social network, on the other hand, have no effect on predicting mobility patterns.

Keywords: online games, mobility networks, online social network (OSN) activities, multiplayers online games, mobility diffusion

OPEN ACCESS

Edited by:
Roberto Interdonato, UMR9000 Territoires, Environnement, Télédétection et Information Spatiale (TETIS), France

Reviewed by:
Fabrizio Marozzo, University of Calabria, Italy
Sabrina Gaito, University of Milan, Italy

***Correspondence:**
Essa Alhazmi
ealhazmi@mail.usf.edu
Adriana Iamnitchi
anda@cse.usf.edu

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *01 April 2019*
Accepted: *03 June 2019*
Published: *25 June 2019*

Citation:
Alhazmi E, Choudhury N, Horawalavithana S and Iamnitchi A (2019) Temporal Mobility Networks in Online Gaming. Front. Big Data 2:21. doi: 10.3389/fdata.2019.00021

1. INTRODUCTION

Online gaming is not only a multi-billion dollar industry (Anderton, 2017) entertaining a large global population, but also a popular form of social interaction among millions of individuals. As online gaming exercises different types of sociability, such as shared activity (Zhuang et al., 2007; Merritt et al., 2013), tie and team formation (Alhazmi et al., 2017), trust formation (Depping et al., 2016), and long-term associations (McEwan et al., 2012; Jia et al., 2015), it becomes a rich source of temporal social interaction data that can be exploited for many computational social science questions. Data from online gaming environments were used to measure otherwise difficult to observe behaviors, such as cheating (Blackburn et al., 2014; Zuo et al., 2016), toxicity (Kwak et al., 2015), gold mining (Ahmad et al., 2009), and measuring online social capital (Molyneux et al., 2015).

Another human behavior that digital records from gaming environments can describe is mobility. Understanding players' mobility between gaming servers is important in multiple aspects, such as server provisioning, traffic redirection in case of server failure, and game promotion. In addition, the migratory patterns of players can be leveraged in modeling information dissemination or behavior adoption. For example, a player may introduce a new set of gimmicks, or may affect the server culture via positive or toxic social behavior.

In real world, human mobility has been shown to be a socially embedded phenomenon (Bilecen et al., 2018), which is affected by both socio-economic factors and the subjectivity of human behaviors (Barbosa Filho et al., 2011). Two important factors have been observed to contribute toward individual's migration decision (Blumenstock and Tan, 2016). Firstly, the extent to which a migrant is connected to communities at home and at the destination, and secondly, the strength and the support of destination ties in providing access to resources available in the destination environment (e.g., job information). The online gaming environment has different characteristics, and it is unclear whether the same arguments apply to player mobility.

This paper quantifies the importance of in-game interactions for a player's decision to migrate from one server to another within the same game. Players move to different servers over time due to various reasons, including technical performance (latency, computation speed), server/game preferences, peer familiarity, or personal endorsements. Previous studies showed that players tend to join games repeatedly with a set of familiar players with whom they shared past experience (Jia et al., 2015; Alhazmi et al., 2017). In this study, we specifically focus on the social interactions as a factor to characterize players' mobility patterns. We develop machine learning-based models to predict, first, the popularity of players over time with respect to the number of neighbors following their mobility patterns, and second, how fast a player moves between servers relative to the others. We present our results using data from two popular online games, Team Fortress 2 (TF2) and Counter Strike: Global Offensive (CSGO), that involve millions of players across a thousand servers over 4 months.

The contributions of this paper are 3-fold: First, it empirically characterizes mobility patterns of players across servers through the temporal mobility networks mechanism built upon their interactions. Second, it identifies the features relevant to the prediction of players' popularity, including early and late movers in the temporal mobility networks. Finally, it shows empirically that the growth and the pace of the mobility can be predicted.

2. DATASET

The gaming dataset used in this study was obtained from two sources: GameMe and the Steam Community. GameMe is a statistical reporting service that monitors real time playing activities on a collection of games. It provides APIs to collect real-time statistics of each player's gaming activity over a thousand gaming servers. The Steam Community is an online social network built on the Steam platform. It also provides APIs to extract players' list of friends, owned games, and game statistics for the most recent 48 h.

We focus on two highly popular games on the Steam platform, CSGO and TF2. CSGO is a tactical combat first person shooter video game where players compete as part of the terrorist or the counter-terrorist team. TF2 is a team-based and objective-oriented first-person shooter game, where players compete on two different teams and can pick a role from different categories, such as pyro, medic, scout, or soldier. The games have similar features including a wide variety of weaponry, maps, in-game voice chat, etc.

We collected data on friendship and temporal gaming interactions in these games through a web crawler that uses the APIs provided by Steam and GameMe. In CSGO, the duration of the collected data range from February 16 to August 9, 2017 (175 days), whereas in TF2, it is from February 16 to April 7, 2017 (51 days). The final dataset recorded over 13 million observations of 1.62 million players and 934 servers in CSGO. For TF2, the dataset contains over two million observations of 231 thousands players in 344 servers. BOT accounts and spectators (i.e., inactive players) were removed from the final dataset.

A game server is an authoritative host of game matches. Online multiplayer gaming environments, such as first-person/third-person shooter games, and role-playing games, provide a list of servers hosting active matches for players. Players can select server(s) and game matches based on different criteria, including server name, player count, match mode, and network latency.

Servers in online gaming have variable lifespans. The lifespan of a particular server is the duration of that server being active excluding intermittent downtime. In CSGO, the average server lifespan was 66 days (maximum 102 days) whereas in TF2, it was 39 days (maximum 51 days). Similarly, the average number of matches in CSGO was 1, 245 (maximum 7, 146) in comparison to 228 (maximum 3, 103) found in TF2. **Figure 1** shows the distributions of players in matches and servers for both CSGO and TF2.

From this dataset we constructed two social networks for each game: a friendship network based on declared relationships in Steam Community, and an interaction network based on the observed activities at gameme.com. The interaction network temporally connects players in the same match. Thus, an edge in the interaction network is undirected, weighted with the number of observed interactions between the players, and labeled with the list of timestamps when the players were observed in game. Only the active players observed in GameMe are included in the friendship network. **Table 1** summarizes the characteristics of interaction and friendship networks in both games.

3. TEMPORAL MOBILITY NETWORKS

In team-based online games, players often follow each other across servers in order to have fun, or to improve their

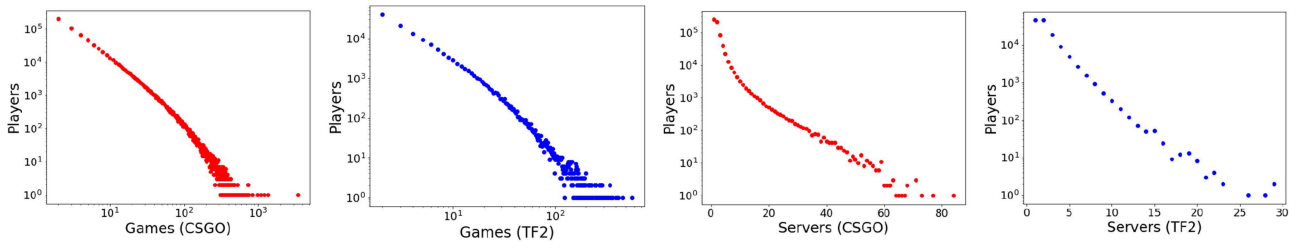

FIGURE 1 | Distribution of players across games and servers in both CSGO and TF2.

TABLE 1 | Data characteristics of the interaction and friendship networks.

Game	Period	Servers	Network	Players	Edges	Density	NCC
CSGO	02/16–08/09/2017	934	Interaction	1,106,652	27,415,330	4.48e-05	4,481
			Friendship	928,863	9,525,587	2.21e-05	2,068
TF2	02/16–04/07/2017	344	Interaction	224,922	6,920,096	2.74e-04	1,636
			Friendship	154,038	832,944	7.02e-05	4,258

NCC, # connected components.

TABLE 2 | Basic statistics of mobility networks in each game.

Games	# Networks	# Nodes per network			# of servers	# Networks per server		
		Min	Mean	Max		Min	Mean	Max
CSGO	2,816	2	202	8,434	705	1	4	15
TF2	1,316	2	51	2,937	323	1	4	10

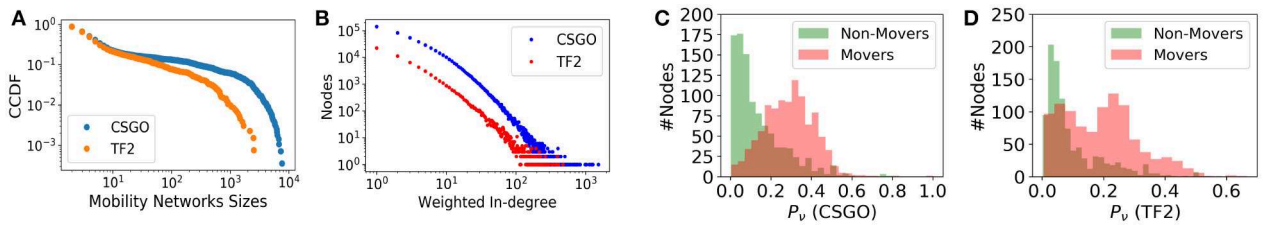

FIGURE 2 | (A) CCDF of temporal mobility networks sizes and **(B)** the average weighted in-degree distribution of mobility networks. Distribution of players' neighborhood ratio (P_{v_i}) in **(C,D)**.

skills and team performance. This study analyzes temporal interaction patterns among players to understand whether co-playing experience has impact on players' movements.

To capture the pattern of players following other players from one server to another, we model players' move as directed networks called temporal mobility networks built on top of the underlying interaction network. Intuitively, players' movements across servers can be explained by social interactions, common experiences related to the characteristics of the home server (e.g., over or under-populated, players' skill, etc.), personal factors (such as the player moving to a different geographical location), and many others. We only capture in this study—due to the inherent limitations of the dataset we collected—the possible reasons due to shared experiences, thus captured by the in-game interactions.

We define a temporal mobility network $G = (V, E)$ in which nodes are players and a directed link from node u to v exists if (i) v moved to server S at time t_m; (ii) u moved to server S at time $t_n > t_m$; and (iii) nodes u and v have preceding interactions at time $t_i < t_m$. In this context, node u is considered to adopt/follow node v in his movement to server S. We build a temporal mobility network based on the player movements in a given server. Therefore, for a given server, in the corresponding mobility network's context, "mover" and "adopter" will be used interchangeably in the rest of the text. The network is acyclic and only the earliest (first) move to a particular server by a pair of players is considered. The edges are time stamped to allow the study of temporal patterns.

Table 2 presents the main statistics on the mobility networks for both games and servers in games. Servers in the mobility networks are the destinations in the mobility process. Each server will attract disconnected networks of players. The number of disconnected groups (temporal mobility networks) per server for the two games are similar: on average, four groups join each server. The maximum number of mobility networks for two games was 15 and 10, respectively. However, larger groups move in CSGO (maximum is above 8,000 players) compared to TF2 (where maximum is under 3,000 players).

The distribution of networks' sizes is highly skewed across servers in both games. **Figure 2A** presents the complementary cumulative distribution functions (CCDF) of the mobility networks' sizes, calculated by considering the total number of nodes per network, and reveals heavy-tailed distributions. **Figure 2B** shows the average weighted in-degree distribution of players in the mobility networks.

In order to understand what might make players move to a different server, we calculated the ratio P_{v_i} for a player v_i between player's neighbors who moved with respect to all his neighbors as depicted in Easley and Kleinberg (2010). We weigh the number of neighbors by the number of interactions. **Figures 2C,D** represent

sampled distributions over 1,000 movers and non-movers in both games. It appears that the players who do not move have a lower ratio of players who moved in their neighborhoods.

4. PREDICTION TASKS

We have two prediction objectives: (i) identify the popular players in the early stage of the mobility networks formation, and (ii) distinguish early and late movers over the lifetime of the mobility networks. The underlying objectives behind these two classification tasks are complementary. First, the identification of popular players helps us detect whether a particular mobility network grows during our observational period. Second, the classification of early/late movers measures the speed of growth. We also examine the features that are most useful for the two prediction tasks.

4.1. Methodology

For the first task, we select temporal mobility networks with lifespans as long as our observation period. We extracted 178 such mobility networks in CSGO and 82 in TF2. We split the network lifespans into four quartiles. We define a node's popularity growth by comparing its in-degree as observed in the first quartile with its in-degree in the last quartile. We consider a node as being popular if its growth is higher than the median of the nodes' growth in that particular mobility network. The classification dataset is constructed by considering each node (player) as a prospective candidate of being popular or non-popular. Each datapoint is described by a set of features (listed in **Table 3**) constructed from the structural properties of each node in the mobility networks in the earlier stage. These features were used as input to a supervised learning algorithm, Random Forest, to predict the popular nodes in the later phase of the mobility network. The ratio of the training and testing datasets was 3:1 (75% training data, 25% testing data out of 140 thousands and 14 thousands instances in CSGO and TF2, respectively). The two datasets are nearly balanced: 57% in CSGO and 59% in TF2 are nodes in the non-popular category.

For the second task, predicting the pace of growth, we classify nodes in the mobility networks as early and late movers. We extracted a set of temporal-paths from each mobility network formed in this study using `pathpy` (Scholtes, 2017). A temporal path consists of a sequence of edges in the network ordered by the node migration time. In **Figure 3** (left), we present the distribution of temporal paths by their size. We notice CSGO consists of relatively longer chains of migrations than TF2. (Note that a node may end up joining multiple mobility networks at different times). We discriminate nodes between early and late considering their delay in movement compared to the median delay of the path they belong to: from the list of nodes in each temporal path, nodes having delays shorter than the median value are considered early movers. **Figure 3** (right) presents the distribution of median delays from all temporal paths extracted from the largest mobility network in each server of the two games. Interestingly, the sequence of movements observed in TF2 occurs at faster rate than in CSGO.

To predict the pace of gamers' movement, we extracted node-specific features described in **Table 3**. These features were used as input to the classifier to predict early (class 0) and late (class 1) adopters. We use a Long-Short Term Memory network for the classification task that consists of two blocks of memory-cells with two different layers of hidden units. The first layer contains 32 and the second one contains 8 units. We used the Adam algorithm with 0.001 learning rate

TABLE 3 | Features used in the pace (P) and growth (G) prediction tasks.

Features	Description	Task
Weight	Weight of edge to the parent node	P
In-degree	Node in-degree	G&P
In-degree$_{NF}$	Node in-degree from non-friends.	G
In-degree$_{F}$	Node in-degree from friends.	G
Out-degree	Node out-degree	G&P
Out-degree$_{NF}$	Node out-degree toward non-friends	G
Out-degree$_{F}$	Node out-degree toward friends.	G
Weighted In-degree	Sum of the weighted in-degree.	G
Adoption Rate	Total #adopters per unit time for the node	G
CC$_{out}$	CC of out-going edges	P
CC$_{in}$	CC of in-coming edges	G&P
CC-NF$_{in}$	CC of in-coming edges from non-friends	G
CC-F$_{in}$	CC of in-coming edges from friends	G
Time Lag/Adoption Duration	Interval between the first and last adoption	G
In-degree$_{parent}$	The in-degree of the node's parent	P
Out-degree$_{parent}$	The out-degree of the node's parent	P
CC-parent$_{out}$	The parent's CC$_{out}$	P
CC-parent$_{in}$	The parent's CC$_{in}$	P
isFriend	If node and its parent are friends	P

CC, clustering co-efficient.

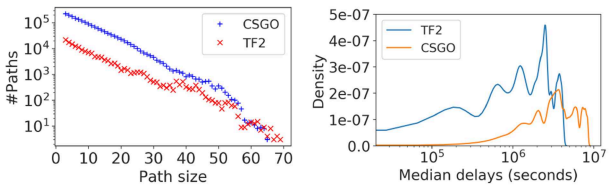

FIGURE 3 | Distribution of path sizes vs. the number of paths and the probability distribution of movement delay by considering temporal paths in mobility networks.

TABLE 4 | Prediction results for the popularity in the mobility networks of both games using Random Forest.

Game	Accuracy	Class	Precision	Recall	F1-score
TF2	0.73	1	0.54	0.72	0.61
		0	0.85	0.73	0.79
CSGO	0.75	1	0.62	0.76	0.68
		0	0.85	0.75	0.80

Class 1 denotes popular nodes and class 0 otherwise.

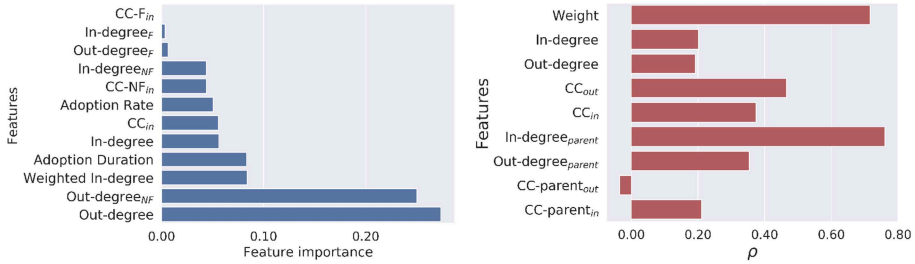

FIGURE 4 | (Left) Ranking of features importance in predicting the popularity. **(Right)** Features importance by Spearman's rank correlation coefficient ρ between the predicted outcome and the ground-truth in predicting the early/late movers.

TABLE 5 | Prediction results for the movement pace in the mobility networks of both games.

Classifier	Game	Accuracy	Class	Precision	Recall	F1-score
LSTM	TF2	0.70	1	0.70	0.72	0.71
			0	0.70	0.68	0.69
	CSGO	0.72	1	0.70	0.77	0.73
			0	0.73	0.76	0.70
RF	TF2	0.66	1	0.67	0.67	0.67
			0	0.65	0.65	0.65
	CSGO	0.69	1	0.67	0.76	0.71
			0	0.71	0.62	0.66

Class 1 denotes late adopters and class 0 otherwise. LSTM denotes Long-Short Term Memory and RF denotes Random Forest.

as optimizer. We split the temporal-paths set of the mobility networks into two sets: the training set includes 60% of the paths out of 1.7 millions and 155,281 paths in CSGO and TF2 consecutively, while the testing set contains the remaining 40% of paths.

4.2. Results

For classifying the popular players from unpopular ones, **Table 4** shows that Random Forest achieved high recall but low precision. Similarly, the prediction performance in CSGO outperformed the performances in TF2. The underlying reasons behind the better performance are the size of the classification datasets and rich feature values without significant overlap between positive and negatively labeled data points. The list of features are ranked according to their importance, calculated by the Random Forest classifier in CSGO, in **Figure 4** (left). It is noteworthy that similar results for TF2 are omitted due to space constraints. The out-degree of a node was found to be the most important feature in predicting the player's popularity. More surprisingly, the out-degree of a node toward his neighbors absent in its neighborhood of the friendship network were found to be most important features in both games. It is evident that friendship has minimal impact in predicting the number of players moving toward a new server following others.

For classifying early adopters from late ones, **Table 5** presents prediction performances demonstrated by both the Random Forest classifier and the LSTM-based neural network. As intuitively expected, the performance demonstrated by the LSTM has outnumbered the performance by the Random Forest classifier. The underlying reason behind the performance improvement by LSTM is its capability of learning the sequence data and consecutive dependency between feature values to successfully classify binary labels. Improved performance by LSTM also proves that in this context, recurrent neural networks can be a better classifier due to the temporal nature of the mobility network paths. Due to the improved performance by the LSTM over Random Forest classifier, the feature importance of the pace prediction tasks for both games were presented as the Spearman's rank correlation coefficient ρ between the predicted outcomes vs. the ground truth of the test data, as shown in **Figure 4** (right). It is noteworthy to mention that similar correlation was observed in TF2. The results demonstrate that the in-degree of a node's parent in the temporal path of the mobility network works as the best performing feature. Alternatively, the weighted interaction between the nodes and their parents with large number of followers are the principal determinants in predicting their pace of movement. On the contrary, the clustering co-efficient of the nodes' parents by considering their out-degree neighbors were found to have negative Spearman correlation in both games. Finally, the friendships between nodes and their parents represent only a small proportion of the instances in both games (2%). Thus, it is irrelevant to measure the correlation of the features incorporating the friendship networks.

5. SUMMARY

This study focused on modeling the temporal mobility patterns of online gamers by tracing the chronological movement of players between two servers. We developed two machine learning-based prediction strategies to predict the growth and pace (speed) in the mobility networks. Our main finding is that a player's mobility decision is affected by the co-players with the maximum number of interactions and not by the declared friends in the friendship network. This study can further be extended to explore the impact of community-level network structure over player's mobility across servers.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

The data were collected by EA. The experiments were conceived by AI and conducted by EA and SH. The data were analyzed and interpreted and the manuscript was written by EA, AI, and NC. All authors reviewed the manuscript.

FUNDING

This research was supported by the U.S. National Science Foundation under Grant No. IIS 1546453. EA was supported by a scholarship from Jazan University.

REFERENCES

Ahmad, M. A., Keegan, B., Srivastava, J., Williams, D., and Contractor, N. (2009). "Mining for gold farmers: automatic detection of deviant players in MMOGs," in *2009 International Conference on Computational Science and Engineering*, Vol. 4 (Vancouver, BC), 340–345.

Alhazmi, E., Horawalavithana, S., Skvoretz, J., Blackburn, J., and Iamnitchi, A. (2017). "An empirical study on team formation in online games," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17 (Sydney, NSW: ACM).

Anderton, K. (2017). *The Business of Video Games: A Multi Billion Dollar Industry*. Jersey City, NJ: Forbes.

Barbosa Filho, H. S., de Lima Neto, F. B., and Fusco, W. (2011). "Migration and social networks-an explanatory multi-evolutionary agent-based model," in *2011 IEEE Symposium on Intelligent Agent (IA)* (Paris: IEEE), 1–7.

Bilecen, B., Gamper, M., and Lubbers, M. J. (2018). The missing link: social network analysis in migration and transnationalism. *Soc. Netw.* 53, 1–3. doi: 10.1016/j.socnet.2017.07.001

Blackburn, J., Kourtellis, N., Skvoretz, J., Ripeanu, M., and Iamnitchi, A. (2014). Cheating in online games: a social network perspective. *ACM Trans. Internet Technol.* 13:9. doi: 10.1145/2602570

Blumenstock, J., and Tan, X. (2016). *Social Networks and Migration: Theory and Evidence From Rwanda*.

Depping, A. E., Mandryk, R. L., Johanson, C., Bowey, J. T., and Thomson, S. C. (2016). "Trust me: social games are better than social icebreakers at building trust," in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '16 (New York, NY: ACM), 116–129.

Easley, D., and Kleinberg, J. (2010). Cascading behavior in networks. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Jia, A. L., Shen, S., Bovenkamp, R. V. D., Iosup, A., Kuipers, F., and Epema, D. H. J. (2015). Socializing by gaming: revealing social relationships in multiplayer online games. *ACM Trans. Knowl. Discov. Data* 10, 11:1–11:29. doi: 10.1145/2736698

Kwak, H., Blackburn, J., and Han, S. (2015). "Exploring cyberbullying and other toxic behavior in team competition online games," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul: ACM), 3739–3748.

McEwan, G., Gutwin, C., Mandryk, R. L., and Nacke, L. (2012). "i'm just here to play games: social dynamics and sociality in an online game site," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12 (Seattle, WA: ACM), 549–558.

Merritt, S., Jacobs, A., Mason, W., and Clauset, A. (2013). "Detecting friendship within dynamic online interaction networks," in *Seventh International AAAI Conference on Weblogs and Social Media* (Boston, MA).

Molyneux, L., Vasudevan, K., and Gil de Zúñiga, H. (2015). Gaming social capital: exploring civic value in multiplayer video games. *J. Comput. Mediat. Commun.* 20, 381–399. doi: 10.1111/jcc4.12123

Scholtes, I. (2017). "When is a network a network?: multi-order graphical model selection in pathways and temporal networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS: ACM), 1037–1046.

Zhuang, X., Bharambe, A., Pang, J., and Seshan, S. (2007). *Player Dynamics in Massively Multiplayer Online Games*. School of Computer Science, Carnegie Mellon University, Pittsburgh, Tech. Rep. CMU-CS-07-158.

Zuo, X., Gandy, C., Skvoretz, J., and Iamnitchi, A. (2016). "Bad apples spoil the fun: quantifying cheating in online gaming," in *Tenth International AAAI Conference on Web and Social Media* (Cologne).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor and reviewer SG declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

frontiers in Big Data

ORIGINAL RESEARCH
published: 26 June 2019
doi: 10.3389/fdata.2019.00022

Link Definition Ameliorating Community Detection in Collaboration Networks

Saharnaz Dilmaghani[1], Matthias R. Brust[1], Apivadee Piyatumrong[2], Grégoire Danoy[1] and Pascal Bouvry[1]*

[1] Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Esch-sur-Alzette, Luxembourg,
[2] National Electronics and Computer Technology Center, A Member of NSTDA, Bangkok, Thailand

Collaboration networks are defined as a set of individuals who come together and collaborate on particular tasks such as publishing a paper. The analysis of such networks permits to extract knowledge on the structure and patterns of communities. The link definition and network extraction have a high impact on the analysis of collaboration networks. Previous studies model the connectivity in a network considering it as a binomial problem with respect to the existence of a collaboration between individuals. However, such a data consists of a high diversity of features that describe the quality of the interaction such as the contribution amount of each individual. In this paper, we have determined a solution to extract collaboration networks using corresponding features in a dataset. We define *collaboration score* to quantify the collaboration between collaborators. In order to validate our proposed method, we benefit from a scientific research institute dataset in which researchers are co–authors who are involved in the production of papers, prototypes, and intellectual properties (IP). We evaluated the generated networks, produced through different thresholds of *collaboration score*, by employing a set of network analysis metrics such as clustering coefficient, network density, and centrality measures. We investigated more the obtained networks using a community detection algorithm to further discuss the impact of our model on community detection. The outcome shows that the quality of resulted communities on the extracted collaboration networks can differ significantly based on the choice of the linkage threshold.

Keywords: network interactions, data-to-network, collaboration network, data analysis, community detection analysis

OPEN ACCESS

Edited by:
Andrea Tagarelli,
University of Calabria, Italy

Reviewed by:
Domenico Mandaglio,
University of Calabria, Italy
Pasquale De Meo,
University of Messina, Italy

****Correspondence:***
Saharnaz Dilmaghani
saharnaz.dilmaghani@uni.lu

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *02 April 2019*
Accepted: *04 June 2019*
Published: *26 June 2019*

Citation:
Dilmaghani S, Brust MR, Piyatumrong A, Danoy G and Bouvry P (2019) Link Definition Ameliorating Community Detection in Collaboration Networks. Front. Big Data 2:22. doi: 10.3389/fdata.2019.00022

1. INTRODUCTION

Collaboration networks are social structures which indicate the relationship between collaborators who perform on the same tasks. Collaboration is an essential component to define the success of today's knowledge sharing ecosystem (Huang et al., 2008) and establishment of innovation. In collaboration networks, nodes represent individuals (aka collaborators) and links between them imply a collaboration. The analysis of collaboration networks can reveal information about the most likely behavior of individuals and groups in the network (Jamali and Abolhassani, 2006) such as discovering the interaction patterns (Akbas et al., 2013; Long et al., 2014; Dilmaghani et al., 2019), the evolution of collaboration communities (Kibanov et al., 2013) and predictive models on the productivity and longevity of collaborations (Chakraborty et al., 2015).

One prominent property studied in the context of collaboration networks is the community structure of nodes (Pan et al., 2014). The discovery of communities, with dense intra-connections and comparatively sparse inter-cluster, can be beneficial for various applications such as discovering common research area of potential collaborators (Bedi and Sharma, 2016). Various network-based community detection algorithms are used for this purpose, e.g., *Louvain*'s algorithm (Blondel et al., 2008), Label Propagation Algorithm (LPA) (Zhu and Ghahramani, 2002).

Most collaboration data are stored in relational databases which are used to extract the collaboration networks to perform network analysis. The context of scientific collaboration networks has been initiated with the studies of Newman (2001a) and Newman (2001b). The network is defined such that the researchers are represented as nodes and the links constructed if at least one paper happened to be published by them. Other studies such as Chakraborty et al. (2015) have followed a similar generative approach to construct the collaboration network from the dataset. In a recent study (Sharma and Bhavani, 2019), a weighted scientific collaboration network has been proposed such that links are weighted by the number of papers. One drawback of previous studies is the elimination of other potential features that represent the collaborations (e.g., date, number of citations). The information which is attached to the data can substantially impact the underlying network representation and, therefore, the outcomes of network analysis (e.g., community detection). Thus the appropriate use of network analysis, substantially depends on choosing the right network representation (Scholtes, 2017), i.e., the definition of nodes and links (Butts, 2009). Besides, in some cases, the definition of the link also requires determining a *threshold* which can significantly alter the outcomes of network properties, e.g., network density (Faust, 2007).

In this paper, we investigated the definition of the fundamental research question of how and which network representation to choose for a given set of data. The drawback of previous studies is that they only consider the existence of a collaboration between individuals to connect them in the network. However, our work proposes a standardized method to produce networks from large and complex datasets. We define a method to construct scientific collaboration networks from the data considering different features describing the collaboration. Furthermore, we benefit from the scientific collaboration dataset of *National Electronics and Computer Technology Center* (NECTEC) to examine our method. Interestingly, our results indicate that identifying a network construction model leads to a less noisy yet well–shaped community structure network with high modularity score.

2. DATASET

We benefit from a particular collaboration database provided by the *National Electronics and Computer Technology Center* (NECTEC) that presents different projects and collaborations in the area of R&D[1]. The whole database is the knowledge management about projects within distinct deliverables where the key information is to know project contributors and contributions. The database consists of three datasets, each indicates a particular deliverable: *PAPER*, *PROTOTYPE*, and *IP* (intellectual property) conducted between July 2013 and July 2018.

The datasets of combined research teams information consist of approximately 8,000 records which correspond to the information of more than 2,300 projects. Detailed statistical information regarding each dataset is provided in **Table 1**. Overall, NECTEC has more than 1,000 members who are contributing to different deliverables with certain features that have been evaluated by the organization. For each researcher who collaborated on a contribution, a contribution percentage has been recorded. Another feature named IC–score which is designed by NECTEC, evaluates the scientific value and the outcome of contributions. For instance, producing a prototype in an industrial stage has a higher impact than one in the laboratory stage. For each project, the IC–score is divided between each contributor considering their individual participation in the project. Overall, each dataset of the deliverables contains (a) project ID, (b) collaborator's ID, (c) contribution percentage of a collaborator for each project, (d) IC–score of a collaborator for each project.

3. METHODOLOGY FOR LINK CONSTRUCTION

We propose a *collaboration score* function that takes into account the combination of features extracted from the dataset. The purpose is to quantify the contribution of researchers considering features describing the collaborations. The collaboration score is the key element to define the link in the network while nodes are co–authors. We introduce a *linkage threshold* (*LT*) on obtained collaboration scores. Thus, multiple networks are produced using various *LT* values.

We define the *collaboration score* function based on the features extracted from the NECTEC datasets which includes (a) the number of projects, (b) the contribution percentage of researchers, and (c) the IC–score of researchers. Given two researchers i and j worked on a mutual project p, i.e., (i, j), let n be the number of projects that i and j have collaborated, and $p_{k,i}$ and $p_{k,j}$ represent the contribution percentage of researcher i and j, respectively, for the kth project. Likewise, $s_{k,i}$ and $s_{k,j}$ indicate the IC–score of each researcher on the *kth* project. Hence, we determine the *collaboration score* function as follows.

$$f_{i,j} = \frac{1}{n}\Big(\frac{1}{2}\sum_{k=1}^{n}(p_{k,i} + p_{k,j}) + \frac{1}{2}\sum_{k=1}^{n}(s_{k,i} + s_{k,j})\Big) \qquad (1)$$

The function takes into account the average of IC–score and contribution percentage between any tuple of collaborators. The

[1]National Electronics and Computer Technology Center (NECTEC) (https://www.nectec.or.th/en/).

TABLE 1 | General overview of the datasets from NECTEC.

Deliverable type	# Researchers	# Projects	Cont. percentage	IC–score
PAPER	576	1717	$\mu = 22.22, \sigma = 19.73$	$\mu = 3.89, \sigma = 4.61$
PROTOTYPE	524	539	$\mu = 15.54, \sigma = 13.73$	$\mu = 9.41, \sigma = 10.75$
IP	489	630	$\mu = 25.15, \sigma = 24.42$	$\mu = 4.08, \sigma = 4.63$
Total	1,056	2,347	$\mu = 20.78, \sigma = 19.82$	$\mu = 5.81, \sigma = 7.73$

Contribution percentage (Cont. percentage) and IC–score are features extracted from the dataset and describe the collaboration.

LT, then, is defined such that it determines different levels of collaboration score in the network. The range of *LT* varies from 0 to 1, which is the normalized range of collaboration score. In a nutshell, increasing *LT* enlarges the number of collaborations.

The threshold values indicate links in the network between the nodes. We produce a set of networks considering various *LT*s. Algorithm 1 shows the pseudocode of the data transformation to networks. A relational dataset of collaborations is the input of the algorithm. The researchers are determined as nodes of the network. For each tuple of researchers, the collaboration score is measured (see line 4). In order to generate a network, links are produced considering a particular *LT* value. All collaborations that are less or equal than the level of the chosen threshold are determined as links in the network (see line 7). Considering various levels of *LT*, a set of networks is generated by the algorithm which is examined in section 4.

Algorithm 1: Network Extraction from Data

Input: D, scientific collaboration dataset
Output: $\mathcal{G}$, a vector of generated networks

```
1:  procedure TRANSFORM-TO-NETWORK(D)
2:      colList ← researchers from D
3:      for tuple(i, j) in colList do
4:          f.append ← collaborationScore(tuple(i, j))
5:          collaboration.append ← Concatenate tuple(i, j) and
    normalize(f)
6:      for LT in range(normalize(f)) do
7:          if collaboration.normalize(f) ≤ LT then
8:              nodes.append([i, j])
9:              links.append([tuple(i, j)])
10:         G ← Network(nodes, links)
11:         G.append G
12:     return G
```

4. RESULTS

Our proposed method has been employed on different deliverable types of the previously described NECTEC collaboration data. As a result of the extraction process, our method returns a set of corresponding collaboration networks. In the first stage, we exploit the distribution of the collaboration score (f) within each dataset. Next, we analyze the topology of the extracted networks given the different values of *LT* by measuring a set of network metrics. Furthermore, for each generated network, we identify the communities using the *Louvain* algorithm and evaluate their quality.

4.1. Data Processing

We exploit the histogram and cumulative distribution function (CDF) of f for each dataset of deliverables from NECTEC. **Figure 1** describes the frequency and distribution of the obtained f after normalization. The average (μ) of f for *PAPER*, *PROTOTYPE*, and *IP* are 0.24 [standard deviation (σ = 0.16)], 0.18 (σ = 0.12), and 0.3 (σ = 0.21), respectively. Furthermore, the figure also shows that the majority of collaborators have relatively low number of contribution. Nevertheless a small number of collaborators are strongly collaborating in various projects.

4.2. Topological Analysis

We analyze the topology and structure of extracted networks from each dataset by calculating a set of network metrics: degree, network density, transitivity, clustering coefficient, betweenness centrality, and closeness centrality. **Figure 2** describes the evolution of these metrics on a set of 41 networks while increasing *LT* from 0 to 1 with the step of 0.025.

The degree of a node in collaboration networks represents the number of direct collaborations for each individual. The average node degree of networks obtained from *PAPER* is 6.59, *PROTOTYPE* is 11.46, and *IP* is 5.71 which indicates that on average, teams in *PROTOTYPE* had significantly higher collaborations compared to others. As illustrated in **Figure 2**, the degree of extracted networks does not change significantly. The reason is after a certain threshold of *LT*, the number of new links which have been added to the network does not grow significantly while the number of nodes stays constant. A similar scenario occurs when measuring network density. The network density calculates the ratio of existing links to the number of all possible links in a network such that a density close to 0 identifies a sparse network while a density equal to 1 is a complete network. With *LT* close to zero, the network mostly consists of isolated nodes which explains why in all three datasets the network density is close to zero. Eventually, the density of the network increases slowly and remains steady. The reason is due to the high number of nodes compared to the number of collaborations between the nodes. This indicates the fact that in real-world collaboration networks each collaborator may only collaborate with a small number of collaborators, hence, the networks are considered as rather sparse.

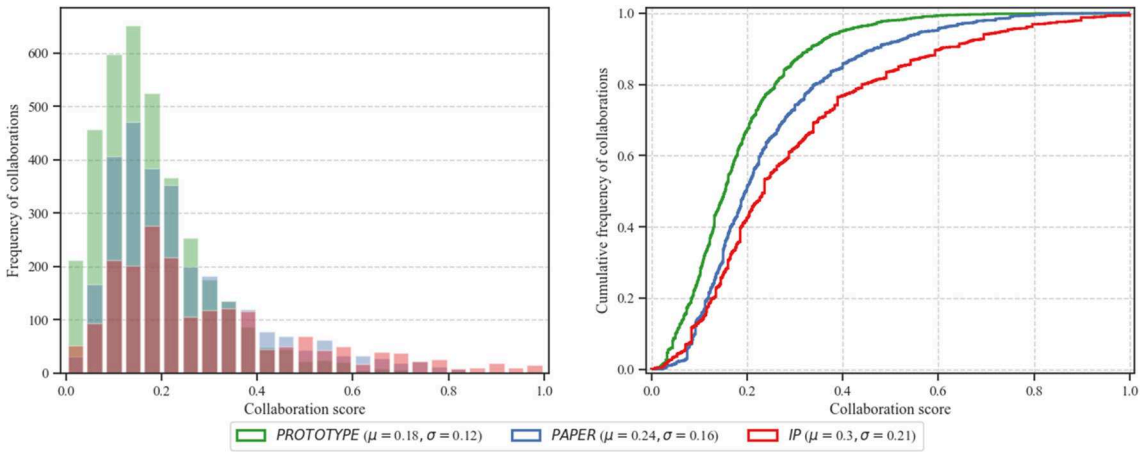

FIGURE 1 | The histogram and cumulative distribution function (CDF) of generated collaboration score (*f*).

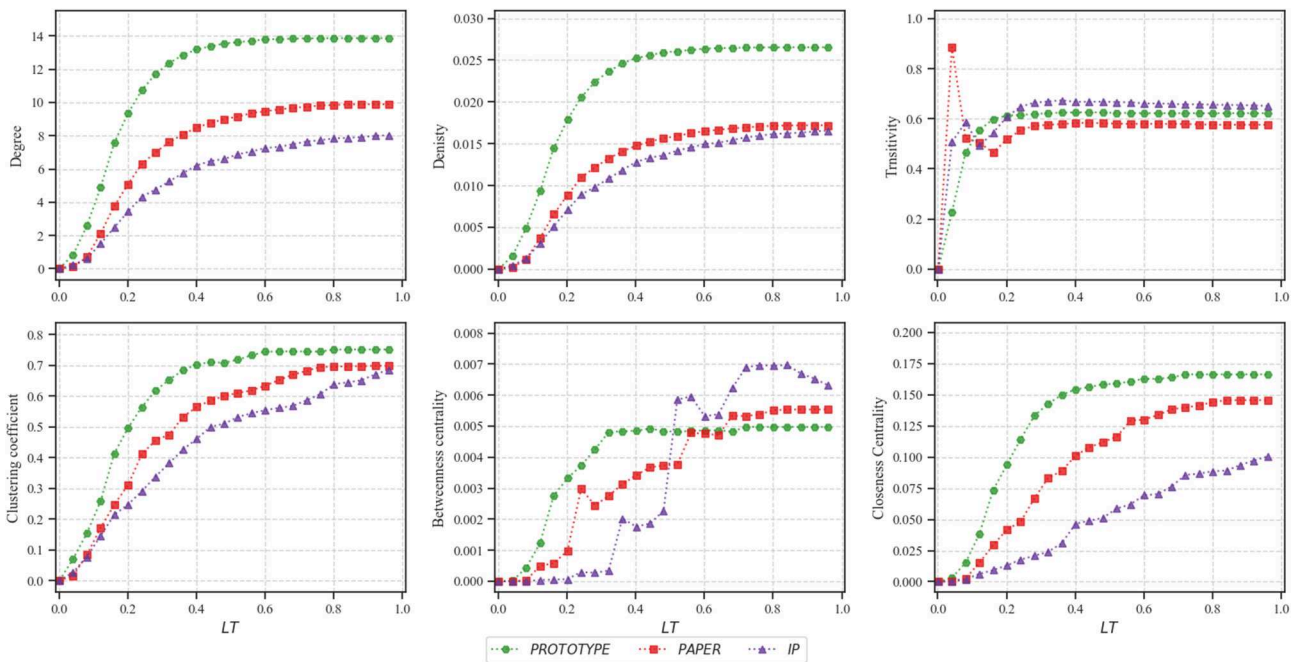

FIGURE 2 | Topological analysis of a set of 41 produced networks from each dataset while increasing *LT* from 0 to 1 by 0.025.

In order to get knowledge on the complexity of collaborations of each dataset, we calculate the transitivity and clustering coefficient of networks. Transitivity refers to the extent to which the relation that relates two nodes in a network that are connected by a link is transitive. Thus, it represents the symmetry of collaborations in our networks and forms triangles of collaborations. **Figure 2** illustrates fluctuations for networks constructed with lower *LT*, however, quickly it approaches a consistent value.

On the other hand, the clustering coefficient describes the likelihood of nodes in a network that tend to cluster together (Watts and Strogatz, 1998). The average clustering coefficient of produced networks is 0.44 for *PAPER*, 0.61 for *PROTOTYPE*, and 0.45 for *IP*. For a relatively high *LT* the clustering coefficient approaches approximately to 0.7. A possible explanation can be that contribution of at least three people happens often in scientific collaboration teams (Newman et al., 2001). Therefore, every collaboration that has three or more co–authors increases the clustering coefficient significantly.

Centrality measures indicate the importance of nodes in the network. We measure betweenness centrality and closeness centrality to analyze datasets. For a node, the betweenness is defined as the total number of shortest paths between every

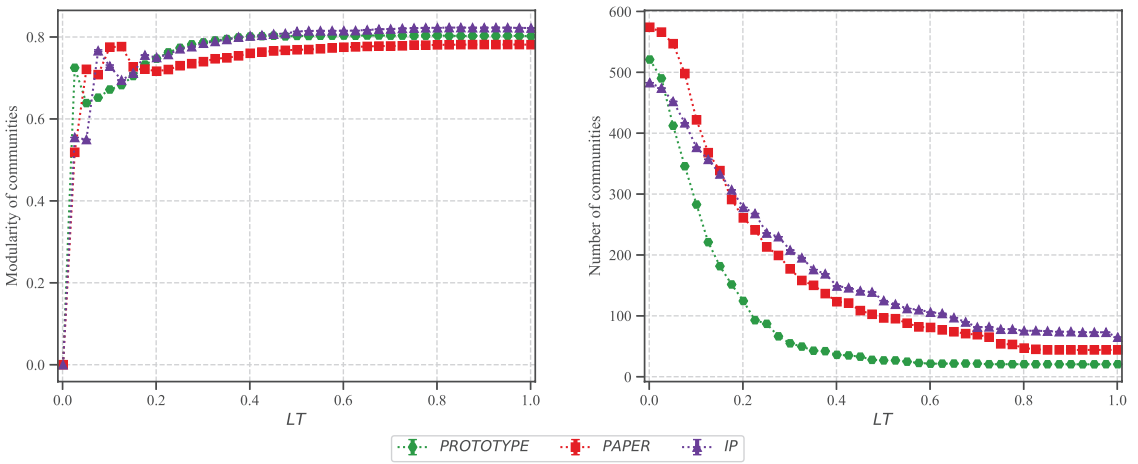

FIGURE 3 | Community detection analysis after implying *Louvain* algorithm on networks produced with different *LT* values. The community modularity score, and the number of clusters are the average of 200 experiments for 41 data points. The error bars are not visible because the standard error is very small.

pair of individuals in the network which pass through the node (Brandes, 2001). In other terms, it highlights collaborators who act as a bridge between different groups in a network.

Moreover, closeness centrality defines the closeness of a node to other nodes by measuring the average shortest path from that node to all other nodes within the network. Hence, the more central a node is, the closer it is to all other nodes (Sabidussi, 1966). All three datasets reach the highest closeness centrality after a certain threshold. However, each dataset reflects a considerably different growth function, such that *IP* follows a linear function after each evolution, *PROTOTYPE*, and *PAPER* are growing exponentially.

4.3. Community Detection Analysis

We imply *Louvain* community detection algorithm to evaluate *LT* on *collaboration score*. We extract communities of each network and measure the modularity and number of clusters. The modularity of communities illustrates the strength of connected nodes inside the same community compare to the community of a random graph (with the same size and average degree). The higher the modularity, the more the network is closer to a well-shaped community structure.

Figure 3 shows the average results of 200 experiments on each dataset including error bars. The figure shows that the modularity of all three datasets converges to relatively a high score of approximately 0.7 after a certain *LT*. It indicates that the produced collaboration networks have well–defined community structure compare to the random network of the same size. As illustrated in this figure, increasing *LT* does not affect the modularity after a particular point. For the lower LT (< 0.4), as also shown in **Figure 2** networks have a considerably lower density, thus, they are sparse. However, the score increases exponentially and becomes steady for all three datasets for $LT > 0.4$. On the other hand, increasing *LT* decreases the number of communities considerably. When networks are sparse (i.e., $LT \leq 0.2$) the number of communities is almost equal to the number of nodes.

Moreover, as illustrated in **Figure 3**, the modularity score increases significantly even for the low values of *LT* and reaches to its highest value before it decreases and becomes steady. On the other hand, the number of communities exponentially decreases. Therefore, the network obtained from $LT < 0.2$ has an extremely high number of communities. In a particular case for *PROTOTYPE*, the modularity increases and becomes steady with $LT > 0.4$, and similarly the number of communities become constant ($= 22$) with $LT > 0.5$. Furthermore, considering the growth of metrics for *PROTOTYPE* from **Figure 2**, all metrics are constant with $LT > 0.4$.

5. DISCUSSION AND CONCLUSION

The approach outlined in this paper infers collaboration networks of researchers within projects of an organization. Our method uses the features describing the collaborations of a research institute and quantifies them by applying a proposed *collaboration score* function.

Our results show that the quality of the detection of communities from the extracted collaboration networks can differ significantly by the choice of the linkage threshold. It turns out that a greedy increase of links and connections can lead to a noisy network structure where the *identity* of nodes could be affected by a large amount of superfluous connections. Consequently, our future work has to focus on the understanding of a networks preference toward a rich network while avoiding a noisy structure (Newman, 2018). Moreover, our experiments on the execution time of community detection indicate that increasing *LT* impacts the execution time of the algorithm. Hence, one option is to generate the network choosing a considerably low threshold while the modularity of communities is still at the highest possible value.

In this study we use a set of network metrics and the modularity score to evaluate communities of obtained networks. However, as future work we are looking at advancing our collaboration score model for network construction from relational data. Moreover, we consider identifying the optimum *LT* in order to recognize high quality communities within the obtained networks.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

SD developed the method and performed the computations and measurements. MB and PB were involved in planning and supervised the work. AP provided the datasets. MB, GD, and AP provided critical feedback.

FUNDING

This work is partially funded by the research programme UL/SnT-ILNAS on Digital Trust for Smart-ICT.

REFERENCES

Akbas, M. I., Brust, M. R., and Turgut, D. (2013). Social network generation and role determination based on smartphone data. *abs/*1305.4133.

Bedi, P., and Sharma, C. (2016). Community detection in social networks. *Wiley Interdiscip. Rev.* 6, 115–135. doi: 10.1002/widm.1178

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science* 325, 414–416. doi: 10.1126/science.1171022

Chakraborty, T., Ganguly, N., and Mukherjee, A. (2015). An author is known by the context she keeps: significance of network motifs in scientific collaborations. *Soc. Netw. Anal. Min.* 5:16. doi: 10.1007/s13278-015-0255-3

Dilmaghani, S. E., Piyatumrong, A., Bouvry, P., and Brust, M. R. (2019). Transforming collaboration data into network layers for enhanced analytics. *arXiv preprint* arXiv:1902.09364.

Faust, K. (2007). 7. very local structure in social networks. *Sociol. Methodol.* 37, 209–256. doi: 10.1111/j.1467-9531.2007.00179.x

Huang, J., Zhuang, Z., Li, J., and Giles, C. L. (2008). "Collaboration over time: characterizing and modeling network evolution," in *Proceedings of the International Conference on Web Search and Data Mining* (Palo Alto, CA: ACM), 107–116.

Jamali, M., and Abolhassani, H. (2006). "Different aspects of social network analysis," in *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)* (Washington, DC: IEEE), 66–72.

Kibanov, M., Atzmueller, M., Scholz, C., and Stumme, G. (2013). "On the evolution of contacts and communities in networks of face-to-face proximity," in *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing* (IEEE), 993–1000.

Long, J. C., Cunningham, F. C., Carswell, P., and Braithwaite, J. (2014). Patterns of collaboration in complex networks. *BMC Health Services Res.* 14:225. doi: 10.1186/1472-6963-14-225

Newman, M. (2018). Network structure from rich but noisy data. *Nat. Phys.* 14:542. doi: 10.1038/s41567-018-0076-1

Newman, M. E. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E* 64:016131. doi: 10.1103/PhysRevE.64.016131

Newman, M. E. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* 64:016132. doi: 10.1103/PhysRevE.64.016132

Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* 64:026118. doi: 10.1103/PhysRevE.64.026118

Pan, G., Zhang, W., Wu, Z., and Li, S. (2014). Online community detection for large complex networks. *PLoS ONE* 9:e102799. doi: 10.1371/journal.pone.0102799

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika* 31, 581–603. doi: 10.1007/BF02289527

Scholtes, I. (2017). "When is a network a network? multi-order graphical model selection in pathways and temporal networks," in *Proceedings of the ACM SIGKDD* (ACM), 1037–1046. doi: 10.1145/3097983.3098145

Sharma, A., and Bhavani, S. D. (2019). "A network formation model for collaboration networks," in *International Conference on Distributed Computing and Internet Technology* (Bhubaneswar: Springer), 279–294.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393:440. doi: 10.1038/30918

Zhu, X., and Ghahramani, Z. (2002). *Learning From Labeled and Unlabeled Data With Label Propagation*. Technical report, Citeseer.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 26 June 2019
doi: 10.3389/fdata.2019.00015

Applying Answer Set Programming for Knowledge-Based Link Prediction on Social Interaction Networks

Çiçek Güven and Martin Atzmueller*

Computational Sensemaking Lab, Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, Netherlands

Link prediction targets the prediction of possible future links in a social network, i. e., we aim to predict the next most likely links of the network given the current state. However, predicting the future solely based on (scarce) historic data is often challenging. In this paper, we investigate, if we can make use of additional (domain) knowledge to tackle this problem. For this purpose, we apply answer set programming (ASP) for formalizing the domain knowledge for social network (and graph) analysis. In particular, we investigate link prediction via ASP based on node proximity and its enhancement with background knowledge, in order to test intuitions that common features, e. g., a common educational background of students, imply common interests. In addition, then the applied ASP formalism enables explanation-aware prediction approaches.

Keywords: modeling social media, social network analysis, link prediction, answer set programming, knowledge-based

OPEN ACCESS

Edited by:
Andrea Tagarelli,
University of Calabria, Italy

Reviewed by:
Cristian Molinaro,
University of Calabria, Italy
Luca Maria Aiello,
Nokia, United Kingdom

****Correspondence:***
Çiçek Güven
c.guven@uvt.nl

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *03 April 2019*
Accepted: *28 May 2019*
Published: *26 June 2019*

Citation:
Güven Ç and Atzmueller M (2019) Applying Answer Set Programming for Knowledge-Based Link Prediction on Social Interaction Networks. Front. Big Data 2:15. doi: 10.3389/fdata.2019.00015

1. INTRODUCTION

Social interaction networks are mediated via social media in various forms and can be modeled using many diverse approaches, particularly using network theory. According to the idea of social interaction networks (Atzmueller, 2014), we adopt an intuitive definition of social media, regarding it as online systems and services in the ubiquitous web, which create and provide social data generated by human interaction and communication (Atzmueller, 2012). Specifically, we target link prediction for predicting future links in a network using background knowledge, formalized by logical formalisms. These allow to provide crucial domain knowledge: in scenarios when historic (link) data is still scarce—similar to the cold-start problem for link prediction— domain knowledge can complement structure-based link prediction. Thus, we utilize domain knowledge to enrich interaction networks, leading to knowledge-based feature-rich networks.

In this paper, we propose to use Answer Set Programming (ASP) for formalizing domain knowledge in order to enable hybrid link prediction (an approach that combines using the network itself as well as background knowledge to predict future links) in a social interaction network. ASP is a form of declarative programming that is used for difficult (NP hard) search problems, c. f., Lifschitz (2008). Here, ASP is relevant since it allows to specify interesting structures and patterns in a compact way, and due to its strength in including background knowledge by facts (and rules) intuitively. The ASP approach involves passing the (graph) structure and the conditions, and returns the (answer) set satisfying the conditions.

The proposed approach is exemplified using a real-world data set capturing networks of face-to-face proximity at a student event. In the interaction network, which is studied for the link prediction task, there are actors (nodes) who only start interacting with the other actors after a

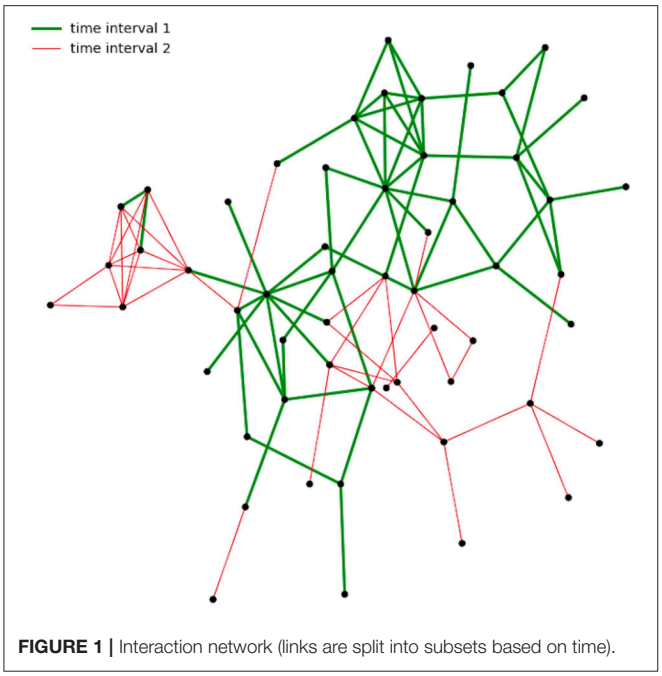

FIGURE 1 | Interaction network (links are split into subsets based on time).

while. In network terms, that means they are disconnected from the rest of the nodes given that a connection is there when there is an interaction. This is known in the literature as the cold start problem, (Leroy et al., 2010). An illustration of this is shown in **Figure 1**; links are split into two classes based on time. The links which correspond to interactions in the earliest interval, namely 'time interval 1' have color green, and are the thicker ones, whereas the color of the edges for the second interval is red.

In this example, we observe that there are nodes which only have connections with red colored edges; this means, that the corresponding interaction happened after the first interval. For those, we cannot apply, e. g., neighborhood features or path-based features for prediction, since no prior links/paths exist between these nodes and the others in the first interval. However, this data is complemented by attributive nodal information, which will be formalized as domain knowledge. Then, these might be informative to make predictions. That is, links between actors can be predicted based on a relation between actors and attributive information. With ASP, it is easy to incorporate such domain knowledge in the form of simple logical predicates and rules. That is why we consider it as an ideal tool in order to incorporate additional information.

It is important to note, that the purpose of this paper is not on analyzing specific patterns and insights on link prediction in social interaction networks, or to show that an ASP approach results in the best performance. Instead, we aim to provide a "proof of concept" of its applicability for link prediction, and to demonstrate its advantages like explainability and enabling a simple formalization and refinement of domain knowledge. The contribution of this paper is thus 2-fold:

1. We introduce the application of ASP as a novel approach for link prediction.
2. We demonstrate how to improve link prediction with contextual domain knowledge modeled using ASP.

The rest of the paper is structured as follows: section 2 discussed necessary background including basic definitions on graphs, and a brief introduction into ASP. After that, section 3 discusses related work. Next, section 4 outlines the proposed method using ASP for link prediction. Then, section 5 presents our results. Finally, section 6 concludes with a summary and outlines interesting directions for future work.

2. BACKGROUND

In this section, we define basic concepts in graph theory that are relevant for this paper. For further background in graph theory we refer the work of Diestel (2017). Next, we provide a brief overview on ASP.

2.1. Basic Definitions: Graph Theory and Link Prediction

A graph G is an ordered pair (V, E) consisting of a set of vertices (nodes) and a set of edges. An edge (u, v) consists of a pair of nodes u, v representing a relationship between them. A social network can be abstracted by a graph, where actors correspond to nodes and the links in between them corresponds to edges. A node v is a neighbor of (adjacent to) a node u if there is an edge (u, v) between them. $\Gamma(u)$ stands for the set of neighbors of a node u. Let $G = \{G_{t=o}, G_{t=1}, \cdots, G_{t=n}\}$ be a temporal sequence of evolving graphs where $G_{t=i} = (V_{t=i}, E_{t=i})$. For link prediction on such sequences, given $t = n$ the goal is to predict the structure of a graph in $t = n+1$, i. e., $G_{t=n+1}$. Specifically, we try to identify pairs (u, v), such that $u, v \in V_{t=n+1}$ and $(u, v) \in E_{t=n+1}$.

Prominent approaches for link prediction consider similarity scores between pairs of nodes, e. g., based on neighborhoods of pairs of nodes. Here, we will enhance link prediction based on neighborhood-based similarity scores with background knowledge. As one prominent neighborhood-based similarity score, we use the *Common neighbors* score: It counts the number of common neighbors of a pair of nodes. Given, (u, v) the pair of nodes under observation, the common neighbors can formally be written as:

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

2.2. Overview on Answer Set Programming

Answer Set Programming (ASP) (Niemelä, 1999) is a declarative problem solving approach; it is one of the three major logic programming families next to Prolog and Datalog. Logic programming is a programming paradigm mainly based on formal logic; such a program consists of facts and rules about the problem domain expressed as sentences in logical form. Given a problem, ASP aims to find one or several possible solutions; these are the so-called answer sets, i. e., all possible sets of facts that are consistent with the facts stated earlier) to the original problem (c. f., e. g., Gebser and Schaub, 2016; Kaufmann et al., 2016). This requires expressing the problem in a formal way. So, we transform and model the problem in the form of a logic program, which consists of rules and

variables. A special program, i. e., the *grounder* then eliminates all instances of the variables and replaces them by ground terms (which can be considered as "values," i. e., propositional atoms) in the language. This facilitates the application of the subsequent step, i. e., applying the answer set *solver*, which typically works on variable-free programs. Finally, the resulting propositional program, which is free of variables, only consists of propositional atoms. This is then the input to the solver which computes the answer sets. Those are all possible sets of facts that are consistent with the facts stated earlier to the original problem. For a more detailed discussion, we refer to e. g., (Niemelä, 1999; Gebser and Schaub, 2016; Kaufmann et al., 2016).

The ASP rules include user defined predicates and variables, as in the following example for common neighbors (*CN*):

```
CN(X, Y, Z) :-edge(X, Y), edge(X, Z),
not edge(Y, Z), Y!=Z.
```

In this notation, "," means "and," ":-" means "if," and "not" stands for negation. Here, "`CN`," and "`edge`" are examples of user defined predicates, which can be true or false for object(s) represented by a specific term replacing a user defined variable(s) such as '`(1,2)`'. The rules without any conditions are called facts. Our example rule is used to formalize the following information: X is a common neighbor of a pair of distinct vertices Y and Z, if there are edges between pairs X, Y and X, Z but not between Y and Z. The if symbol '`:-`' is omitted for the facts, so that '`edge(1,2).`' is a fact.

The solution to a problem is called an "answer set", which consists of propositions that are supposed to be true in the answer set. A solution to the above rule and the two facts '`edge(1,2).`', and '`edge(1,4).`' is the answer set containing these facts and the propositions '`CN(1,2,4).`', and '`CN(1,4,2).`'.

We used ASP to enhance link prediction in a network with background knowledge and used a small data set for this proof of concept. However, ASP is designed for NP-hard problems as stated earlier and finds its applications in large instances of industrial problems, since it offers a rich representation language and high performance solvers; some recent applications are listed in Falkner et al. (2018). Some examples of ASP solvers that are considered to be efficient are Smodels (Syrjänen and Niemelä, 2001), WASP (Dodaro, 2013), Clasp (Gebser et al., 2012) and Clingo (Gebser et al., 2014b). Clingo[1] itself combines a powerful grounder (Gringo) with Clasp (for solving) into an integrated system. For ease of use, and due to its efficiency (e. g., Guyet et al., 2018; Schäpers et al., 2018), we utilized Clingo in the context of this paper.

3. RELATED WORK

The focus of link prediction is the dynamics and mechanisms in the creation of links between the parties in social networks (Liben-Nowell and Kleinberg, 2003). The purpose is to learn a model for predicting the links accurately. There is already a large body of research for link prediction concerning *online* social networks, e. g., (Katz, 1953; Adamic and Adar, 2003; Liben-Nowell and Kleinberg, 2003; Murata and Moriyasu, 2007; Lü and Zhou, 2010; Scholz et al., 2013, 2014) considering neighborhood-based and path-based measures. A first comprehensive fundamental analysis was done by Liben-Nowell and Kleinberg (2003), where the link prediction problem was defined as the search to carefully predict edges that will be added to a given snapshot of a social network during a given interval, using network proximity measures. This shows a strong connection to the approach to this paper, while we apply a novel approach, i. e., ASP for performing the search. In addition, we also include domain knowledge for a knowledge-based link prediction approach, also tackling the common cold start problem in link prediction (Leroy et al., 2010).

Link predictions can be used for different prominent applications: recommending and suggesting promising interactions between two individuals in such a social network (Li and Chen, 2009; Papadimitriou et al., 2011), the prediction of missing links, (Liben-Nowell and Kleinberg, 2003), and improving collaborative filtering (Huang et al., 2005). In this paper, we mainly focus on the perspective of utilizing link prediction for recommendation and collaborative filtering, while also target explainability and transparency of the predictions which is also facilitated by our proposed approach.

To the best of the authors' knowledge, the idea of merging Answer Set Programming and link prediction in the context of social networks is new. De Raedt et al. (2007) studied a probabilistic version of Prolog, to discover links in large network of biological concepts. The probabilistic Prolog would then aim to compute the success probability for the existence of a link between nodes such as genes and diseases. Furthermore, there have been earlier studies relating ASP and social network analysis: Jost et al. (2012) modeled a way to suggest new interactions related to events in a social network for a personal assistant of the network platform (EasyReach) which monitors interactions. A study relating social networks with ASP in the privacy and security context is described in Hu et al. (2013). There, multiparty access control for online social networks is studied. Marra et al. (2014, 2016) studied properties of social networks, and information diffusion in Social Network Analysis. They applied ASP for analyzing properties of social networks, in a multi-social-network setting. The study of Seo et al. (2013) also combines social network analysis and logic programming. In that study a high-level graph query language SociaLite based on Datalog is proposed, due to its expressive power and efficiency, an tested on real life social graphs. We have a similar motivation in terms of the ease of use, and of expressivity, where we target explicative link prediction in the context of social networks, utilizing topological network information as well as attributive relations. Furthermore, explainable social network analysis is a further feature of the ASP-based approach, where first approaches in the context of explicative data mining (Atzmueller, 2017, 2018) have been discussed by Masiala and Atzmueller (2018a,b).

[1] Available at: https://potassco.org/

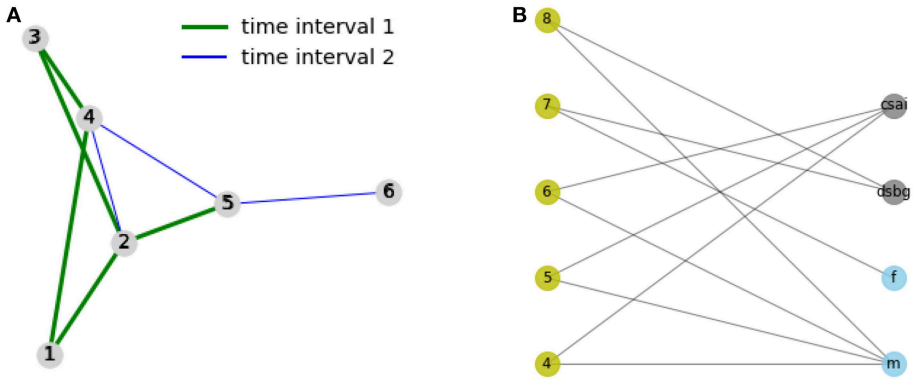

FIGURE 2 | Example: interaction and attributive/knowledge-based networks. **(A)** Interaction network (links are split into subsets based on two different time intervals. **(B)** Graph capturing attributive data: students on the left, and attributive information on the right.

4. METHODS

In the following, we outline our method for link prediction using ASP. The main strength of ASP is its intuitive way to state a problem, also allowing to scale the problem up easily, and the availability of computationally powerful ASP solvers. For this study, the former two points are more relevant since in our application context we utilize a relatively small data set so far. As an ASP solver, we use Clingo (Gebser et al., 2014a) embedded in Python.

Below, we will first illustrate our approach via a small hypothetical example. Then we will describe the data set, and finally we will discuss our findings on the data set.

4.1. Example

We consider a social interaction network between students as actors, and attributive information collecting information such as gender, affiliation, and area of study of the students. For those, we provide two according network structures: one indicating the interactions, the other (bimodal) one modeling information of the students as actors in the network. Regarding the left network, the graph G shown in **Figure 2** represents interaction between actors at an event, split into two time frames. The edges $E_1 = \{(1,2), (2,3), (1,4), (3,4), (2,5)\}$ represent the interactions in the first interval T_1, $E_2 = \{(2,4), (1,5), (3,5), (2,6)\}$ represent the interactions that happened in the second time interval T_2 afterwards. The bipartite graph G_A shown on the right of **Figure 2** represents the choices of the attributive information provided as background knowledge. The nodes $4, 5, 6, 7, 8$ represents students, and the nodes f, m represent their gender (f: female, m: male). The nodes *dsbg*, *csai* are standing for the master programs the students are enrolled to, e. g., "Cognitive Science and Artificial Intelligence" or "Data Science for Business and Governance". The edges in E_2 are aimed to be predicted by using information coming from prior interactions captured by E_1 as well as captured by background knowledge given G_A.

The following code predicts a link between a pair of nodes in G for T_2 if they have two common neighbors in G during T_1 or G_A. That is, a link is predicted for a node pair u, v without an existing link in the interaction graph for T_1 $[(u,v) \notin E_1]$ when they are similar in terms of their neighbors, or when they are similar based on their respective attribute values, in this case having the same gender and following the same program are necessary. Then the code compares the links in G for T_2, E_2 (which we can see as a test set), and returns the matches between the predicted links E_{2pred} and the test set. The ASP program is composed of two parts: The *facts* describing the networks, and the *rules* for inferring the prediction.

```
#const n=2.
#const n_attrib =2.

% ASP facts
% Defining the networks/graphs
node(1..6). % Nodes of the interaction graph
edge(1,2). edge(1,4). edge(2,5). edge(2,3). edge(3,4). % Edges,
    first time interval
test(2,4). test(4,5). test(5,6). % Edges, second time interval (test
    set)
% Nodes and edges of the attributive graph:
node_attrib(4..8). node_attrib(csai). node_attrib(dsbg). node_attrib(f).
    node_attrib(m).
edge_attrib(5,csai). edge_attrib(8,dsbg). edge_attrib(7,dsbg).
edge_attrib(4,csai). edge_attrib(6,csai). edge_attrib(5,m).
    edge_attrib(4,m).
edge_attrib(8,m). edge_attrib(7,f). edge_attrib(6,m).

% ASP rules
% This is an undirected graph, hence there is symmetry in edges.
edge(Y,X) :- edge(X,Y).
edge_attrib(Y,X) :- edge_attrib(X,Y).
% X is a common neighbor of Y and Z where they are not connected.
c(X,Y,Z) :- edge(X,Y), edge(X,Z), not edge(Y,Z), Y!=Z.
c_attrib(X,Y,Z) :- edge_attrib(X,Y), edge_attrib(X,Z), not
    edge_attrib(Y,Z), Y!=Z.
% a link is predicted when there are 2 common neighbors in the
    interaction graph
cn_lp(Y,Z) :- node(Y), node(Z), not edge(Y,Z), Y!=Z, n=#count{X:c(X,
    Y,Z)}.
```

```
% a link is predicted when there are 2 common neighbors in the
    attributive graph
cn_lp(Y, Z) :- node(Y), node(Z), not edge(Y, Z), Y!=Z, n_attrib =#count{
    X: c_attrib (X, Y, Z)}.
test (Y, X) :- test (X, Y).
% The match rule compares the predicted set of links with the test set
match(X, Y) :- test (X, Y), cn_lp(X, Y).

#show cn_lp/2.
#show match/2.
```

This example is designed in such a way that, there is 100% overlap between the predicted links and the test set. Thus, the output is:

```
match(2,4) match(4,5) match(5,6) match(6,5) match(5,4) match(4,2) cn_lp
    (4,2) cn_lp (2,4) cn_lp (1,3) cn_lp (3,1) cn_lp (4,6) cn_lp (5,6) cn_lp
    (6,4) cn_lp (5,4) cn_lp (6,5) cn_lp (4,5)
```

It is easy to see that–depending on the formalization of the predicates and rules used in the ASP program, the answer set itself can accommodate helpful explanations of why a link was predicted. This can be supported by a trace of the applied rule structure, e. g., utilizing a reconstructive explanation methodology (Wick and Thompson, 1992; Atzmueller and Roth-Berghofer, 2010), complemented by further background knowledge and/or context information from the network structure.

Since the graph derived from the attributive information connects the students to other parameters, a prediction based on its common neighbors will predict links between students when constructed as above. The rules can be modified in such a way that for a constant n, where $\Gamma_A(x)$ stands for the neighborhood of node x in G_A, E_{2pred} stands for the predicted edges for T_2:

$\forall u, v, x, y \in V \mid (u, v) \notin E_1, |\Gamma_A(u) \cap \Gamma_A(v)| = n \implies \forall x \in \Gamma_{G_1}(u) \backslash \Gamma_{G_1}(v)$, $(x, v) \in E_{2pred}$ and $\forall y \in \Gamma_{G_1}(v) \backslash \Gamma_{G_1}(u)$, $(x, u) \in E_{2pred}$.

4.2. Data Set Description

For this study, we utilized a real life data set, which had been collected during a student event. This included information on face-to-face interactions and attributive information including gender, academic degree, age group, area of studies.[2] For that, active proximity tags based on Radio Frequency Identification technology (RFID-chips) developed by the SocioPatterns Collaboration[3] were applied. These are able to detect face-to-face interactions at large scale, using the radio packets exchange between two devices provided that the devices are in a distance of 1–1.5 m, and the parties remained in contact for at least 20 s. An interaction ends, when no packets are detected within a 20 s interval. The sensor data is used to construct social interaction networks capturing offline interactions between people. For more details on the data preprocessing, we refer to Barrat et al. (2010).

[2]Participants were invited to wear RFID proximity tags. Study participants also gave their written informed consent for the use of their data in scientific studies. Data were collected in an anonymous way.

[3]http://www.sociopatterns.org/

TABLE 1 | Network characteristics: Attributive network, and the interaction network in two time intervals.

Characteristics	G_A	G_1	G_2
Number of nodes	124	47	40
Number of edges	456	59	38
Density	6%	5.5%	4.9%

For constructing feature-rich networks, we utilized the data set focussing on its two components: One is capturing the interactions collected via sensors between students, and the other one is based on the given attributive information. The interaction data set contains data from 56 students attending the student event. First, using the proximity contacts, we generated a social interaction network. Then, an edge $\{u, v\}$ is created, if a face-to-face contact with a duration of at least 20 s among participants u and v was detected. There were 340 interactions with the lower bound of 20 s, the maximal interaction length being 1,042 s (on average 69.5 s), over the course of 8 hours. After removing duplicate edges (only the first interactions are kept between parties in case there were more than one interaction), only 97 edges are left. These edges are split into two subsets E_1, and E_2 with corresponding graphs G_1, G_2 while the order based on time is preserved with ratio $(6:4)$.

The attributive data set is relevant to capture the similarities based on the attribute values, which is structured as a bipartite graph G_A. One of the partitions consists of the student ids (anonymous) and the other partition consists of attributes about gender, age group, academic degree, area of studies. For instance, there is a node corresponding to value 'female' for the gender attribute, "Data Science" for the area of studies. There is an edge between the node representing a student and the nodes representing the attribute. This resulted in a data set consisting of two columns corresponding to the sets of nodes representing the partition, where each row represents an edge. There are 456 rows in this data set, and 124 vertices partitioned into two sets as described above for students and attributive information of respective sizes 76 and 48. Some characteristics of the graphs G_A, G_1, G_2 can be seen in **Table 1**. The sparsity in the interaction graphs makes link prediction a hard problem there.

5. RESULTS AND DISCUSSION

We first focused on the cold start problem. There are 9 nodes which showed up in the second time interval. There are 14 edges for these nodes in E_2. For any pair of vertices in the graph, if there is an edge between them in the test set, then that is an actual positive, otherwise actual negative. A match between the predicted and actual positive is a true positive. We predicted edges for the newcomers based on a simple similarity measure in G_A. We predicted an edge between a pair of students if there had been no edge between them in G_1, and they had n common

neighbors in G_A graph where n is in $\{4, 5\}$. This implied 7 true positives, and 65 predicted positives out of 315 possible edges in G_2. These imply a precision of 10.7%, a recall of 50% and an F1 score of 18%.

The following rules are used to augment the common neighbor method described by an example above, with the formalized background knowledge coming from the attributive information. An edge is predicted between a pair of vertices in V_2, if there is no such edge in G_1, these vertices are distinct and they have four or five common neighbors in G_A.

```
#const n_attrib1 =4.
#const n_attrib2 =5.
attributive_edge (Y,X):- attributive_edge (X,Y).
c_attrib (X,Y,Z) :- attributive_edge (X,Y), attributive_edge (X,Z), not
    attributive_edge (Y,Z),Y!=Z.
pn(Y,Z) :- e_2_node(Y), e_2_node(Z), Y!=Z, not e_1_edge(Y,Z), n_attrib1
    =#count{X: c_attrib (X,Y,Z)}.
pn(Y,Z) :- e_2_node(Y), e_2_node(Z), Y!=Z, not e_1_edge(Y,Z), n_attrib2
    =#count{X: c_attrib (X,Y,Z)}.
```

We chose the number of common neighbors as the similarity metric, since it is s standard metric, and it is also very explainable and interpretable, as also discussed above. Using ASP we first predicted links based on common neighbors only–utilizing the interaction network. We predicted a link between a pair of non-adjacent nodes in G_1, when they have n common neighbors, for different values of n, and compared these with $G_2 = (E_2, V_2)$, treated as the ground truth for this problem. Given that the network G has low density all edges considered (i. e., the data is not balanced across classes) accuracy is not a good metric, hence we look into precision recall and F1 score only, see **Table 2**. There are 38 edges in G_2, which is the size of actual positives, $\binom{40}{2} = 780$ possible edges, and 742 actual negatives, that is the difference between possible and existing edges.

We see in **Table 2**, link prediction solely on interaction data does not work well with the common neighbors metric: We only achieve an F1 score of 11.0%. We noted earlier, one limitation of using interaction data is the cold start problem. Here $V_1 \backslash V_2 = 16, V_1 \backslash V_1 = 9$. That is a big community change, 16 people left and 12 new people arrived. That is a potential explanation to the performance. However, even if we neglect the cold-starters, focusing on the intersection of nodes in G_1 and G_2 then we still obtain rather comparable bad results, which we also verified using the linkpred package[4] using the standard common neighbors, preferential attachment and rooted pagerank metrics. When we start adding new information based on the attributive information in G_A, the number of true positives starts increasing as well. In our results, we see an increase on the cold-starters of 18%, leading to an overall F1 measure of 15.4% which clearly outperforms the baseline. A refined exploitation of the background knowledge can then lead to further improved evaluation metrics, e. g., by including social theories and extending our applied simple common neighbors strategy.

Link prediction is quite difficult for this data set, due to sparsity and the cold start problem. Given the results, we can argue common neighbors is not a very strong predictor for future links for this data set. With the attributive information data we see an increase in false positives (wrongly predicted links) decreasing the precision, and F1 but since correctly predicted links also increased, recall increases slightly. It is important to note that we so far applied only a simple strategy for formalizing background knowledge: The purpose here is to propose an approach to the link prediction problem, not to find the best performing method. We aim to refine the model using the attributive information by formalizing appropriate background knowledge, in order to explore options for improving link prediction in future work.

TABLE 2 | Link prediction evaluation metrics.

Number of common neighbors	Graph used for prediction	True positives	Predicted positives	Precision	Recall	F_1
≤ 4	G_1	6	31	19.4%	7.7%	11.0%
≤ 4	G_1	16	170	9.4%	42.1%	15.4%
$\in \{4, 5\}$	G_A					

We treated any attribute value equally here, where as in reality, some attributes will be more informative than others. Also, more common attribute values might be less informative. The results can then be improved by exploring those. Overall, ASP remains an ideal way to incorporate and test that additional background knowledge with its flexibility. For example, ASP can be used to incorporate further insights about the population studied by looking further into background data. Some observations whose impact into link prediction could be tested here are the following: for students who consider becoming an entrepreneur, other common characteristics are: being Male, being between 18 and 25 years old, and having a degree in Data science Bachelor. Also among people who are between the ages 26 and 35, "paid job at an existing company" is a more common feature than for example "consider becoming an entrepreneur."

A further advantage of the proposed approach is given by its explainability: The answer set itself describes the "solutions" for link prediction. By tracing back the applied rules used for inferring the answer set, specific choices can be illustrated for link prediction, i. e., which factors were responsible for establishing a specific link. In that way, ASP provides a transparent and interpretable approach for link prediction, integrating feature-rich networks complemented by background knowledge. In section 4.1, a hypothetical example showcasing link prediction enhanced with an attributive graph is given. That is, pairs of nodes in the interaction network are predicted to be linked, if they are similar in terms of their past behavior (captured by the existing number of common neighbors) or sharing attributes such as gender or area of study in the attributes network. This requires considering the topological information of both graphs, i. e., the list of nodes and edges, as well formalizing the rules defining common neighbors. Other rules then define link prediction based on the number of common neighbors in both graphs, as below.

[4]https://github.com/rafguns/linkpred/

```
#const n=2.
#const n_attrib =2.

% a link is predicted when there are 2 common neighbors in the
    interaction graph
cn_lp(Y, Z) :- node(Y), node(Z), not edge(Y, Z), Y!=Z, n=#count{X:c(X,
    Y, Z)}.
% a link is predicted when there are 2 common neighbors in the
    attributive graph
cn_lp(Y, Z) :- node(Y), node(Z), not edge(Y, Z), Y!=Z, n_attrib =#count{
    X: c_attrib (X, Y, Z)}.
```

These rules simply state for a pair of distinct nodes Y, Z, which are not linked by an edge, a link is predicted between them when they have `n` (or `n_attrib`)common neighbors in the interactions or the attributive graph, respectively. Of course, the names can always be chosen to be more descriptive so that the logical statement resembles natural language more (`link_predicted_based_on_common_neighbors` instead of `cn_lp`). Basic understanding of logical expressions is enough to make sense of the rules. The answer set then itself captures the respective `cn_lp` facts, together with all those (new) facts that were applied in the solving process. Taken together, this then supplies an explanation as a trace of the applied rules, which can of course be complemented with further information such as, e. g., topological features in the form of statistical network descriptors.

6. CONCLUSIONS

In this paper, we proposed using ASP to incorporate background knowledge to the link prediction problem, which is not possible using some other approaches, for example, using standard social network analysis methods, e.g., proximity-based or path-based methods. In that way, we also introduced the application of ASP as a novel approach for link prediction. We explored that using a real-world data set capturing networks of face-to-face proximity at a student event: The dataset is relatively sparse, thus the link prediction problem is quite difficult, and becomes even more challenging in the context of the cold start problem. Therefore, the application of background knowledge proved to be especially relevant.

Our experiments using a standard common neighbors approach for link prediction showed, that providing background knowledge considerably improved the prediction performance. Furthermore, we showed how ASP can be conveniently applied in such a knowledge-based approach, in particular also relating to explanation-aware techniques since the result of ASP, i. e., the answer set, can be directly mapped to extensive explanations on the link prediction method. In this paper, we thus specifically demonstrated how to improve link prediction with contextual domain knowledge modeled using ASP – as a "proof of concept" of its applicability for link prediction. Furthermore, we demonstrated its advantages like explainability and enabling a simple formalization and refinement of domain knowledge.

For future work, we aim to extend and refine the model further, investigating different theory-based formalizations, like structural holes and social capital (Burt , 2002), and social roles (Scripps et al., 2007). Further future directions include the characterization of unpredicted links and extending the features used for the prediction toward temporal relationships, the order of the interactions, and information coming from the duration of conversations, as well as the existence of multiple edges–toward advanced link prediction in feature-rich complex interaction networks (Interdonato et al., 2019).

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

ÇG and MA conceived of the idea, interpretation of the data and wrote the manuscript. ÇG implemented the method and ran the experiments.

ACKNOWLEDGMENTS

This work has been partially supported by the German Research Foundation (DFG) under grant AT 88/4-1.

REFERENCES

Adamic, L. A., and Adar, E. (2003). Friends and neighbors on the web. *Soc. Netw.* 25, 211–230. doi: 10.1016/S0378-8733(03)00009-1

Atzmueller, M. (2012). Mining social media: key players, sentiments, and communities. *WIREs Data Min. Knowl. Disc.* 2, 411–419. doi: 10.1002/widm.1069

Atzmueller, M. (2014). Data mining on social interaction networks. *arXiv:1312.6675v2*.

Atzmueller, M. (2017). "Onto explicative data mining: exploratory, interpretable and explainable analysis," in *Proceedings of Dutch-Belgian Database Day* (Eindhoven).

Atzmueller, M. (2018). "Declarative Aspects in Explicative Data Mining for Computational Sensemaking," in *Proceedings of International Conference on Declarative Programming (DECLARE)* (Heidelberg: Springer).

Atzmueller, M., and Roth-Berghofer, T. (2010). "The mining and analysis continuum of explaining uncovered," in *Proceedings of SGAI International Conference on Artificial Intelligence (AI-2010)* Cambridge.

Barrat, A., Cattuto, C., Colizza, V., Pinton, J., den Broeck, W. V., and Vespignani, A. (2010). High resolution dynamical mapping of social interactions with active RFID. *PLoS ONE* 5:e11596. doi: 10.1371/journal.pone.0011596

Burt, R. S. (2002). "The social capital of structural holes," in *The New Economic Sociology: Developments in an Emerging Field*, Vol. 148, eds F. Mauro, R. C. GuillÈn, E. Paula, and M. Marshall (New York, NY: Russell Sage Foundation), 90.

De Raedt, L., Kimmig, A., and Toivonen, H. (2007). "ProbLog: a probabilistic prolog and its application in link discovery," in *Proceedings of IJCAI* (Hyderabad), Vol. 7, 2462–2467.

Diestel, R. (2017). *Graph Theory*. *Graduate Texts in Mathematics*, 5th Edn. Berlin; Heidelberg: Springer.

Dodaro, C. (2013). Engineering an Efficient Native ASP Solver. *TPLP* 13.

Falkner, A., Friedrich, G., Schekotihin, K., Taupe, R., and Teppan, E. C. (2018). Industrial applications of answer set programming. *KI-Künstliche Intel.* 32, 165–176. doi: 10.1007/s13218-018-0548-6

Gebser, M., Kaminski, R., Kaufmann, B., and Schaub, T. (2014a). Clingo=ASP + control: preliminary report. *CoRR* abs/1405.3694.

Gebser, M., Kaminski, R., Kaufmann, B., and Schaub, T. (2014b). Clingo= ASP+ control: preliminary report. *arXiv [preprint]. arXiv:1405.3694*.

Gebser, M., Kaufmann, B., and Schaub, T. (2012). Conflict-driven answer set solving: from theory to practice. *Artif. Intel.* 187, 52–89. doi: 10.1016/j.artint.2012.04.001

Gebser, M., and Schaub, T. (2016). Modeling and language extensions. *AI Mag.* 37, 33–44. doi: 10.1609/aimag.v37i3.2673

Guyet, T., Moinard, Y., Quiniou, R., and Schaub, T. (2018). "Efficiency analysis of ASP encodings for sequential pattern mining tasks," in *Advances in Knowledge Discovery and Management* (Berlin; Heidelberg: Springer), 41–81.

Hu, H., Ahn, G.-J., and Jorgensen, J. (2013). Multiparty access control for online social networks: model and mechanisms. *IEEE Trans. Knowl. Data Eng.* 25, 1614–1627. doi: 10.1109/TKDE.2012.97

Huang, Z., Li, X., and Chen, H. (2005). "Link prediction approach to collaborative filtering," in *Proceedings of 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York, NY: ACM), JCDL '05, 141–142.

Interdonato, R., Atzmueller, M., Gaito, S., Kanawati, R., Largeron, C., and Sala, A. (2019). Feature-rich networks: going beyond complex network topologies. *Appl. Netw. Sci.* 4:4. doi: 10.1007/s41109-019-0111-x

Jost, H., Sabuncu, O., and Schaub, T. (2012). "Suggesting new interactions related to events in a social network for elderly," in *Proceedings of International Workshop on Design and Implementation of Independent and Assisted Living Technology.*

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika* 18, 39–43. doi: 10.1007/BF02289026

Kaufmann, B., Leone, N., Perri, S., and Schaub, T. (2016). Grounding and solving in answer set programming. *AI Mag.* 37, 25–32. doi: 10.1609/aimag.v37i3.2672

Leroy, V., Cambazoglu, B. B., and Bonchi, F. (2010). "Cold start link prediction," in *Proceedings SIGKDD* (New York, NY: ACM), 393–402.

Li, X., and Chen, H. (2009). "Recommendation as link prediction: a graph kernel-based machine learning approach," in *Proceedings of ACM/IEEE JCDL* (New York, NY: ACM), 213–216.

Liben-Nowell, D., and Kleinberg, J. M. (2003). "The link prediction problem for social networks," in *Proceedings of CIKM* (New York, NY: ACM), 556–559.

Lifschitz, V. (2008). "What is answer set programming?" in *Association for the Advancement of Artificial Intelligence* (Palo Alto, CA), 1594–1597.

Lü, L., and Zhou, T. (2010). Link prediction in weighted networks: the role of weak ties. *EPL Europhy. Lett.* 89:18001. doi: 10.1209/0295-5075/89/18001

Marra, G., Ricca, F., Terracina, G., and Ursino, D. (2014). "Exploiting answer set programming for handling information diffusion in a multi-social-network scenario," in *Proceedings of JELIA* (Berlin; Heidelberg: Springer), 618–627.

Marra, G., Ursino, D., Ricca, F., and Terracina, G. (2016). Information diffusion in a multi-social-network scenario: framework and ASP-based analysis. *Knowl. Inf. Syst.* 48, 619–648. doi: 10.1007/s10115-015-0890-z

Masiala, S., and Atzmueller, M. (2018a). "First perspectives on explanation in complex network analysis," in *Proceedings of BNAIC* (Den Bosch: Jheronimus Academy of Data Science).

Masiala, S., and Atzmueller, M. (2018b). "Towards explainable complex network analysis," in *Proceedings of Dutch-Belgian Database Day* (Belgium: Hasselt University).

Murata, T., and Moriyasu, S. (2007). "Link prediction of social networks based on weighted proximity measures," in *Web Intelligence* (Fremont, CA: IEEE), 85–88.

Niemelä, I. (1999). Logic programs with stable model semantics as a constraint programming paradigm. *Ann. Math. Artif. Intel.* 25, 241–273. doi: 10.1023/A:1018930122475

Papadimitriou, A., Symeonidis, P., and Manolopoulos, Y. (2011). "Friendlink: link prediction in social networks via bounded local path traversal," in *Proceedings of CASoN* (Fremont, CA: IEEE), 66–71.

Schäpers, B., Niemueller, T., Lakemeyer, G., Gebser, M., and Schaub, T. (2018). "ASP-based time-bounded planning for logistics robots," in *Proceedings of International Conference on Automated Planning and Scheduling (ICAPS).*

Scholz, C., Atzmueller, M., Barrat, A., Cattuto, C., and Stumme, G. (2013). "New insights and methods for predicting face-to-face contacts," in *Proceedings of ICWSM* (Palo Alto, CA: AAAI Press).

Scholz, C., Atzmueller, M., and Stumme, G. (2014). "On the predictability of recurring links in networks of face-to-face proximity," in *Proceedings of WWW 2014 (Companion)* (New York, NY: IW3C2/ACM).

Scripps, J., Tan, P.-N., and Esfahanian, A.-H. (2007). "Exploration of link structure and community-based node roles in network analysis," in *Proceedings of 7th IEEE International Conference on Data Mining (ICDM)* (Washington, DC: IEEE Computer Society), 649–654.

Seo, J., Guo, S., and Lam, M. S. (2013). "SociaLite: datalog extensions for efficient social network analysis," in *Proceedings of IEEE International Conference on Data Engineering (ICDE)* (Washington, DC: IEEE), 278–289.

Syrjänen, T., and Niemelä, I. (2001). "The smodels system," in *Proceedings of International Conference on Logic Programming and NonMonotonic Reasoning* (Berlin; Heidelberg: Springer), 434–438.

Wick, M. R., and Thompson, W. B. (1992). Reconstructive expert system explanation. *Artif. Intel.* 54, 33–70. doi: 10.1016/0004-3702(92)90087-E

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 28 June 2019
doi: 10.3389/fdata.2019.00019

Including Vulnerable Populations in the Assessment of Data From Vulnerable Populations

Latifa Jackson[1,2]*, *Caitlin Kuhlman*[3], *Fatimah Jackson*[2,4] and *P. Keolu Fox*[5]

[1] Department of Pediatrics and Child Health, College of Medicine, Howard University, Washington, DC, United States, [2] W. Montague Cobb Research Laboratory, College of Arts and Sciences, Howard University, Washington, DC, United States, [3] Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, United States, [4] Department of Biology, College of Arts and Sciences, Howard University, Washington, DC, United States, [5] Department of Pediatrics, College of Medicine, University of California, San Diego, San Diego, CA, United States

Data science has made great strides in harnessing the power of big data to improve human life across a broad spectrum of disciplines. Unfortunately this informational richesse is not equitably spread across human populations. Vulnerable populations remain both under-studied and under-consulted on the use of data derived from their communities. This lack of inclusion of vulnerable populations as data collectors, data analyzers and data beneficiaries significantly restrains the utility of big data applications that contribute to human well-ness. Here we present three case studies: (1) Describing a novel genomic dataset being developed with clinical and ethnographic insights in African Americans, (2) Demonstrating how a tutorial that enables data scientists from vulnerable populations to better understand criminal justice bias using the COMPAS dataset, and (3) investigating how Indigenous genomic diversity contributes to future biomedical interventions. These cases represent some of the outstanding challenges that big data science presents when addressing vulnerable populations as well as the innovative solutions that expanding science participation brings.

Keywords: inclusion, genomics, COMPAS, African Americans, algorithmic fairness, Samoa, criminal justice

OPEN ACCESS

Edited by:
Yelena Mejova,
Institute for Scientific Interchange, Italy

Reviewed by:
Kyriaki Kalimeri,
Institute for Scientific Interchange, Italy
Daniela Paolotti,
Institute for Scientific Interchange, Italy
Rumi Chunara,
New York University, United States

***Correspondence:**
Latifa Jackson
latifa.f.jackson@gmail.com

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *02 April 2019*
Accepted: *03 June 2019*
Published: *28 June 2019*

Citation:
Jackson L, Kuhlman C, Jackson F and Fox PK (2019) Including Vulnerable Populations in the Assessment of Data From Vulnerable Populations. Front. Big Data 2:19. doi: 10.3389/fdata.2019.00019

INTRODUCTION

The past several decades have seen great improvements in the scale of data collected, analyzed and used to improve human life. This data expands our understanding of social science, business and biomedical science among other disciplines (Murdoch and Detsky, 2013). It is able to find patterns in extensive data sets, and use those observations to test hypotheses and predict phenomena. Unfortunately, the vast majority of this data has been focused on a small subset of global ethnic diversity and culture. In particular, the dominance of European and Asian data science culture has skewed both data science analysis and inference. In the analysis of social science data, the global consequences of social exclusion are costly, including exacerbating poverty, reducing human capital and diminishing culturally coherent solutions which could be more easily adopted in communities (Tangcharoensathien et al., 2018). Meanwhile, a scan of biomedical data shows consistent inequalities in the inclusion of those vulnerable populations that are at most risk for having health disparities (Popejoy and Fullerton, 2016). Indeed the need for better data collection, reporting, analysis and interventions on the environmental and social determinants of health is pressing, and improvement may influence patient health outcomes (Lu et al., 2018).

There is however a critical absence of discussion around the role that vulnerable populations themselves play in articulating the data science problems. This perspective is crucial for designing analytical solutions and most relevantly in interpreting findings from their culturally competent lens. This lack of engagement leads to loss of agency in problem identification, under-representation in the analytical data science space and ultimately poorer solutions that fail to take into account the lived experiences of vulnerable populations.

We seek to use three case studies to explore ways that data scientists, human geneticists, and biological anthropologists can collaborate to encourage the participation of vulnerable populations in data science to address locally relevant questions, generate novel datasets, and learn how to address systemic biases in currently existing datasets. Here we highlight three approaches to involving members of vulnerable populations in the collection, analysis, and interpretation of data derived from vulnerable populations. Each case study explores how vulnerable/ethnic minority populations can be engaged to contextualize data inference within a social context to bring better understanding. In the first example, we introduce work by researchers at Howard University, in remediating the paucity of genetic knowledge about African-descended groups and ameliorating their consequent health vulnerabilities. The second example describes our experience training vulnerable populations about criminal justice data to gain their insights into what that data might mean for their communities. The third example looks at the impact of the exclusion of vulnerable Polynesian populations in variant identification for obesity pharmacogenomics based on biomedical sample collection. Each case study highlights how vulnerable population can make meaningful contributions to the assessment and interpretation of big data. While these case studies do not provide a complete solution to the lack of participation of vulnerable populations in their well-being, they do chart a roadmap that show how engagement can lead to higher quality data generation, new dataset construction, and community trust-building and empowerment in data science.

CASE 1: GENERATING GENOMIC DATA EQUITY IN VULNERABLE POPULATIONS

In spite of the origins of humanity in continental Africa and the ancient, historical, and contemporary dispersions of African peoples via at least four major Diasporas worldwide (Zeleza, 2005), very little is known about the genesis, extent, and duration of African genetic variability. This scientific reality increases the vulnerability of modern African-descended populations and limits their ability to benefit from new advances in genomic sciences (Sirugo et al., 2019). The benefit of modern genomics is primarily through the development of comprehensive and inclusive reference databases to which newly discovered variants can be compared and contextualized (Jackson, 2018). Effective genetic medicine depends upon such reference databases. Without appropriate reference standards, the push for subpopulation relevant precision medicine invariably falls short and the targeted population remains under-served and sometimes dis-served. Furthermore, there is an ongoing urgent need to see Africa on its own terms as terrain of the endogenous and the indigenous, a locale of emergence whether its genetics, morphology, ecology, language/linguistics or culture (writ large) (Keita personal communication 2019). This can only be done by integrating scientists and other scholars from the understudied indigenous communities to actively participate in the collection, analysis, interpretation, and dissemination of genetic knowledge about their own people.

Currently, the reference databases are predominantly Eurocentric, as are the genomic priorities in mainstream western science. This is expected and not problematic in and of itself since the majority of researchers are of European descent. However, this imbalance presents issues when the client base is ethnically and geographically diverse and decidedly non-European. These groups can only benefit from the existing databases to the extent that they maintain genomic profiles congruent with North Atlantic European patterns. In other cases, there may also be population-specific mutations in understudied populations that cause health disparities that go under-diagnosed in African-descended groups (Sirugo et al., 2019) primarily because they differ in mutation patterns from the majority European population. Finally, the interpretation of African-derived genomic data suffers if knowledgeable African and African-descent scholars are not involved in the analysis, contextualization, and practical application of the resulting data.

At Howard University, we have launched three African Genome Projects. In the Atlantic African Diaspora Genome Project our aim is to provide historically-informed, geospatially diverse sampling to the study of African-descended peoples in the America hemisphere. The Atlantic African Diaspora Genome Project aims to collect samples from North, South, and Central America and the Caribbean ($N = 1{,}000$ samples) (Mann, 2001). The second of the African Genome Projects focuses on continental Africa ($N = 10{,}000$ samples). This project aims to effectively capture the magnitude of genomic variability in the homeland of humanity by focusing on the various terrestrial biomes on the continent and sampling proportionately from each based on the level of existing ecological complexity. The third phase of our data base development efforts is the Red Sea African Diaspora Genome Project ($N = 1{,}000$). This effort aims to trace the migration pathways of African-descended groups eastward across the Red Sea and Indian Ocean (Harris, 1971, 2003; Cooper, 1977; Alpers, 1997; Ewald, 2000). This database will allow researchers to track relevant African signals to the east of Africa, following the many well-established historical routes out of the continent. The W. Montague Cobb Research Laboratory has been in the forefront of the development of augmented genomic data bases to characterize African genomic diversity. Our hope is that by acquiring and interpreting representative African genomic diversity, we will develop the capacity to reconstruct the evolutionary history of African descended peoples worldwide and that of our species, and in so doing, increase the access of African-descended populations to the immediate and long term benefits of genomic knowledge.

As the largest and most well-known historically Black university, Howard University is uniquely poised to initiate

this study. In this preliminary collection effort, several weeks were devoted to community education and recruitment. We assembled a team of primarily African-descended interdisciplinary researchers to design and implement the project. These scholars included colleagues in the life sciences, medical sciences, social sciences, humanities, and computational sciences. On the day of collection, within 8 h, 463 non-hospitalized individuals freely provided informed consent for access to their DNA, salivary microbiome, ancestral background, and general health information. African Americans from North America and the Caribbean and continental Africans were the pre-identified target populations. While a total of 25 nationalities and 35 ethnicities were represented in this first sample, 260 of our participants (56.2%) self-reported as North American Black or African American. Participant data were subdivided based upon ancestral origins. Three hundred forty-eight participants (75.2%) contribute to the Atlantic African Diaspora Genomes Database, 31 participants (6.7%) from continental Africa will be included in the Continental African Database, and 75 participants (16.2%) will go into the Red Sea African Diaspora Database. Nine participants (2.0%) identified their ancestral origins in Eurasia or Oceania and were assigned to a Control cohort.

The vulnerability of African-descended populations to missing insights and benefits of advances in genomic sciences is particularly acute for continental Africans. These populations retain high levels of regionally specific genetic diversity. Yet, the efforts to date have generally been based on opportunistic sampling of Africans. Consequently, for more continental Africans, current genomic knowledge is particularly non-illuminating. Without carefully constructed reference genomic databases that integrate ecological, anthropological, and historical data what is currently known presents a weak profile of continental African substructure, population stratification, and migration history. The ability to reconstruct the biological histories of Africans remains limited and with only a few selected African populations studied, our knowledge of continental African diversity lacks the nuanced regional and ethnic specificity that characterizes European reference databases. If African genetic diversity was studied systematically, we expect it to yield as much, if not more, geospatial and ethnic complexity as Europe. In particular, since humans have had a protracted residence in Africa, there have been ample opportunities for regional adaptations to emerge, and extensive migrations throughout the continent have occurred over hundreds of thousands of years.

Very limited genomic studies of indigenous Africans have been done and even fewer are publicly available and integrated in general reference databases for comparative research purposes. Although the 1000 Genomes Project (1000GenomesConsortium, 2015) reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping, Africa was not adequately represented given its status as the homeland of our species, continent of longest residence, and therefore the indigenous peoples with the greatest expected collective accumulations of acquired mutations. Although the 1000 Genomes Project characterized a broad spectrum of genetic variation, in total over 88 million variants [84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants], all phased onto high-quality haplotypes, coverage of the non-European populations from whom many American lineages can be traced remains insufficient, particularly given the long presence of African-descended individuals in this hemisphere, the extensive opportunities for gene flow with non-Africans, and the continentally diverse origins of these early Africans to America. This was noted over 20 years ago Jackson (1996, 1997, 1998), yet the deficiency in our databases persists.

For Diaspora African populations such as Legacy African Americans who have been in the country for 11–16 generations and are an amalgamation of African peoples with modest gene flow from non-Africans, more information is known about the European-derived components of their genomes than is revealed about their larger, residual African components. This limits the value of current genomic medicine in these individuals. Furthermore, since much of this admixture with Europeans occurred within the context of African enslavement in the seventeenth, eighteenth, and nineteenth centuries, the European-derived segments in the genomes of African Americans tend to be truncated in length and random in their dispersion in the genome. Although an estimated 30% of Legacy African American men carry Y-chromosome haplogroups found more commonly in North Atlantic Europe, the rest of their genomes also reflect this historical European admixture, but the distribution of these genes is non-uniform and piecemeal.

The historically most important diaspora for African people has been inadequately studied. This is the intra-African diaspora. Unfortunately, however, knowledge of the genomic and demographic ramifications of intra-African migrations, adaptations, and admixtures are lacking. For the vast majority of continental Africans and African descended people outside of Africa, the more African their lineage, the less current genomic knowledge is able to reveal about their disease vulnerabilities, ancestry, and phenotypic markers. The ramifications of inadequate studies of African genomic diversity are not limited to individuals of African descent. In previous studies we have shown that personalized genomic testing can have multiple beneficial educational ramifications for tested individuals (Johnson and Jackson, 2015). Even a small amount of data on one African ancestry has been shown to stimulate additional interest in this history and the science behind it. In the absence of relevant information, these opportunities, for example in enhanced interest in STEM, are diminished.

Our approach can remediate this situation and bring equity to our genomic knowledge by capturing a wider diversity of human variability. A first step has been to increase the number of diverse non-European individuals in the reference databases, creating truly comprehensive and representative databases for meaningful world-wide comparisons and as a platform for broadly beneficial precision medicine. A particular need is to capture the high variability of indigenous Africans in each of the terrestrial biomes of the continent, since much of this genomic diversity is not yet characterized. This has to be done in an intentional model-based sampling method, not haphazardly or

simply opportunistically. Sampling should also not be biased toward hunter-gatherer groups to the exclusion of agriculturalists and post-agriculturalists in Africa. We need clear hypothesis-driven sampling strategies for studying genomic diversity in non-European peoples and these need to be coupled with relevant historical, anthropological, ecological, and geospatial data. These data should be integrated using computational biology to generate algorithms that accurately characterize the populations under study, reconstruct their histories, and provide predictive data for their enhanced survival.

To generate sophisticated bioinformatic profiles of African genomic diversity, we need to identify the salient population substructure of African and African-descended donors so that their genomics can be appropriately contextualized. Using ethnogenetic layering in the Atlantic African Diaspora, we have hypothesized that microethnic groups such as the Gullah/Geechee of the South Carolina Lowcountry may retain unique genomic markers as a consequence of their antiquity (compared to other African American groups), relative geographic and cultural isolation (Jackson, 2008), and endogamous mating preferences (Caldwell, personal communication). This is not only due to the geographical distances between these groups, but also because of their differing population histories, migration stories, admixture patterns, dietary exposures, and other relevant variables.

In collaboration with Helix and National Geographic, our strategy is to divide the completed data bases equally into Discovery and Replication cohorts for the integrative testing of hypotheses regarding admixture, ancestry, migration, selection, disease susceptibility/resistance. Once completed, these databases will provide the scientific community with greater referencing depth with expected positive ramifications for a public increasingly interested in and dependent upon the results of genomic interpretations for their health and well-being. This case emphasizes the need to form substantive collaborations with institutions such as Howard University that are addressing questions related to the health of underrepresented populations. We are interested in forming collaborative relationships with data scientists to develop appropriate analytical algorithms for population inference.

CASE 2: ENCOURAGING PARTICIPATION OF VULNERABLE GROUPS IN DATA SCIENCE FOR ALGORITHMIC FAIRNESS

In this case study, we describe our experience working with criminal justice recidivism data to design a tutorial for the Broadening Participation in Data Mining workshop (BPDM). The tutorial on algorithmic fairness in the criminal justice system took place at BPDM 2019, a 3-day standalone workshop for 65 underrepresented gender, ethnicity and ability minorities from undergraduate through early career data scientists held at Howard University. This algorithmic fairness tutorial was first introduced by Dr. Falvio Calmon from Harvard University at BPDM2017 in Halifax, Nova Scotia. The 2017 co-location of BPDM with SIGKDD and the Fairness Workshop increased BPDM participant exposure to the topics of algorithmic fairness and data mining. Each tutorial has been preceded by a panel discussion on algorithmic fairness and the role of data scientist derived from vulnerable populations in recognizing the underlying biases inherent in large data sets such as the COMPAS dataset. The tutorial introduces the topic of algorithmic fairness, which attempts to identify and mitigate unfair bias against vulnerable groups in automated decision making procedures, and investigates in-depth the application of one such automated tool within the criminal justice system in the US.

We feel there are a number of benefits to focusing the hands-on tutorial for the BPDM workshop on this topic. Teaching tools that incorporate social good topics (in this case social justice, criminal justice, and algorithmic fairness) have been identified as having potential for broadening participation in computing (Buckley et al., 2008) where women and ethnic minorities have been woefully underrepresented. Students motivated by their interests and values, and engagement with non-traditional students can tap into this by demonstrating ways that computer science can have a positive social impact and "make a difference" (Goldweber et al., 2013). In addition to appealing to their interests, exposing students to the topic of algorithmic fairness can advance their research skills, exposing them to cutting-edge research practices for real world competency, and ethical application of data mining skills.

Furthermore, creating a more inclusive body of data analysts looking into this type of problem data can help ensure a diverse and inclusive critical perspective on the use of AI in society. Participants at the BPDM workshop are members of underrepresented groups in computing, representing ethnic, ability and gender minorities identified as vulnerable populations on both side of the data analysis pipeline. A key consideration of our tutorial development was to avoid putting the burden of addressing unfair structural biases onto the very members of the populations who are being made vulnerable. However, given the recent research interest in judicial fairness, our workshop provided the opportunity for trainees from vulnerable populations to use data as a means to both identify structural inequalities and to address those inequalities using algorithmic fairness nested within a social equity construct.

Interest in the impact of big data on society has been growing recently in the data mining and machine learning community, with input from legal scholars (Barocas and Selbst, 2016). Of particular concern is the use of algorithmic decision making procedures in regulated domains such as lending, housing and criminal justice. The study of algorithmic fairness seeks to address the fact that structural inequities which exist in our society can be encoded in subtle ways in the data we collect and analyze, allowing discriminatory practices to be perpetuated or even exacerbated by predictive models trained on historic data. Recent research in the AI community has focused on identifying bias against protected groups, as defined by sensitive data attributes such as age, gender, disability and ethnicity. Many tests have been proposed for assessing fair outcomes (Hardt et al., 2016; Chouldechova, 2017; Kleinberg et al., 2017). Identifying unfair treatment of these vulnerable populations is a paramount and challenging task, given the widespread use of

sophisticated, difficult to interpret, and often proprietary models for decision making.

Recent reporting has, in part, been fueled by a high profile expose (Angwin et al., 2016) published by ProPublica, a Pulitzer prize-winning investigative journalism organization. The article investigates risk assessment tools widely used in the US criminal justice system. These tools are algorithms developed by private companies and purchased by states to evaluate defendants. Judges are presented with risk assessment scores rating the dangerousness of defendants, which they can use in decisions such as setting bail or deciding sentencing. The authors show that one popular tool, called COMPAS, assessed African Americans and European Americans differently when assigning risk scores to be used at bail hearings. Their analysis showed that "black defendants were nearly twice as likely to be misclassified as higher risk compared to their white counterparts." In response, the company which developed the algorithm published a counter analysis, using a different statistical test to demonstrate fairness with respect to ethnic inference of a vulnerable population. Computer science researchers then picked up the investigation, publishing numerous results, including showing that the different standards used to determine fairness were impossible to satisfy concurrently Chouldechova (2017); Kleinberg et al. (2017).

The data released with this article has become the de facto benchmark for "fair" algorithms seeking to ensure equal treatment of different groups. The COMPAS dataset is unusual in that it contains real world data demonstrating a direct impact of algorithmic decision making on individuals. The data were available as part of public records, and include sensitive data attributes of race, gender, and age, as well as identifying information. Its popularity and availability have meant it has been used extensively by researchers in a very short time. Choosing this dataset for a hands-on tutorial session at the broadening participation workshop created an opportunity for discussion and reflection on the role of members of vulnerable populations as both data points and as data scientists. The tutorial presented a brief overview of the topic, introducing the concepts of protected groups defined by sensitive data attributes such as race and gender. In our workshop discussion we considered the problematic nature of such datasets and their increased role in decision making in our society, alongside other examples. We discuss subtle ways that data have historically been used to enforce discriminatory practices, for example in the redlining practices in which zip code was used as a proxy for race to enforce residential segregation in housing. Then we discuss ways that unfair bias can enter a modern data mining pipeline.

Typical data mining models train on data collected in the past, and then are used to make decisions about the future. If there are historical inequalities inherent in the training data, they will be perpetuated, and possibly even exacerbated by our predictive model. Skewed training data can lead to better accuracy for some groups vs. others. We discussed the example of gender stereotypes encoded in word embeddings used in natural language processing (Bolukbasi et al., 2016), and the example of facial recognition tools trained on majority white, male faces (Buolamwini and Gebru, 2018). These examples demonstrate cases where fairness research had a real world impact, as these papers have prompted companies to improve facial recognition software, and the development of bias mitigation techniques for text analysis. We discussed questions to consider when developing/applying new method, e.g., "Who will use this technology, and will it work equally well for everyone?" and "Is my dataset representative of all groups?"

The learning objectives of the tutorial are to examine some examples of structural inequality in society that is buttressed by data mining practices including developing ways to recognize ways in which unfair bias might be introduced into a data mining pipeline. Because vulnerable populations are often placed in the position of being whistleblowers for structural inequalities, we discussed how to perform analyses to verify whether a predictive model is fair or unfair and what outcomes should be considered when developing data mining techniques beyond accuracy. To address these concerns we have to develop tools to democratize the development of data mining techniques and technologies using open and transparent methods with clearly reproducible findings. This tutorial demonstrates one approach to doing this [i.e., with interactive Jupyter notebooks (Kluyver et al., 2016)] and give students hands-on experience with open software tools.

The Algorithmic Fairness for Vulnerable Populations tutorial steps through a typical data analysis pipeline. First the data is cleaned and preprocessed according to the steps taken in the ProPublica analysis. Then a number of statistical and visualization methods are applied to allow participants to assess the attributes in the training dataset and understand whether there is any unfair bias present. Finally, three notions of group fairness are introduced, covering state-of the-art bias detection metrics from the recent literature:

- Disparate Impact. This legal concept is used to describe situations when an entity such as an employer inadvertently discriminates against a certain protected group. This is distinct from disparate treatment where discrimination is intentional. To demonstrate cases of disparate impact, the Equal Opportunity Commission (EEOC) proposed "rule of thumb" is known as the 80% rule.
- Calibration. This statistical test was used to verify the fairness of the COMPAS model by the company Northpoint that created the tool. The basic idea behind calibrating a classifier is to have the confidence of the predictor reflect the true outcomes. In a well-calibrated classifier, if 100 people are assigned 90% confidence of being in the positive class, then in reality, 90 of them should actually have had a positive label. To use calibration as a fairness metric we compare the calibration of the classifier for each group.
- Equalized Odds. The last fairness metric we consider is based on the difference in error rates between groups. The equalized odds criterion (Hardt et al., 2016) proposes to look at the difference in the true positive and false positive rates for each group. This aligns with the analysis performed by ProPublica.

The goal of this tutorial's implementation was to allow for hands-on analysis right away, without requiring any heavy overhead from installing many tools or having to clean and pre-process the data. At the same time, all analysis was fully transparent and available for experimentation. Participants could step through

the notebook and simply follow along, or dig deeper and edit the code directly to experiment with the data. Suggestions for possible further experimentation are provided throughout. Links to datasets, research papers, Wikipedia entries, and Python data mining tools provide context and avenues for deeper investigation into the topics and methods described. A clear outcomes was that trainees who undertook the data manipulation and assessment felt empowered to identify the limitations of data resulting from structural inequalities and to identify mechanisms to address those biases in data.

CASE 3: INVESTIGATING VULNERABLE POPULATION SPECIFIC VARIATION USING GENOME EDITING TOOLS

Indigenous communities represent a classic example of a vulnerable population for whom territorial rights, educational attainment and health status are all under stress. Nevertheless, they remain a subject of keen genomic interest to western scientists. Unfortunately, these largely one-sided cross cultural scientific interactions between Indigenous populations and European ancestried scientists have long been steeped in misunderstanding and mistrust. Cases like the Havasupai Nation's inclusion in stigmatizing mental health research against their will have helped to drive many Indigenous peoples to reassess their willingness to work with non-Indigenous scientists (Garrison, 2013). The development of novel large scale data generation tools have emphasized the voluntary exclusion of Indigenous populations and the paucity of data upon which to gain meaningful insights on Indigenous communities' health and well-being.

The utility of data analysis has been readily adopted by human geneticists, who have willingly accepted the tools of big data to better understand the features of the genome including variable sites across the genome, chromosomal arrangements, and population level variation.

This pursuit of ever increasing data has lead to breakthroughs in ancestry assessments, multi-omic precision medicine models and has spurred molecular breakthroughs like the Crispr-Cas9 system of gene editing. Crispr-Cas9, most recently made infamous by the ethically condemned modification of Chinese twins (Schmitz, 2019).

While genome sequencing is a great tool for identifying genetic variation that might be involved in disease mechanisms, correlation does not equal causality. Gene editing tools offer the population geneticists the opportunity to identify population-specific variation derived from large scale sequencing experiments and to conduct further assessment of the functional significance of genome sequence variation, thus potentially identifying the changeable sites underlying traits or disorders. For example, gene editing technologies can be used to investigate population-specific, positively selected point mutations implicated in a range of diseases (Komor et al., 2016). In addition to using these tools that are already in existence to functionally investigate individual variants in clonal cell lines, multiple laboratories have begun to develop new editing tools to simultaneously introduce multiple mutations in the human genome via multiplex nucleotide editing of population specific haplotypes under selection, or multiple point mutations on different chromosomes in human genome.

Engineering new tools to functionally investigate single nucleotide changes is an exciting prospect for two primary reasons: (A) Creating accountability. Culturally competent empirical evidence and detailed theoretical considerations should be used for evolutionary explanations of phenotypic variation observed in humans (especially Indigenous populations). Population genetics investigators frequently overlook the importance of these ethnographic criteria when associating observed trait variation with evolutionary analysis. Functional investigation of population specific variation has the potential to empower the population genetics community by holding evolutionary explanations accountable (Gould and Lewontin, 1979). This need for mechanistic insight is framed by problematic narratives and exacerbated by correlation based studies that fail to properly functionally investigate single nucleotide changes. Because Indigenous populations are vulnerable (i.e., at risk populations), it is the genomic technology development community's responsibility to take these potentially problematic narratives to task (Neel, 1962). Not to just reclaim Indigenous history through the population genetics projects we champion, but potentially empower Indigenous history with genome editing tools. (B) Democratizing tools. Indigenous peoples are under-represented in both population-based genomic studies, and as primary investigators in academia. For Indigenous researchers, this leads to questions as to how Indigenous peoples will meaningfully participate in human population genetics, and how to address the disparities currently existing in Indigenous communities? One way that Indigenous scientists are addressing this is the formation of an educational consortium that is focused on educating Indigenous genetics, such as the Summer Internship for Indigenous Peoples in Genomics (SING Consortium). This research consortium works with Indigenous communities to generate large scale data to address the genomic and health disparity questions that those communities have Claw et al. (2018).

In addition to standard metrics of academic success such as grant awards and paper publications, Indigenous researchers must transition our research focus to understanding how independent research programs will become actionable. If participating Indigenous communities are not presented with tangible benefits to collaborating with non-Indigenous scientists, such as access to medicine, developments to infrastructure, or capacity building, then research focusing on Indigenous communities could potentially continue a legacy of colonial exploitation. Technological independence, self-governance, and democratization of the tools should always be the long-term goal of ethical partnerships in genomic sample collection, large scale data analysis and inference generation. Some easy solutions to address these concerns include engaging Indigenous communities in educational seminars within Indigenous spaces including Native American Reservations, Hawaiian Heiau, and

Maori Marae. Another priority must be to transition genomic research toward focusing on the development of biomedical tools to make gene editing of deleterious genomic changes more affordable, empowering Indigenous populations across the globe to gain agency over their own future.

DISCUSSION

Each of these case studies demonstrates how vulnerable ethnic and justice status individuals can be involved, not just as the objects of proposed studies of vulnerable populations but in the study design, implementation, and importantly the analysis and inferential assessment of results. In each case, including vulnerable populations can yields better more inclusive results with populations becoming invested in the outcomes of evidence-based analysis. Among the lessons derived from these cases are that partnering with institutions that serve vulnerable populations is crucial to the collection of bias free data. This collaboration must include clear benefit for vulnerable communities. Another lesson learned is that where data on vulnerable populations exists, partnering with data scientists derived from those vulnerable populations can help to disentangle an algorithm's inferential ability from a manifesting of implicit bias in data collection. Finally overcoming generational reluctance to participate in research on underserved populations requires both educational trust building and dialogue with a collaborative spirit. Data science must include vulnerable populations in the research design, analysis and inference of data findings in order to make interpretations that are valuable and meaningful to those populations. Whether focused on social science, biomedical applications or preventing the harvesting of large scale genomic data from vulnerable populations with no clear reciprocal benefit to them, the inclusion of these diverse population and perspectives can improve data science. In addition, the continuing need for broad educational access and enhanced ability to make sense of the increasing complexity of big data requires that more vulnerable community perspectives be included.

While we focus on the role that vulnerable populations can play in addressing the information, health and social justice disparities in their communities, it is equally important to identify the role that intersectionality plays in the lived identities of vulnerable populations. We believe that this is an area that needs to be further addressed in the data science research literature. Taken together, these case studies present illustrative examples of how vulnerable populations, researchers, and the institutions that serve them can contribute to improving data science by their participation, not just as study subjects, but as robust intellectual research partners.

DATA AVAILABILITY

The COMPAS tutorial datasets analyzed for this manuscript can be found in the GitHub repository https://github.com/caitlinkuhlman/bpdmtutorial.

ETHICS STATEMENT

The African American Genome Projects was carried out in accordance with the recommendations of the Human subjects guidelines of the Howard University Institutional Review Board with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Howard University Institutional Review Board.

AUTHOR CONTRIBUTIONS

LJ conceptualized the collaborative manuscript theme. LJ, PF, FJ, and CK each wrote sections of this manuscript. LJ and CK edited the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

1000GenomesConsortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Alpers, E. A. (1997). The african diaspora in the northwestern indian ocean: reconsideration of an old problem, new directions for research. *Comp. Stud. South Asia Africa Middle East* 17, 62–81. doi: 10.1215/1089201X-17-2-62

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). *Machine Bias*. New York, NY: Pro Publica.

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. *Cal. Law Rev.* 104:671. doi: 10.2139/ssrn.2477899

Bolukbasi, T., Chang, K-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *arXivpreprint* arXiv:1607.06520.

Buckley, M., Nordlinger, J., and Subramanian, D. (2008). "Socially relevant computing," in *ACM SIGCSE Bulletin - SIGCSE 08,* Vol. 40 (New York, NY: ACM), 347–351. doi: 10.1145/1352322.1352255

Buolamwini, J., and Gebru, T. (2018). Accountability, and transparency gender shades: intersectional accuracy disparities in commercial gender classification. *Proc. Mach. Learn. Res*. 81, 1–15.

Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5, 153–163. doi: 10.1089/big.2016.0047

Claw, K. G., Anderson, M. Z., Begay, R. L., Tsosie, K. S., Fox, K., Nanibaa, N. A., et al. (2018). A framework for enhancing ethical genomic research with indigenous communities. *Nat. Commun.* 9:2957. doi: 10.1038/s41467-018-05188-3

Cooper, F. (1977). *Plantation Slavery on the East Coast of Africa*. New Haven: Yale University Press.

Ewald, J. J. (2000). Crossers of the sea: slaves, freedmen, and other migrants in the northwestern indian ocean, c. 1750–1914. *Am. Hist. Rev.* 105, 69–91. doi: 10.2307/2652435

Garrison, N. (2013). Genomic justice for native americans: impact of the havasupai case on genetic research. *Sci. Technol. Human Values* 38, 201–223. doi: 10.1177/0162243912470009

Goldweber, M., Barr, J., Clear, T., Davoli, R., Mann, S., Patitsas, E., et al. (2013). A framework for enhancing the social good in computing education: a values approach. *ACM Inroads* 4, 58–79. doi: 10.1145/2432596.2432616

Gould, S. J., and Lewontin, R. C. (1979). The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 205, 581–598. doi: 10.1098/rspb.1979.0086

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Adv. Neural Info. Process. Syst.* 2016, 3315–3323.

Harris, J. (1971). *The African Presence in Asia: Consequences of the East African Slave Trade*. Evanston, IL: Northwestern University Press.

Harris, J. E. (2003). Expanding the scope of african diaspora studies: the middle east and india, a research agenda. *Rad. History Rev.* 87, 157–168. doi: 10.1215/01636545-2003-87-157

Jackson, F. (1996). Concerns and priorities in genetic studies: insights from recent african american biohistory. *Seton Hall Law Rev.* 27, 951–970.

Jackson, F. (1997). "Assessing the human genome project: an african american and bioanthropological critique," in *Plain Talk about the Human Genome Project* (Cambridge, MA), 95–104.

Jackson, F. (1998). Scientific limitations and ethical ramifications of a non-representative human genome project: African american response. *Sci. Eng. Ethics* 4, 155–170. doi: 10.1007/s11948-998-0046-6

Jackson, F. (2008). Ethnogenetic layering (el): an alternative to the traditional race model in human variation and health disparity studies. *Ann. Human Biol.* 35, 121–144. doi: 10.1080/03014460801941752

Jackson, F. (2018). Genomic testing among african americans—problems, limitations, and solutions. *Gene Watch* 2018, 1–31. Available online at: http://www.councilforresponsiblegenetics.org/GeneWatch/GeneWatchPage.aspx?pageId=589

Johnson, J., and Jackson, F. (2015). Use of multiple intelligence modalities to convey genetic and genomic concepts in african american college biology students. *Nat. Sci.* 7:299. doi: 10.4236/ns.2015.76033

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2017). "Inherent trade-offs in the fair determination of risk scores," in *Proceedings of the 8th Conference on Innovation in Theoretical Computer Science* (Ithaca, NY).

Kluyver, T., Ragan-Kelley, B., P'erez, F., Granger, B. E., Bussonnier, M., Frederic, J., et al. (2016). "Jupyter notebooks-a publishing format for reproducible computational workflows," *in Positioning and Power in Academic Publishing: Players, Agents and Agendas*, eds L. Fernando and S. Birgit (Southampton: IOS Press), 87–90.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., and Liu, D. R. (2016). Programmable editing of a target base in genomic dna without double-stranded dna cleavage. *Nature* 533:420. doi: 10.1038/nature17946

Lu, J. B., Danko, K. J., Elfassy, M. D., Welch, V., Grimshaw, J. M., and Ivers, N. M. (2018). Do quality improvement initiatives for diabetes care address social inequities? Secondary analysis of a systematic review. *BMJ Open* 8:e018826. doi: 10.1136/bmjopen-2017-018826

Mann, K. (2001). Shifting paradigms in the study of the African diaspora and of Atlantic history and culture. *Slavery Abol.* 22, 3–21. doi: 10.1080/714005179

Murdoch, T. B., and Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA* 309, 1351–1352. doi: 10.1001/jama.2013.393

Neel, J. V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am. J. Human Genet.* 14:353.

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nat. News* 538:161. doi: 10.1038/538161a

Schmitz, R. (2019). *Gene-editing Scientist's Actions are a Product of Modern China*. Available online at: https://www.npr.org/2019/02/05/690828991/gene-editing-scientists-actions-are-a-product-of-modern-china (accessed April 01, 2019).

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell* 177, 26–31. doi: 10.1016/j.cell.2019.02.048

Tangcharoensathien, V., Mills, A., Das, M. B., Patcharanarumol, W., Buntan, M., and Johns, J. (2018). Addressing the health of vulnerable populations: social inclusion and universal health coverage. *J. Global Health* 8:20304. doi: 10.7189/jogh.08.020304

Zeleza, P. T. (2005). Rewriting the African diaspora: beyond the black atlantic. *Afr. Affairs* 104, 35–68. doi: 10.1093/afraf/adi001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 08 August 2019
doi: 10.3389/fdata.2019.00018

Global Awareness Landscape for Ailments—A Twitter Based Microscopic View Into Thought Processes of People

Durga Toshniwal[1], *Soumya Somani*[2]*, *Rohit Aggarwal*[3] and *Preeti Malik*[1]

[1] Department of Computer Science, Indian Institute of Technology Roorkee, Roorkee, India, [2] Department of Computer Science, Symbiosis Institute of Technology, Symbiosis International University, Pune, India, [3] Department of Computer Science, Indian Institute of Information Technology, Allahabad, India

In this day and age, people face a lot of stress due to the fast pace of life. Due to this, people in today's digital age, suffer from a plethora of ailments. It is universally accepted that a greater awareness of ailments and their corresponding symptoms leads to an increased lifespan and better quality of life. Early detection and screening can help doctors nip diseases in their natal stages. However, not everyone is aware of them, which makes it a global issue. The study of the degree of disease awareness amongst people belonging to different nations and continents is a matter of great interest. One method that is suitable for this purpose is using clinical data. But, this data is not readily available. However, today a plethora of platforms are available to people to share their thoughts and experiences. People post about many of the important events in their lives on social media. Their posts offer a microscopic view into their lives and thought processes. Based on this intuition, twitter data pertaining to various chronic and acute diseases has been collected. Tweets for 30 deadly ailments have been collected over a period of 3 months amounting to a total of 19 million. A feature extraction approach is proposed which is used to identify the disease awareness levels across different nations. Deriving the global awareness landscape for ailments can help to identify regions which are well aware and also those that need to get aware. Clustering has been used for this purpose.

Keywords: data mining, world health, social media, epidemics, ailments, twitter data analytics

OPEN ACCESS

Edited by:
Yelena Mejova,
Institute for Scientific Interchange, Italy

Reviewed by:
Daniela Paolotti,
Institute for Scientific Interchange, Italy
Kyriaki Kalimeri,
Institute for Scientific Interchange, Italy
Rumi Chunara,
New York University, United States

***Correspondence:**
Soumya Somani
smaheshwari1234@gmail.com

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *25 March 2019*
Accepted: *03 June 2019*
Published: *08 August 2019*

Citation:
Toshniwal D, Somani S, Aggarwal R and Malik P (2019) Global Awareness Landscape for Ailments—A Twitter Based Microscopic View Into Thought Processes of People. Front. Big Data 2:18. doi: 10.3389/fdata.2019.00018

INTRODUCTION

With the success of Web-2.0, it has become a quotidian task for web users to express their views on a myriad of issues. Web 2.0 has given an opportunity to its users so that they can interact and collaborate to create texts of their thinking and understanding on this virtual platform. This has given birth to many web applications, social-networking sites, video sharing sites, blogs, hosted services, and wikis among other things.

Consequently, many social platforms are available to people for sharing their thoughts on a variety of topics, events and products. Most of these posts chronicle their daily activities and struggles. People post about all the relevant and irrelevant events in their lives. Not all of these are useful but many of them can be used to gain an insight into society. These can be collected and the useful information can be selected by applying multiple data analytic and mining techniques.

The field of world health can vastly benefit from analyzing this data. A number of people share their health struggles and their opinions on health concerns around them on social media. Many compulsively post regular updates on the diseases that they themselves or their close relatives suffer from. People also express their concern about the diseases that are currently widespread in their localities on their social media. An analysis of these posts can be very helpful in finding a disease's spreading pattern or at the very least help us in determining disease awareness patterns among the citizens of various countries. This information can be used as a preparatory measure by the government and citizens of various countries.

Twitter has become a popular source of data in the last decade. The posts are brief, and therefore, they effectively convey a person's opinion in just a few words making it useful for the purpose of research. Twitter is also very convenient for all internet users and since the internet is ubiquitous in today's day and age, it can be called the virtual realization of all thoughts prevalent in today's society at a given point in time. Many researchers have already established Twitter as a useful source of information or data while working on many topics including public health (Sriram et al., 2010).

In this paper, ailments have been classified into chronic and acute ailments (which are to be identified differently) thereby forming two sets of the problem. Each disease also needs to be worked upon individually as all of them are different from each other in one way or another.

The rest of the paper is organized as follows: Section Literature Review gives the literature review of related topics. Section Proposed Work gives the details of the proposed work for the paper. Section Experiments and Discussion comprises of the dataset details, experiments conducted on them and the results from them. Section Conclusion and Future Work is the conclusion and future work of the report. And the references are given in the last section.

LITERATURE REVIEW

In this section of the paper, some of the related research works in this field have been described briefly.

Not much work has been done to analyze twitter data for the purpose of determining awareness levels of diseases in various countries across the globe. So far research works have focused on some particular ailments or on the observations from a specific country. There is a need to perform a study that spans across a large set of common ailments in order to generate a complete picture of the awareness levels of various diseases in different countries around the world.

Research by Paul and Dredze (2011) gives an analysis of the health issues that can be studied using data from Twitter. This work focuses on the tweets collected from the United States. Results from this work show that most of the ailments that are studied can be predicted with accuracy in terms of the location of the patient, except for the deadly ones in which the patients' relatives and not the patients themselves might be tweeting. So, the tweets may be from a location which is different from the patients' location thereby reducing the accuracy.

There are a few studies regarding the occurrence of influenza in the United States during different years, pertaining to different kinds of work in the fields of disease pattern, location pattern etc. Influenza occurs in all the seasons with different intensity and different regions making it an interesting subject.

One approach given by Signorini et al. (2011) has combined the analysis of the occurrence of H1N1 and influenza on a weekly cross-validated dataset. The results of the prediction were cross-checked with the actual statistics of occurrence of the two diseases with an average error of 0.28% and standard deviation of 0.23%.

Another approach given by Aramaki et al. (2011) focuses on separating negative tweets that show the person not having influenza, from positive tweets which actually indicate influenza occurrence. Results show that it could successfully filter out negative tweets with f-measure $= 0.76$ and it detects influenza with a high correlation ratio of 0.89.

Yet another approach given by Smith et al. (2016) using Twitter data from the influenza epidemic of 2012-2013 in the United States, majorly works on distinguishing between the tweets that show awareness toward the disease and the tweets that actually show an ailment. Results from the model show that occurrence of disease has very different trends than that of its awareness. It has also shown that disease trends vary on a regional basis but awareness trends do not vary as much across different regions. Similarly, some other diseases like Dengue, HIV, H1N1, Zika etc. are also discussed using twitter data.

Based on the detailed literature survey done, it can be observed that most of the existing research works are based on the analysis of social media data from specific locations or sets of locations. Thus, there is a great scope to develop techiniques that work on data collected on a national or global level.

Also there is no centralized data on occurance patterns of diseases on a national or global level collected by the governments or other agencies.

Further, most of the research works done so far are targeted toward the analysis of a set of few ailments only. However, there is a lack of research work that holistically covers a broad spectrum of ailments.

In the present work, we aim to address some of the limitations as mentioned above.

PROPOSED WORK

In order to determine the awareness levels about various ailments in various countries across the globe, the following framework (as per **Figure 1**) has been proposed.

Proposed Framework

To find the awareness levels using any kind of epidemiological data is not feasible as the data is not readily available due to privacy issues and the data that is available is insufficient to cover the entire globe. Thus, an approach to determine a worldwide analysis of public health awareness amongst people using twitter data has been given here.

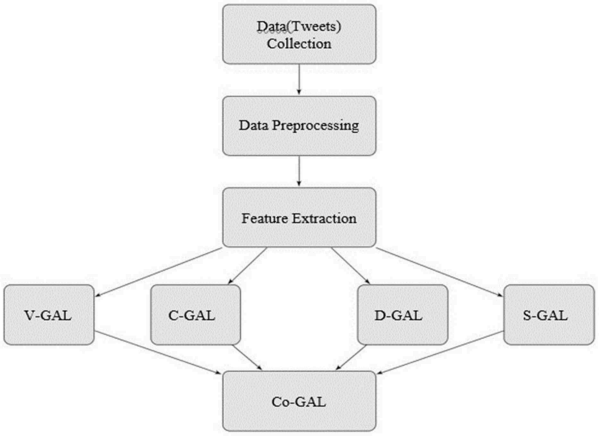

FIGURE 1 | Proposed Framework, V-GAL is Visual Global Awareness Landscape, C-GAL is Continent based Global Awareness Landscape, D-GAL Disease based Global Awareness Landscape, S-GAL Similarity based Global Awareness Landscape, Co-GAL Consolidated Global Awareness Landscape.

The proposed framework for the work is illustrated in **Figure 1**.

Data Collection

The data was collected from twitter live stream using the twitter API over a period of 3 months. Twitter live stream allows us to connect to the twitter API and open a pipeline for selected data to be delivered to us. A total of 30 ailments were chosen based on the level of severity and spread. Only those ailments which were being discussed on twitter were considered. Different keywords regarding each ailment were used to collect this data from the stream. The Tweepy library was used to access the twitter API.

Data Biases

Only English language tweets were collected. This was done to avoid transation of non english tweets since such translation will yeild noisy data. Therefore, this work does not include any expression done by people on non english languages. There was no thresholding applied to the volume of tweets from a nation. Also, various nations of the world will have largely varying population and hence bigger countries will have more tweets. This will introduce a bias toward such countries in awareness levels. To prevent this, normalization of the number of tweets from a country with respect to its population has been done.

Data Preprocessing

Not all of the collected tweets have the location attribute in them. The location tagged tweets are thus separated out for further analysis. This is achieved by filtering out the tweets that had null or garbage values as their location values. The tweets are processed using Google Geo-coding API to determine the country from where the tweets are posted. The corpus of the tweets is then segmented based on their country. It is also segregated into tweets about chronic and acute ailments based on their keywords and noisy tweets i.e., tweets containing non-english words, very few words etc. are filtered out.

Feature Extraction

Feature Vectors are derived to give clusters of countries with similar awareness.

The feature vectors are derived as follows:

Let C be the set of countries given as per Equation 1:

$$C = \{C_1, \ldots, C_i, \ldots, C_n\} \tag{1}$$

Further, let the set of chronic ailments be denoted by A_{chj}:

$$A_{chj} = \{A_1, \ldots, A_j, \ldots, A_c\} \tag{2}$$

And let the set of actue ailments be denoted by A_{aj}:

$$A_{aj} = \{A_{c+1}, \ldots, A_j, \ldots, A_m\} \tag{3}$$

Let the set of all ailments be denoted by A:

$$A = A_{chj} \mathbf{U} A_{aj}$$

Thus,

$$A = \{A_1, \ldots, A_j, \ldots, A_m\} \tag{4}$$

The corpus of tweets, T is given as per Equation 5:

$$T = \sum_{i=1}^{m} T_i \tag{5}$$

Where T_i is the total number of tweets from country C_i given as per Equation 6:

$$T_i = \sum_{j=1}^{m} T_{ij} \tag{6}$$

Where,

T_{ij} = The number of tweets from country i about ailment j.

Let the population of country C_i be denoted by P_i and P be the world population[1]. Then:

$$P = \sum_{i=1}^{n} P_i \tag{7}$$

The tweets have been segregated based on location coordinates as discussed in the Data Preprocessing section.

The proposed approach for awareness level indication using feature vectors is

Let the Feature Vector for a country C_i be denoted by FV_i as per Equation 8:

$$FV_i = (A'_1, \ldots, A'_j, \ldots, A'_m) \tag{8}$$

Where,

$$A'_j = (T_{ij} / T_i)$$

[1]https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)

Feature vector of a country C_i for chronic diseases is called the Chronic Ferature Vector and is given as per Equation 9:

$$CFV_i = (A'_1, \ldots, A'_j, \ldots, A'_c) \quad (9)$$

Similarly, the Acute Feature Vector for a country C_i is given as per Equation 10:

$$AFV_i = (A'_{c+1}, \ldots, A'_j, \ldots, A'_m) \quad (10)$$

Thus,

$$FV_i = AFV_i \, \mathbf{U} \, CFV_i \quad (11)$$

After the Feature Vectors are derived, Link Based, and Agglomerative Clustering methods are applied to get clusters of countries with similar awareness.

The aim of clustering is as follows:

- Given an ailment, the aim is to determine a group of countries showing similar awareness levels for it.
- Given a country, the objective is to find the top ailments being discussed.
- And lastly, we need to determine the countries that have similar top scoring ailments.

Visual Global Awareness Landscape (VGAL)

A Tweet Index has been defined to create the Visual Global Awareness Landscape (VGAL). It gives the level of awareness about various diseases for every country based on its normalized population. It is defined as per Equation 12.

$$\text{Tweet Index} = ((T_i/P_i)x(P/T)) \quad (12)$$

Continent Based Global Awareness Landscape (CGAL)

In this landscape, a discussion has been given regarding the diseases that people are most aware of in each continent. Acute and Chronic diseases have been discussed separately. So, the top scoring acute and chronic diseases for each continent have been determined in this landscape.

Disease Based Global Awareness Landscape (DGAL)

A disease based discussion has been presented regarding the countries that have the most awareness about each disease. Also, the top scoring diseases being discussed in each country are compared to the most prevalent ailment in that country. Acute and Chronic diseases have been considered separately for this purpose.

Similarity Based Global Awareness Landscape (SGAL)

In this landscape, clustering algorithms have been applied on CFV and AFV sets to determine similarity based groups of countries. Clusters of countries are formed such that within a cluster, similar awareness levels exist for a common set of diseases. Two methods of clustering which are inspired by Guha et al. (2000) and Kaufman and Rousseeuw (1990) have been applied to the CFV and AFV sets. The methods are: Link Based and Agglomerative Clustering.

TABLE 1 | Dataset description.

Sr. no.	Attributes	Values
1	No. of chronic diseases	10
2	No. of acute diseases	20
3	Duration	July-Sept 2017
4	Total tweets collected	19,301,623
5	Total countries	244

Consolidated Global Awareness Landscape (Co-GAL)

The CVF and AVF sets have further been analyzed to give the Consolidated Global Awareness Landscape which comprises of:

- Holistic Awareness Profile (HAP): This consists of countries that have awareness about all the ailments considered. Such countries are not in immediate need of awareness campaigns for diseases. The can also mentor other countries to help them in becoming more aware against diseases.
- Specific Awareness Profile (SAP): It consists of countries that have awareness about some specific ailments.
- Negligible Awareness Profile (NAP): This consists of countries that have the least awareness. These countires are in immediate need of awareness campaigns against various diseases.

Lastly, geographical aspects have been considered to deternmine the geographical closeness of countries lying in the same cluter. Also, the actual occurance of ailments has been considered to determine the correlation between the occurance and awareness levels of ailments.

EXPERIMENTS AND DISCUSSION

The following section contains the dataset description that gives us the total number of ailments considered in this work along with the number of acute and chronic diseases. This section also presents the results obtained in this work.

Dataset Description

The data was collected from twitter live stream using twitter API over a period of 3 months. As per **Tables 1**, **2**, 30 ailments in total were chosen based on the level of severity and spread.

Ailments have been classified into two categories: Chronic and Acute. An ailment that develops over a longer period of time and lasts for more than a period of 3 months is known as a chronic ailment and an ailment that comes rapidly and lasts for a short period of time is categorized as an acute ailment.

Discussion

After the data has been processed and the various steps specified in the proposed framework have been carried out, the following results have been obtained. These results give us a holistic picture of the global awareness levels of various ailments.

TABLE 2 | Ailment description.

Sr. no.	Type of ailment	No. of ailments	Namely
1	Chronic	10	Cancer, chikungunya, diabetes, heart diseases, hepatitis, HIV, leprosy, RHD, TB and toxoplasmosis
2	Acute	20	Chickenpox, cholera, dengue, diarrhea, ebola, H1N1, influenza, Japanese encephalitis, lassa fever, malaria, measles, mumps, pertussis, rift valley fever, Smallpox, syphilis, typhoid, typhus, yellow fever, zika

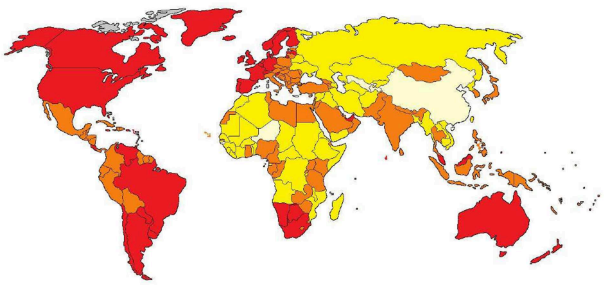

FIGURE 2 | Visual Global Awareness Landscape.

Visual Global Awareness Lanscape (VGAL)

Figure 2 shows the awareness levels of each country based on the normalized tweets per person (given by the Tweet Index). The most aware countries are represented in red and the least aware are represented in light yellow. The awareness for each color is:

Red = 1; 0.2 ≤ orange < 1; 0.1 ≤ yellow < 0.2 and 0 ≤ light yellow < 0.1

Red denotes high awareness, orange denotes medium awareness, yellow denotes low awareness, and light yellow represents the least aware countries.

Continent Based Global Awareness Lanscape (CGAL)

Based on location of generation of the tweets, they can be divided amongst the seven continents. The statistics for each of the continents can be seen in **Table 3**. The % Column depicts the percentage of tweets from the continent with respect to the world. The top chronic column gives the top scoring chronic ailment for that particular continent. Similarly, the top acute column gives the top scoring acute ailment for that continent.

Out of the chronic diseases, cancer is prevalent in all of the continents except for South America. Tuberculosis (TB) is the most prevalent chronic ailment in South America. This can be explained by the fact that Brazil has a high occurrence of TB and most of the tweets from South America (around 73%) are from Brazil.

Out of the set of the acute ailments, Cholera, Dengue and Zika have the most awareness in various continents (refer **Table 3**).

Disease Based Global Awareness Lanscape (DGAL)

All the ailments have been classified into Acute and Chronic ailmets. The top scoring ailments from each category are given in **Tables 4**, **5**. The % column in the **Tables 4**, **5** give the percentage of tweets for each disease with respect to the total number of tweets from the world. Top scoring countries column gives the countries that have the highest number of tweets for a given ailment. The most prevalent ailment as per **Tables 4**, **5** signify the most commonly occuring ailment in that specific country.

TABLE 3 | Continent based statistics based on twitter data collected.

Sr. no.	Continent	% of tweets	Top chronic	Top acute
1	North America	51.1	Cancer	Zika
2	Europe	16.52	Cancer	Cholera
3	South America	14.88	TB	Cholera
4	Asia	11.13	Cancer	Cholera
5	Africa	4.73	Cancer	Cholera
6	Oceania	1.61	Cancer	Cholera
7	Antarctica	0.01	Cancer	Dengue, cholera

TABLE 4 | Top chronic ailments and the countries that majorly discuss about them.

Ailment	% of tweets	Top scoring countries (%)		Most prevalent ailment
Cancer	45	Unites States	60	Cancer (for all)
		United Kingdom	9	
		France	3.56	
		India	2.95	
		Canada	2.79	
TB	26	Brazil	40	TB for Brazil, Spain, Portugal and cancer for UK and US
		Unites States	22.5	
		Spain	4.36	
		Portugal	4.22	
		United Kingdom	3.25	
HIV	11.7	Unites States	45.6	Cancer (for all)
		United Kingdom	9.62	
		South Africa	5.89	
		Nigeria	3.78	
		India	3.41	
Diabetes	10.7	United States	47	Cancer for UK, US, India, TB for Brazil and diabetes for Indonesia
		United Kingdom	9.35	
		Brazil	5.96	
		Indonesia	3.78	
		India	3.5	

Top Chronic Ailments

Cancer has the highest % of tweets among all the chronic ailments, making it the most talked about disease all over the world. Other top scoring chronic ailments are TB, HIV and Diabetes.

Table 4 gives the top chronic ailments along with the top scoring countries for each ailment. For example, Brazil, Spain

TABLE 5 | Top Acute Ailments and the countries that majorly discuss about them.

Ailment	% of tweets	Top scoring countries (%)		Most prevalent ailment
Dengue	1.85	India	28	H1N1 in India, dengue in Pakistan and Mexico, cholera in US and zika in Brazil.
		Pakistan	14.6	
		United States	13	
		Brazil	6.64	
		Mexico	5.21	
Zika	1.77	United States	40	Dengue in US and Mexico, zika in Brazil and Venezuela, H1N1 in India
		Brazil	26	
		India	4.36	
		Mexico	3.66	
		Venezuela	2.95	
Cholera	1.66	United States	31.6	Cholera for all
		Kenya	12.16	
		United Kingdom	11.18	
		Poland	6.62	
		Nigeria	3.91	
Measles	0.8	United States	40	Cholera in US, UK and Australia, measels in Indonesia and H1N1 in India
		Indonesia	16.4	
		United Kingdom	9.85	
		India	6.08	
		Australia	3.28	

and Portugal have the maximum number of tweets about TB making them areas of high concern of TB. The most prevalent chronic ailments in the top scoring countries have also been given in **Table 4**.

Top Acute Ailments

Table 5 gives the top scoring acute ailments along with the top scoring countries for each ailment and the most prevalent acute ailments in those countries.

Out of all the countries discussing about dengue, only Pakistan and Mexico have it as the most prevalent acute disease.

However, all of the five countries most concerned about cholera have it as their most prevalent acute ailment.

Similarity Based Global Awareness Lanscape (SGAL)

To determine a similarity based global awareness landscape, clustering has been done on the set of country wise Feature Vectors, FV. Acute and chronic ailments have been considered separately for this landscape.

A total of 22 clusters of countries having similar awareness levels for chronic diseases have been generated. The major results have been presented in **Table 6**. It gives the size of the cluster, some of the important countries in that cluster and the similarity traits for that cluster.

TABLE 6 | Clusters of countries for chronic ailments.

Cluster No.	No. of countries	Countries	Ailment(s)
1	56	Algeria, Fiji, Greece, Canada, Australia, United Kingdom, United States, Nigeria, India etc.	Cancer 50–70%
2	47	Afghanistan, Kazakhstan, Japan, Hong Kong, China, Tajikistan, Hungary, Romania etc.	Cancer and TB 35–40%
3	19	Switzerland, Denmark, Nepal, Bangladesh, Sri Lanka, Vietnam etc.	Cancer, TB, HIV and diabetes 10–40%

TABLE 7 | Clusters of countries for acute ailments.

Cluster No.	No. of countries	Countries	Ailment(s)
1	8	Albania, Mauritiana, North Korea, Montenegro, Norway, Tunisia etc.	Cholera, ebola
2	7	Algeria, Turket, Samoa, Vanuatu, India, Myanmar (Burma) etc.	Dengue, cholera
3	5	Afgahnistan, Djibouti, Pitcairn Islands, Cook Islands and Equitorial Guinea	Cholera

For acute ailments, 38 clusters have been generated. They have sizes ranging from 2 to 12. The clusters consist of groups of countries that have similar levels of awareness for acute ailments. The most important results have been presented in **Table 7**. For example, Albania, Mauritiana, North Korea etc., have similar awareness for cholera and ebola.

This landscape gives similarity based groups of countries, which are clusters of countries that have similar awareness levels about similar diseases.

When the feature vectors for ailments are compared, a few of them show very similar awareness spread. Syphilis, Pertussis and Small Pox have a similar spread in terms of the number of tweets. Similarly, the awareness spread for HIV and Ebola, Chickenpox and Mumps are quite similar.

Consolidated Global Awareness Lanscape (Co-GAL)

Highly aware countries are countries that have awareness about all the considered ailments i.e. countries having citizens tweeting about all the considered ailments. Only seven countries, namely Australia, Canada, France, India, Thailand, UK and US, are highly aware countries these can be classified into HAP. Countries like Argentina, Brazil, Nigeria etc lack in awareness of some ailments despite having a large number of total tweets. These are classified into SAP. Such countries must not be mistaken for highly aware countries since they lack in awareness about some of the considered ailments.

TABLE 8 | Incidence and awareness comparision.

	Countries having awareness (HA)	**Countries not having awareness (NA)**
Countries having incidence (HI)	Countries having incidence and awareness of ailments (HIHA)	Countries having incidence but no awareness about ailments (HINA)
Countries not having incidence (NI)	Countries not having incidence but having awareness about ailments (NIHA)	Countries neither having incidence nor awareness about ailments (NINA)

TABLE 9 | Incidence and awareness comparision for TB.

	Countries discussing about TB	**Countries not discussing about TB**
Countries under high TB burden[a]	Brazil, India, Philippines	Indonesia, China, Nigeria, Pakistan, South Africa, Bangladesh, DR Congo, Ethiopia, Myanmar, UR Tanzania, Mozambique, Vietnam, Russian Federation, Thailand, Kenya, Uganda, Afghanistan, Cambodia and Zimbabwe
Countries not under TB burden[a]	US, Spain, Portugal, UK, Argentina, Chile and France	Rest of the world

[a]*www.who.int/tb/publications/global_report/en/.*

The awareness and actual occurrence of ailments can be compared and the countries can be divided into four groups based on this comparision. The groups are as follows:

- Countries with both occurrence and awareness.
- Countries that have awareness but no occurrence.
- Countries that have occurrence but no awareness.
- Countries that have neither occurrence nor awareness.

This has been illustrated in **Table 8**.

As an example, consider **Table 9** which gives the occurrence and awareness comparision for TB in various countries of the world. **Table 9** gives us various countries that have both awareness and occurrence of TB and also countries that have neither.

CONCLUSION AND FUTURE WORK

In the present work, data has been collected from a twitter live stream. A set of analytics and processing has been applied to the collected data to determine the awareness levels in each country or continent regarding each ailment. An approach for feature extraction has been proposed. The feature vectors hence derived are used for clustering. The primary aim of clustering is to determine clusters of countries with similar awareness levels. Various aspects namely, Visual Global Awareness Landscape (VGAL), Continent based Global Awareness Landscape (CGAL), Disease based Global Awareness Landscape (DGAL), Similarity based Global Awareness Landscape (SGAL), and Consolidated Global Awareness Landscape (Co - GAL), have been determined to present a holistic picture of the global awareness landscape of various ailments. This work has revealed that discussion or awareness about ailments and their incidence is not necessarily co-occurring. The analysis has also revealed that the countries can be divided into four groups namely:

- Countries having incidence and awareness of ailments.
- Countries not having incidence and awareness of ailments.
- Countries having incidence and no awareness of ailments.
- Countries niether having incidence and nor awareness of ailments.

The results of this work can be used by the governments of various nations and also international agencies like WHO to determine the countries that need immediate awareness drives for various diseases. Also, the nations that are highly aware can mentor other nations to spread awareness about these ailments. There is no centralized repository of global data available hence a direct comparitive study may not be possible. In the present work emphasis is placed on spatial analysis. A temporal analysis can also be done, which can also be seen as the future scope of the work.

DATA AVAILABILITY

The datasets for this study will not be made publicly available because The datasets are a part of sponsored research project and therefore cannot be made available directly in form of open data.

AUTHOR CONTRIBUTIONS

DT: conceptualization of the proposed methodology, idea, and guidance. SS, RA, and PM: partial implementation and documentation.

FUNDING

The partial funding for this work has been provided by Department of Science and Technology -Interdisciplinary Cyber Physical Systems (DST-ICPS), New Delhi, India and by Ministry of Human Resource Development (MHRD), India. We would also like to thank Indian Institue of Technology, Roorkee for supporting this research.

REFERENCES

Aramaki, E., Maskawa, S., and Morita, M. (2011). "Twitter catches the flu: detecting influenza epidemics using Twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processin* (Stroudsburg, PA: Association for Computational Linguistics).

Guha, S., Rastogi, R., and Shim, K. (2000). "ROCK: a robust clustering algorithm for categorical attributes," in *Proceedings 15th International Conference on Data Engineering* (Sydney, NSW: Information Systems), 345–366.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: Wiley.

Paul, M. J., and Dredze, M. (2011). "You are what you Tweet: analyzing Twitter for public health," in *ICWSM 20* (Barcelona).

Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza and H1N1 pandemic. *PLoS ONE* 6:e19467. doi: 10.1371/journal.pone.0019467

Smith, M. C., Broniatowski, D. A., Paul, M. J., and Dredze, M. (2016). "Towards real-time measurement of public epidemic awareness: monitoring in influenza awareness through twitter," in *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content* (Stanford, CA).

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). "Short text classification in twitter to improve information filtering," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. (Geneva: ACM).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

PERSPECTIVE
published: 28 August 2019
doi: 10.3389/fdata.2019.00029

Reflections on Gender Analyses of Bibliographic Corpora

Helena Mihaljević[1]*, *Marco Tullney*[2], *Lucía Santamaría*[3] and *Christian Steinfeldt*[1]

[1] Hochschule für Technik und Wirtschaft Berlin, University of Applied Sciences, Berlin, Germany, [2] Technische Informationsbibliothek (TIB), Hanover, Germany, [3] Amazon Development Center, Berlin, Germany

The interplay between an academic's gender and their scholarly output is a riveting topic at the intersection of scientometrics, data science, gender studies, and sociology. Its effects can be studied to analyze the role of gender in research productivity, tenure and promotion standards, collaboration and networks, or scientific impact, among others. The typical methodology in this field of research is based on a number of assumptions that are customarily not discussed in detail in the relevant literature, but undoubtedly merit a critical examination. Presumably the most confronting aspect is the categorization of gender. An author's gender is typically inferred from their name, further reduced to a binary feature by an algorithmic procedure. This and subsequent data processing steps introduce biases whose effects are hard to estimate. In this report we describe said problems and discuss the reception and interplay of this line of research within the field. We also outline the effect of obstacles, such as non-availability of data and code for transparent communication. Building on our research on gender effects on scientific publications, we challenge the prevailing methodology in the field and offer a critical reflection on some of its flaws and pitfalls. Our observations are meant to open up the discussion around the need and feasibility of more elaborated approaches to tackle gender in conjunction with analyses of bibliographic sources.

Keywords: gender, reproducibility, data science, bias, societal issues, science studies, automatic gender recognition

OPEN ACCESS

Edited by:
Katja Mayer,
University of Vienna, Austria

Reviewed by:
Doris Allhutter,
Austrian Academy of Sciences (OAW), Austria
Claire Donovan,
Brunel University London, United Kingdom

***Correspondence:**
Helena Mihaljević
helena.mihaljevic@htw-berlin.de

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *08 August 2019*
Accepted: *13 August 2019*
Published: *28 August 2019*

Citation:
Mihaljević H, Tullney M, Santamaría L and Steinfeldt C (2019) Reflections on Gender Analyses of Bibliographic Corpora. Front. Big Data 2:29. doi: 10.3389/fdata.2019.00029

1. INTRODUCTION

Despite the increasing number of women entering the Science, Technology, Engineering and Mathematics (STEM) fields, gender inequities persist. Women leave academia at a higher rate than their male colleagues, leading to significant female underrepresentation, particularly in permanent academic positions. A successful academic career has long been inextricably tied with a prolific scholarly record; scientific publications are not only the major outlet for scholarly communication, they are regarded as a proxy for a researcher's scientific credo and are one of the key factors in achieving and maintaining a flourishing career in academia. A natural question arises whether women and men differ in their publication practices in a way that contributes to the observed gender gap in STEM.

With the digitization of bibliographic metadata it became possible to approach this matter on a large scale using algorithmic, statistical, and computational methods. Several studies have leveraged existing databases to investigate the role of gender in academic publishing, either with a general focus (Larivière et al., 2013; West et al., 2013) or for particular disciplines, such as mathematics (Mihaljević-Brandt et al., 2016) or biology (Bonham and Stefan, 2017). In

Mihaljević-Brandt et al. (2016), we analyzed the scholarly output of about 150,000 mathematicians who authored over 2 million research articles since 1970. We showed that women abandon academia at a larger rate then their male counterparts, at different stages of their careers. We focused on aspects known to have a strong impact on career development, and concluded that, on average, women mathematicians publish in less prestigious journals and appear less frequently as single authors while they collaborate with a comparable-sized network of peers. These results prompted the interest for extending this line of analysis to other disciplines, work that is being continued in an ongoing interdisciplinary project[1].

Within the course of our investigations we have faced a number of critical aspects that are worth examining more closely. While we are certain that our results are relevant and reliable, we believe that some of the underlying assumptions and methods, though deemed valid and adequate given the available resources, deserve to be examined in more detail. Our ultimate goal is to foster a discussion on critical and sensitive topics that may potentially be encountered when making statements about individuals and existing societal issues based on publication metadata.

In this article we review a series of concerns that arise after critical examination of the core assumptions that ordinarily underlie gender inference from bibliographic data sources. We inspect common biases induced by gender assignment algorithms and other common data processing steps applied to bibliographic records. Finally, we discuss the reception and interplay of this kind of research within the field, and reflect on the issue of data and code availability and its effect on scientific standards like reproducibility. We discuss potential alternatives in order to foster a debate about best practices for subsequent projects.

2. CRITICAL ASPECTS OF THE ANALYSIS OF GENDER IN SCHOLARLY PUBLICATIONS

2.1. Assessing Humans

In bibliometric studies, the author's name is often the only piece of information susceptible of providing an indication of their gender. Name-to-gender inference is typically performed using a combination of multiple steps that usually involve querying name repositories like censuses or birth lists as well as applying insights from sociolinguistics. This is precisely how we approached the gender inference task in Mihaljević-Brandt et al. (2016). Recent analogous studies are increasingly making use of web services that continuously gather data from multiple sources. The results are sometimes augmented by applying, e.g., face recognition software to images retrieved when using a search engine to look up the author's name string.

Many issues arise in connection with said approaches. The resulting processes are seldom transparent, reproducible, or transferable; most studies relying on name-based gender inference fall short on thoroughly evaluating potential biases (Santamaría and Mihaljević, 2018). Enhancing name-based gender inference by facial analysis algorithms might incur an additional significant bias, particularly against darker-skinned women (Buolamwini and Gebru, 2018). Moreover, such approaches only allow for a binary definition of gender, which fundamentally excludes individuals that do not conform to this societal concept. This topic is typically not further discussed in the relevant literature. Ultimately and from a statistical point of view, this exclusion is considered "bearable": the estimated share of transgender and other non-binary authors is considered low enough that the binary gender simplification does not significantly distort the results. And yet, this enormously diminishes the needs and practices of transgender authors. Moreover, from the perspective of an individual who identifies outside the binary model, every such study is another manifestation of a "misgendering" practice in which the person is refused to be considered as part of the target group. In fact, automatic misgendering from an algorithm tends to be perceived as even more harmful than if it originated from another person (Hamidi et al., 2018).

The problem lies in the basic idea of inferring a person's gender form an attribute, such as the name string: personal names are assigned to individuals at birth as part of a schema based on a binary, immutable, and physiologically determined definition of gender (Keyes, 2018), much like other automatic gender recognition systems based on features, such as face, body, movement, or voice (Hamidi et al., 2018). Hence any approach that automates gender recognition (AGR) through a third-party mechanism, be it algorithmically or via human judgment, denies the view that one's gender identity is subjective (Butler, 1988), and embodies an old concept: an "incongruous pairing of futuristic AGR technology with old-fashioned conceptualizations of gender and its value to society" (Hamidi et al., 2018, 7), or as D'Ignazio (2016) puts it: "Non-binary genders will always be outliers."

Gender-inclusive bibliometric analyses can become possible only when no names or photographs are used as proxies for gender, allowing authors to define their gender autonomously instead. We have frequently thought about different approaches toward self-identification. A first idea was to draw a sample of authors and ask them to volunteer their gender. The drawbacks quickly become apparent, since authors can only be contacted via information taken from the publication's metadata. This introduces several issues: not every author provides their e-mail addresses, as often only the lab's or research group's PI is listed as corresponding author; then, only part of the contacted researchers would respond to such a request, which further prevents the creation of a random subsample; finally, the legal ramifications of using e-mail addresses for this purpose are far from clear. Moreover, the procedure would have to be repeated for every new study, leading to an unfeasible approach. Especially the latter argument begs for a sustainable and scalable solution. A second idea was to provide a web service to facilitate gender self-identification. If taken seriously, such an infrastructure should not be part of a time-limited research project, but instead exist

[1]"A Global Approach to the Gender Gap in Mathematical, Computing, and Natural Sciences: How to Measure It, How to Reduce It?" https://gender-gap-in-science.org

as a persistent service, preferably run by a suitable organization. Such a service would presumably take a long time to become widespread in the scientific community, even if researchers considered it meaningful enough to provide data.

It is therefore impossible to accurately assign a gender to all authors without misgendering certain groups of individuals, and it seems difficult to design and implement a service for self-identification to generate a solid database that could be utilized for sound statistical analyses. This begs the question of whether such analyses are in fact necessary and what benefit they provide to societal development. Every analysis bears the risk of reinforcing gender stereotypes and binary gender models. External attribution of properties like gender is not only difficult and biased, it is an infringement of the autonomy of the people who are subjected to it: "Simply starting with the assumption that all data are people until proven otherwise places the difficulty of disassociating data from specific individuals front and center" (Zook et al., 2017). There should be a good reason to conduct analyses that require assigning gender to individuals; we decided to perform them because academia is notoriously not gender-agnostic and because gender differences can be observed and need to be explained. Yet there is a fine line between analysing gender inequalities and reinforcing gender as a category, and we still would like to see processes like publishing and hiring become as gender-agnostic as possible.

2.2. Simplification and Selection Biases

The preparation of bibliographic records involves various algorithmic routines, which might be rule-based (e.g., comparison of affiliation strings with geo-databases), rely completely on third-party sources (e.g., usage of name-to-gender probabilistic assignments from commercial web services), or involve non-trivial machine learning models (e.g., linkage of authorship records to author entities). Thus, the resulting data set is the product of multiple data preprocessing steps and as such naturally susceptible to errors. It is best practice to estimate the inaccuracies of the involved procedures as realistically as possible, in particular when modeling social phenomena. However, this is often a highly complex and resource-consuming task that unsurprisingly falls short on many occasions, not only in commercial data science projects but also in scientific studies.

Large data sets typically require more preprocessing work. On the positive side, and in contrast to empirical work based on small samples, researchers can afford to exclude data points that do not contain sufficient information for the subsequent data mining steps (or, in other words, contain missing values in relevant variables that cannot be adequately inferred). At the same time, removal of data points induces bias. An illustrative example is the exclusion of the majority of Chinese names: these can stem from thousands of characters whose multiple meanings frequently reflect certain gender stereotypes. Much of this information is lost through romanization, which normally takes place when Chinese authors publish in Western journals.

The example above illustrates two kinds of biases often encountered in bibliographic analyses (Ridge, 2015): selection bias, which describes the tendency to skew data sources toward the most accessible subsets, and sampling or exclusion bias, which introduces a distortion of the data sets toward certain subgroups. Analogous examples abound: record linkage algorithms work worse for authors with very common names; author profiles of women are more often incomplete due to larger probability of family name changes; researchers with names of East-European origin are harder to cluster due to varying spellings from different name transliterations. This list is far from complete but already indicates that a precise specification and quantification of the biases induced through preprocessing is practically impossible.

While bias is typical for projects and applications from data science or machine learning, it is regularly left unaddressed in many business applications and scientific projects. This is somewhat surprising given the fact that data science practitioners often have a background in traditional sciences, where the identification and removal of bias when reasoning about the world are of high importance (Ridge, 2015). Luckily, there is a growing number of research communities, such as "Fairness, Accountability, and Transparency in Machine Learning" (FATML) that address the transparency of algorithmic decisions and the reduction of induced biases, partly in reaction to recent examples of discrimination caused especially by computer vision software amplifying existing societal prejudices.

Recommendations on how to recognize and avoid bias in data science are increasingly becoming mandatory, leading to the formulation of judicious best practices that ought to be implemented regardless of the concrete task at hand. In order to make research as transparent and reproducible as possible, one should at the very least track raw data sources comprehensively; provide quantitative and qualitative information about them; record and summarize data processing pipelines; describe all data transformations and explore their effect; and write and publish reproducible code (Ridge, 2015). Recent work by Gebru et al. (2018) formalizes this in a sense by proposing a framework to document data sets with data sheets containing a list of standardized questions: why a data set was created, who funded it, what preprocessing has been done, and in case it relates to real people, whether they agreed to the data usage. Still, these best practices will be challenged in many projects, especially in those that make use of closed data not available for secondary analyses.

2.3. Interaction With the Field

An intriguing and partly surprising result in Mihaljević-Brandt et al. (2016) is the underrepresentation of female authors in high-ranked journals, evaluated with respect to two prominent ranking schemes. In mathematics, as well as in other fields, it is commonsensical to expect the perceived quality of the journals where authors publish to be relevant for their scientific career. However, we cannot quantify how relevant it is. The available data does not allow us to transfer our found correlation between gender and journal rank into a model for the observed gender gap in mathematics. Modeling female mathematicians' careers would require much more information beyond publication data, thus no inference or predictive model can be produced based solely on studying bibliometric records.

Yet in fact, we are certain that the observed inequality regarding top-journal publications is causally related to the

higher drop out of women mathematicians, but we cannot prove it. A causal link seems probable, but has not been found: "An interesting pattern, by definition, is one that has a non-negligible subjective or logical probability of being potentially explicable, at least in part. It is possible to judge that a pattern has an underlying explanation even if we are unable to find it" (Good, 1983). The proof of a causal effect usually requires some sort of experiment, but the most one can really expect from working with observational data is correlation. As argued further in Villa (2018), there are still certain benefits of talking about causality explicitly even if it may not be demonstrable. For one thing, we constantly operate like this without being able to perform confirmatory experiments, but, more importantly, it suits the purpose of the undertaken data analysis: "When you analyze data [it] is because you want [to] arrive to some conclusions to take further actions. If you think in that way, is because you think those actions affect (and thus are a cause of) some quantity of interest. So, even [when] you talk about correlations for technical correctness, you are going to use those insights in a causal way" (Villa, 2018).

Although we are able to exclude the choice of subfield as a relevant factor, we cannot conclusively deduce why women publish less in high-ranked journals. Are women simply less likely to submit an article to them, or are they more frequently rejected? To fill the "causality gap" we resorted to a different data source. We recently conducted a global survey of scientists in STEM, in which participants were also asked to quantify the number of their publications submitted to a renowned journal within the last 5 years. A preliminary evaluation of the responses indicates that, on a global scale, women and men perceive that their submission practices in that respect are comparable.

Considered as part of the big picture, our result is thus a good example of what Tukey (1962) calls "approximate knowledge," referring to the maxim that data analysis progresses by offering approximate answers to the right questions. It also showcases the importance of exploratory analyses, which are essential to be able to formulate appropriate discussion points and to plan further data acquisition (Tukey, 1993). Presently it seems sensible to demand more transparency from publishers regarding their publication acceptance data. Journal rejection rates split by gender should be openly shared, since that would ultimately help elucidate the reasons for the underrepresentation of women in "renowned" journals. The formulation of such demands, though, would position one's own work within a system of institutional decision-making, moving it further away from a descriptive approach which rather focuses on revealing differences between genders within academia. While a descriptive approach might appear more "objective" and pure, it is arguable whether bibliometric research can be isolated in that way at all. As discussed in Angermuller and van Leeuwen (2019), who studied the societal role of bibliometric and scientometric research from Michel Foucault's perspective on science as power-knowledge, descriptive research that uses numbers to represent social realities is necessarily a constitutional part of such realities. As such, bibliometric research "cannot simply render a given state of the social world reality without intervening in it."

Certainly, our research can be used to compare groups of individuals, and it is challenging to estimate the exact effect it might have on academic decision-making. For instance, the conclusion that women publish less than men in a given period of time can be used to justify the lack of women among professors or grant recipients. Thus, without placing results within the right context and formulating clear goals, research on effects of gender on publication practices could help objectify and justify already existing inequalities between groups of academics. We believe, however, that this demands domain-specific expertise, which is crucial to be able to formulate relevant research questions for different fields or "to balance appropriate assumptions with computationally efficient methods" (Blei and Smyth, 2017). As posed in Good (1983), "even an exploratory data analyst cannot expect to obtain truly deep results in a science with which he is unfamiliar unless he cooperates with a scientific specialist."

One other obstacle when communicating results of data-driven research is the non-availability of data, code, and other artifacts that would enable reproducibility of the findings, identification of errors, or creation of derived investigations. Making research openly available includes providing open data and openly published software code. This is especially important if working on big data sets when far-reaching preprocessing steps are applied. In fact, reproducibility is one of the key requirements of (at least) future research (Donoho, 2017) (less critical are Shiffrin et al., 2017). Many data sources are not open. In our research we used paywalled databases, especially the large zbMATH corpus. We archived data and code and ensured that it can be accessed—if the rightholders of the database allow. This is not optimal, yet it is a first step. But in a general sense and for a broader public, our research is not reproducible—as it is the case of many data science projects.

When research results shall influence people's lives, every necessary step should be taken to make studies as reliable as possible. Data needed to reproduce the findings has to be archived, and its long-term availability ought to be guaranteed (Waltman et al., 2018). When working with open data, a data repository has to be found. When working with closed data, additional steps are necessary to ensure that other researchers will be able to access it. Relying on data not available for secondary analyses should be the very last resort, and researchers shall always try to make their data and software accessible. This might include negotiating with rightholders of databases. These efforts should at least be documented, if working with non-open data and code seems inevitable in some cases. At the very least it should be possible for other researchers to have a way to check the original results. A special meaning comes to this question when we talk about bias in research designs, data and algorithms. A middle ground that could be used more is the provision of aggregated data and visualizations, including interactive ones that offer researchers and other interested parties a better insight into the data and findings (we are following this path in our current project).

3. DISCUSSION

With each publication of their research findings, scientists expose their work to the public. But scientists themselves might become data points for measurements or analyses of scientific practices, often without being aware of the concrete usage of their data and without the possibility to interact or exert any influence on it. This is in particular the case when demographical features, such as gender or country of origin, are the subject of investigations. It is thus of the utmost importance for data scientists working in this field to "recognize the human participants and complex systems contained within their data and make grappling with ethical questions part of their standard workflow" (Zook et al., 2017).

We have discussed some troublesome but fundamental aspects frequently encountered in analyses of bibliographic records with respect to gender. We have problematized the process of inferring an author's gender solely from metadata like a name string, which is not only in stark contrast with a subjective and internal perception of gender but also runs the risk of misgendering individuals who do not conform to the gender binary. Due to a lack of alternatives that do not infringe the subject's autonomy, and the risk of reinforcing gender stereotypes and binary gender models, we find it important to keep questioning the necessity of any given gender-related data analysis and to compare the objectives and effects of our own research (to disclose gender inequalities) with the methodological compromises we make (e.g., reinforcing a binary gender model). For research like ours that lies at the intersection of data science and sociology, it is paramount to reflect on the interpretations and usages of one's research within the field. We believe that it is almost impossible to treat such research as solely descriptive or exploratory; we would instead propose considering the research context more closely and formulating the goals in a transparent way in order to minimize the risk of misusage for objectification or reinforcement of existing inequalities. In our opinion, a solid contextualization of analyses involving social phenomena and human participants demands domain-specific expertise, ultimately leading to interdisciplinary collaborations. Such collaborations, especially those involving qualitative methods, might be able to shed some light on the mechanisms that lead to the observed differences between male and female authors.

In Mihaljević-Brandt et al. (2016), we highly benefited from our expertise in mathematics and gender studies, in data science and in working with bibliometric data. We believe that previous domain knowledge helps to address shortcomings, such as the recognition of biases induced through data selection and processing and their potential effects. This topic, while often neglected in studies based on exploratory data analyses, is of high relevance for the actual conclusions that follow from the obtained results. The difficulty of specifying and quantifying the bias more precisely, but also the natural demand for reproducibility of research, make it all the more important to provide open access to raw data plus the software code. The analysis of bibliographic data is often based on closed data sources stored in paywalled corpora. Since such research has the potential to influence people's lives, we believe that scientists in this field should put considerable efforts into finding acceptable solutions and compromises with the rightholders of databases.

These hurdles are not easy to overcome. Domain expertise can be ensured by inviting researchers from the field to collaborate, thus fostering multidisciplinary research. This, however, might lead to difficulties, e.g., due to mainstream expectations in a discipline. Given the ubiquity of commercial bibliographic databases, ensuring sustainable access to comprehensive open bibliographic data will need additional and combined efforts of researchers and others (e.g., librarians).

AUTHOR CONTRIBUTIONS

HM and MT conceived the idea for the report and wrote the first draft of the manuscript. CS and LS contributed to the design and to specific sections. HM, LS, and MT edited and corrected the text. All authors read and approved the submitted version.

ACKNOWLEDGMENTS

This work was informed by the authors' participation in the project A Global Approach to the Gender Gap in Mathematical, Computing, and Natural Sciences: How to Measure It, How to Reduce It? funded by the International Science Council (ISC).

REFERENCES

Angermuller, J., and van Leeuwen, T. (2019). "On the social uses of scientometrics: the quantification of academic evaluation and the rise of numerocracy in higher education," in *Quantifying Approaches to Discourse for Social Scientists, Postdisciplinary Studies in Discourse*, ed R. Scholz (Cham: Springer International Publishing), 89–119.

Blei, D. M., and Smyth, P. (2017). Science and data science. *Proc. Natl. Acad. Sci. U.S.A.* 114, 8689–8692. doi: 10.1073/pnas.1702076114

Bonham, K. S., and Stefan, M. I. (2017). Women are underrepresented in computational biology: an analysis of the scholarly literature in biology, computer science and computational biology. *PLoS Comput. Biol.* 13, 1–12. doi: 10.1371/journal.pcbi.1005134

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Volume 81 of Proceedings of Machine Learning Research*, eds S. A. Friedler, and C. Wilson (New York, NY: PMLR), 77–91.

Butler, J. (1988). Performative acts and gender constitution: an essay in phenomenology and feminist theory. *Theatre J.* 40, 519–531. doi: 10.2307/3207893

D'Ignazio, C. (2016). *A Primer on Non-binary Gender and Big Data*.

Donoho, D. (2017). 50 years of data science. *J. Comput. Graph. Stat.* 26, 745–766. doi: 10.1080/10618600.2017.1384734

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach III, H. D., and Crawford, K. (2018). Datasheets for datasets. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1803.09010 (accessed April 8, 2019).

Good, I. J. (1983). The philosophy of exploratory data analysis. *Philos. Sci.* 50, 283–295. doi: 10.1086/289110

Hamidi, F., Scheuerman, M. K., and Branham, S. M. (2018). "Gender recognition or gender reductionism?: the social implications of embedded gender recognition systems," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18* (New York, NY: ACM), 8:1–8:13.

Keyes, O. (2018). The misgendering machines. *Proc. ACM Hum. Comput. Interact.* 2, 1–22. doi: 10.1145/3274357

Larivière, V., Ni, C., Gingras, Y., Cronin, B., and Sugimoto, C. R. (2013). Bibliometrics: global gender disparities in science. *Nature* 504, 211–213. doi: 10.1038/504211a

Mihaljević-Brandt, H., Santamaría, L., and Tullney, M. (2016). The effect of gender in the publication patterns in mathematics. *PLoS ONE* 11:e0165367. doi: 10.1371/journal.pone.0165367

Ridge, E. (2015). *Types of Bias and How to Avoid Bias in Data Science*.

Santamaría, L., and Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Comput. Sci.* 4:e156. doi: 10.7717/peerj-cs.156

Shiffrin, R. M., Brner, K., and Stigler, S. M. (2017). Scientific progress despite irreproducibility: a seeming paradox. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1710.01946 (accessed April 8, 2019).

Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Stat.* 33, 1–67. doi: 10.1214/aoms/1177704711

Tukey, J. W. (1993). *Exploratory Data Analysis: Past, Present and Future*. Technical Report 302. Princeton, NJ: Princeton University.

Villa, A. R. D. (2018). *Why Do We Need Causality in Data Science?—Towards Data Science*.

Waltman, L., Hinze, S., Scharnhorst, A., Schneider, J. W., and Velden, T. (2018). Exploration of reproducibility issues in scientometric research part 1: direct reproducibility. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1804.05024 (accessed April 8, 2019).

West, J. D., Jacquet, J., King, M. M., Correll, S. J., and Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLoS ONE* 8:e0066212. doi: 10.1371/journal.pone.0066212

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., et al. (2017). Ten simple rules for responsible big data research. *PLoS Comput. Biol.* 13, 1–10. doi: 10.1371/journal.pcbi.1005399

Conflict of Interest Statement: LS was employed by company Amazon.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

PERSPECTIVE
published: 11 September 2019
doi: 10.3389/fdata.2019.00032

AI for Not Bad

*Jared Moore**

Wadhwani Institute for Artificial Intelligence, Mumbai, India

Hype surrounds the promotions, aspirations, and notions of "artificial intelligence (AI) for social good" and its related permutations. These terms, as used in data science and particularly in public discourse, are vague. Far from being irrelevant to data scientists or practitioners of AI, the terms create the public notion of the systems built. Through a critical reflection, I explore how notions of AI for social good are vague, offer insufficient criteria for judgement, and elide the externalities and structural interdependence of AI systems. Instead, the field known as "AI for social good" is best understood and referred to as "AI for not bad."

Keywords: artificial intelligence, social good, not bad, ethics, data science, critical reflection

INTRODUCTION

We have begun to apply artificial intelligence (AI) to areas that claim to interact with "social good." New academic centers and initiatives label themselves as such. Cornell and Berkeley work on human-compatible AI[1] and Stanford's Human-Centered AI initiative aims "to advance AI research, education, policy, and practice to improve the human condition."[2] The University of Hong Kong claims to work on "beneficial AI."[3] The University of Washington and the University of Chicago offer programs on "data science for social good,"[4] while Harvard and the University of Southern California call it "AI for social good."[5]

These efforts carry over into conferences. At the prestigious AI conferences NeurIPS, ICML, and ICLR this past year, one group led workshops on "AI for social good."[6] Bloomberg News has held an annual "Data for good exchange" conference since sponsoring a "special event" at ACM KDD in 2014, a year where the overall conference had the theme "Data Science for Social Good," defined as "applying data science to improve civic and social outcomes."[7] A 2018 talk at ACM SIGIR used same term (Ghani, 2018) and is similar to non-academic conferences like "AI on a social mission"[8] and the "Rework AI for Good Summit."[9] Philosophers, too, have asked, "For The Public Good? Values and Accountability in AI and Data Science."[10]

The world outside of universities has not been quiet. Google, Facebook, IBM, and Intel have pages on "AI for social good"[11] and Microsoft has one about "AI for good."[12] AI research labs like

OPEN ACCESS

Edited by:
Katja Mayer,
University of Vienna, Austria

Reviewed by:
Hemank Lamba,
Carnegie Mellon University, United States
Luke Stark,
Microsoft Research, United States

***Correspondence:**
Jared Moore
jared@jaredmoore.org

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *31 March 2019*
Accepted: *30 August 2019*
Published: *11 September 2019*

Citation:
Moore J (2019) AI for Not Bad. Front. Big Data 2:32. doi: 10.3389/fdata.2019.00032

[1]https://research.cornell.edu/research/human-compatible-ai; https://humancompatible.ai
[2]https://hai.stanford.edu/
[3]https://caire.ust.hk/about/about-caire
[4]https://dssg.uchicago.edu/; https://escience.washington.edu/dssg/
[5]https://cyber.harvard.edu/story/2018-12/embracing-ai-social-good; https://www.cais.usc.edu/about/mission-statement/
[6]https://aiforsocialgood.github.io/2018; https://aiforsocialgood.github.io/icml2019/; https://aiforsocialgood.github.io/iclr2019/
[7]https://www.bloomberg.com/company/d4gx/; https://www.kdd.org/kdd2014/bloombergpress.html
[8]http://iaenmissionsociale.com/
[9]https://www.re-work.co/events/ai-for-good-summit-san-francisco-2019
[10]https://psa2018.philsci.org/en/74-program/program-schedule/program/124/psa2018-public-forum-for-the-public-good-values-and-accountability-in-ai-and-data-science
[11]https://ai.google/social-good; https://fbaiforindia.splashthat.com; https://www.ibm.com/watson/advantage-reports/ai-social-good.html; https://www.intel.ai/ai4socialgood/
[12]https://www.microsoft.com/en-us/ai/ai-for-good

AI2, WadhwaniAI, and MILA, respectively discuss AI for "common good," "social good," and "humanity."[13] Government initiatives from India, the U.S., and China do similarly[14].

"Social good" shifts between social responsibility, societal impacts, society, common good, the good, development, and ethics. Its proposals come in similar forms: calls for more data, better data, broader application, more diverse voices, reflexivity, transparency, changes to funding priorities, more education, more regulation—more.

The meaning of artificial intelligence shifts as well. It may mean "algorithmic systems," or "automated decision making" (Harris and Davenport, 2005)—other times, it is synonymous with "data science" or "big data." It also could be the case that AI does not truly exist and only refers to some yet-to-come future (Walch, 2018) when, presumably, this "social good" will actually be achieved. To others, that AI does not exist is misleading (Schank, 1987; Bringsjord and Schimanski, 2003). To such technical minds, AI would chiefly refer to a set of techniques like machine learning, deep learning, active learning, or reinforcement learning[15].

"AI for the good" de-politicizes the problems addressed. Many of these problems, like poverty, recidivism, and the distribution of resources, are ones of institutional failure. Technology-based approaches, when not aimed at the root of problems, divert attention from the proper recourse: structural change.

In this paper, I offer a critical perspective on the use of language of AI practitioners like myself who, from practice to theory, apply their work to some definition of "good." I use discursive analysis to explore the space between the notion of such projects and their actuality. In so doing, I follow Green (2018) in identifying AI systems as inherently political. Vague terms are the wagons of a modern gold rush into the promised riches of a mythic AI frontier. Like the California gold rush, this expansion may bring environmental degradation, concentrations rather than distributions of wealth, and the oppression of marginalized populations.

It is not the primary aim of this paper to synthesize a definition of AI, social good, or their combination. Chiefly, I theorize about what the apparent use omits. Nonetheless, I do offer and argue for a preliminary definition of good in section three. I use the term "data science" to loosely denote AI systems. For clarity hereafter and unless otherwise noted, "AI for the good" or "AI for social good" will encompass the above uses as they exist today and will refer to the projection of the computational discipline onto some definition of public or societal good. AI itself means, and will be used to mean in this paper, more than just the application of a statistical model like logistic regression to a dataset: it will mean the notions associated with such systems, the specifics of which I will explore below.

This paper proceeds in four parts. First, I review relevant literature. Second, I argue why "AI for the good," as it is used, is inappropriate. Third, I address possible critiques of my approach. Fourth, I suggest directions for those who aim to work in "AI for the good."

LITERATURE REVIEW

Many have already studied the components of "AI for the good." I review these attempts in four parts. First, I establish the precedent for practitioners to reflect on data science. Second, I summarize critiques of AI systems and language. Third, I review promising directions for the field. Fourth, I present attempts AI practitioners have made to improve elements of "AI for the good."

First, following Agre (1997) and Iliadis and Russo (2016), I critically reflect on data science. I draw on science and technology studies and discursive analysis to bolster the integrity of scientific knowledge through "socially robust knowledge" (Nowotny, 2003). I speak to practitioners of AI as well as to those who study the use of such tools.

Second, existing works provide or analyze the meanings beyond the underlying functioning of AI systems. There are claims that these data-focused technologies might overcome theory (Anderson, 2008) or transform modern life (Mayer-Schönberger and Cukier, 2013). In examining "the algorithm as a thing and the algorithm as a word," I choose words rather than the content of techniques as the site of critique (Beer, 2017, p. 9). Words are crucial because "by definition, a technological project is a fiction, since at the outset it does not exist, and there is no way it can exist yet because it is in the project phase" (Latour and Porter, 1996). "AI for social good" is one such project—if it already existed, why say so? Even AI alone, "evokes a mythical, objective omnipotence, but it is backed by real-world forces of money, power, and data" (Powles, 2018). Here Beer"s dichotomy between the algorithm as a word and as a physical manifestation becomes evident. Associating other words with AI—like intelligent, good, or society—creates notions of efficiency, neutrality, and progress, like how many technological metaphors (Stark and Hoffmann, 2019) "are myths that suffuse modern society" (Dalton and Thatcher, 2014) wielding power.

In rebutting common notions of neutrality from practitioners, Green focuses on the political nature of AI technologies. He thoroughly argues that data science should be seen of as political and, responding to the frequent practitioner argument that "We should not let the perfect be the enemy of the good," states, "data science lacks any theories or discourse regarding what "perfect" and "good" actually entail" (Green, 2018, p. 19). The pro-technology argument takes "for granted that technology-centric incremental reform is an appropriate strategy for social progress" (Green, 2018, p. 19) without having to worry about how (or whether) this actually occurs. This belief that the introduction of a technology is sufficient to yield a positive end is often, like by Dalton and Thatcher (2014), called technological determinism.

Third, there are promising approaches to define "good" with regard to AI. Social work provides one application. Tambe and Rice propose a union between social workers and AI practitioners, because "AI can be used to improve society and

[13] https://allenai.org/; http://wadhwaniai.org/; https://mila.quebec/en/ai-society/

[14] https://www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf; https://www.whitehouse.gov/ai/; http://www.baai.ac.cn/blog/beijing-ai-principles

[15] Schank (1987) and Bringsjord and Schimanski (2003) allow that AI exists given a narrow functional definition and explore the complications of intelligence.

fight social injustice" (Tambe and Rice, 2018, p. 3). Patton, a social worker academic, finds footing for such a union and identifies ways AI practitioners can engage well—largely by privileging those with whom they work (Patton, 2019). D'Ignazio adds to this by applying a social work code of ethics to data scientists, making explicit the principles to which data scientists seldom commit, like commitments to social justice and to the communities with whom they work (D'Ignazio, 2018).

Fourth, AI practitioners use terms like "AI for good" seemingly without regard to their notional or metaphorical value, but some engage with what might constitute "good." Practitioners, like Niño et al. (2017), use "social good" as a *domain* from which to solve problems ("the field of social good") (Niño et al., 2017, p. 896). These projects are designed for "serving the people who are in need globally, improving the society we live in and people's conditions within it" and make up application areas like health care, ecology, human rights, child welfare, etc. (Niño et al., 2017, p. 897). Niño et al. characterize key areas in projects for "social good" in a framework including data ownership, ethics, sustainability, assessment, stakeholder engagement, etc. Nevertheless, they do not mention what makes a project constitute "social good" except as existing in one of the application areas[16], as described by Green"s and D'Ignazio"s critiques. Using "social good" as a domain risks allowing the constituent projects to be seen of as good even if they fail to meet principles espoused by others (like by having poor data management practices), use no principles at all, or, more importantly, meet a set of principles that actively violate the principles of social justice (but retain the term "good"). I will henceforth refer to this understanding of "social good" as the *domain definition*.

For example, Green questions the focus on crime prediction systems at USC's AI for Social Good initiative. He argues that the initiative bolsters racist and oppressive policing instead of working to address the structural problems which lead to police action (Green, 2018). Similarly, Palantir, a big data company that produces crime prediction systems for clients like the U.S. government, recently partnered with the United Nation's World Food Program (WFP) (World Food Program, 2019). One might argue that such an endeavor is "social good" given that WFP is a not-for-profit aimed at reducing poverty. Nonetheless, this partnership met a significant outcry from groups like the Responsible Data List (Easterday, 2019). Clearly, these groups interpret "social good" quite differently. Their disagreement indicates the insufficiency of the domain definition.

Other practitioners working in "AI for good" recognize limitations of their efforts. Researchers at IMB call for a shift to produce open AI platforms to mitigate one-off projects (Varshney and Mojsilovic, 2019). Maxmen questions the worth of the Big Data for Good project from a global telecommunications group in its use of call detail records to respond to disasters because governments might (mis)use the same data for surveillance (Maxmen, 2019). Along the same lines, but largely not using the term "AI for good," recent work in fairness, accountability, and transparency (FAT*) has aimed to define best principles and practices for AI systems. Like Greene et al. (2019) and Lipton (2016) note that such technical efforts occur in too limited a manner; they present reforms to structures that might better be replaced. Selbst et al. expand on these critiques to note how FAT* as a field misses the broader social context and might be better served focusing on process and collaborating deeply with domain experts (Selbst et al., 2019).

In an examination of the entire field of AI, as opposed to individual projects, Floridi et al. identify principles for the creation of a "good AI society" regarding under-use, mis-use, and over-use (Floridi et al., 2018). Improving on others, they use the term "AI for social good" just once and not in the context of a discipline, but rather to identify the application of their framework[17]. Notably, they focus on potential harms (like those possible from a general artificial intelligence) on an equal, if not greater, degree than current harms (like threats to individual privacy). This corresponds to their inclusion of under-use of AI as a risk. Given the current harms of AI, their "good AI society" may just be a "good bad society," or, the best of the worst.

Prominent AI practitioners have acknowledged some of the inherent risks and ambiguities of AI technologies (Dietterich and Horvitz, 2015; Horvitz and Mulligan, 2015), but they do so in a way that appears to just pay lip service to, and thus avoid, fundamental critiques. To paraphrase, they argue that the risks of AI technologies are important, but that the risks can only be solved by further development of AI technologies. The utopic notion of economic liberalism employs the same sort of rhetoric: because the free-market ideal has never been achieved, one can always argue that its failures are due to insufficiently free markets (Polanyi, 2001). Likewise, data scientists, instead of addressing critiques, focus on how to realize the ideal of datafication in society (Rouvroy et al., 2013); they reinforce a technological determinism. In this way, the use of "AI for the good," given the domain definition, appears to strategically avoid consideration that the risks of AI may be too great to consider any further development of the technologies.

Many arguments for and notions of AI technologies sit on lose ground. Critiques of these technologies highlight their limitations, often in the sense of technological determinism and the avoidance of structural problems. A greater focus on these political problems and an engagement with communities might reorient the field. With these in mind, I examine whether "AI for good" is appropriate to classify the field.

THE GOOD AND BAD OF "AI FOR GOOD"

When working for "the good" we must ask *which* good and *for whom*. By committing to definitions of what constitutes "good" and "bad" with regard to AI technologies, I examine the appropriateness of labeling the field as "AI for the good." I described in the literature review how the clearest criteria for "AI for good" is based on the domain with which an AI

[16]Their framework does address a "lack of standardized good practices to leverage the power of data" (Niño et al., 2017, p. 897).

[17]"This should involve a clear mission to advance AI for social good, to serve as a unique counterbalance to AI trends with less focus on social opportunities" (Floridi et al., 2018, p. 704).

technology interacts (the domain definition). We are meant to accept that because a project works on health, with not-for-profit organizations, in the space of climate change, on poverty-reduction, etc., that it is "for social good." In this section, my argument is as follows:

1. I provide an alternative definition of good according to the *capability approach* and social justice.
2. Following 1, there are projects that are good, but that are not labeled as such.
3. Following 1, AI technologies carry inherently bad externalities.
4. Following 3, in order to consider net goods, "AI for social good" must engage with and balance out these bads.

First, I offer a functional definition of "good" for an AI system using the capability approach and social justice. Green (2018, p. 4) cites (Collins, 2002) in defining a social justice project as "an organized, long-term effort to eliminate oppression and empower individuals and groups within a just society" and advocates for such projects in data science. Such a project can work in complement with the capability approach, a theoretical framework predicated on context-dependent individual freedom and well-being as defined by people's capabilities or real opportunities to act. This approach, particularly as evoked in the areas of information and communication technologies (Johnstone, 2007; Kleine, 2010), provides an operational lens for AI technologies. I use the capability approach with a particular focus on accountability and individual control over private information to highlight voices from historically marginalized communities Of course, one might disagree with my definition on many grounds—mine is neither radical (e.g., anti-capitalism) nor conservative (e.g., a defense of the status quo) enough and remains vague. My point is not so much to advance *this* definition as to advocate for discussion of *which* definition is most appropriate just as Green, Patton, and D'Ignazio do. Such a frame will then allow us to analyze claims of "social good." Suffice to say, a "good" intervention should be empowering (particularly of basic human functioning), address structural conditions of oppression, and perform at least as well as interventions using similar amounts of resources.

For example, imagine a project designed in partnership with a community in a specific West-African country with little access to health care. The project uses a computer vision application on a smartphone to screen babies for birth defects. This project might be viewed of as "good" given that it specifically works with a marginalized community and increases their capability to access health care. Further, the community might not have achieved the same access to care with a similarly-resource-intensive effort to train more medical professionals.

Second, using the definition of good from 1, there are projects which do not use the label "AI for the good" that might be classified as such. For example, consider recent efforts in federated learning to decentralize and distribute the computations constituent in the training of a model (McMahan and Ramage, 2017). These efforts address some concerns about the privacy of user data: such data might not need to be collected in the same centralized manner. Furthermore, one can imagine a fully-specified federated learning project that meets the criteria of 1. Despite this, the concept of federated learning does not carry the moniker social good.

Third, inherently bad externalities arise with AI technologies. Recent work has shown that model training creates a significant carbon footprint (Strubell et al., 2019). In order to create an AI system, one must employ many engineers and scientists and set-up infrastructure, all of which are costly—perhaps more so than other interventions. Even more significantly, enormous invisible and unacknowledged labor goes into labeling data for training purposes, much of which occurs under potentially or explicitly exploitative conditions (Gray and Suri, 2019). Datafication names the creep to record more of life in a manner that can be processed by a computer (Mayer-Schönberger and Cukier, 2013). It undergirds the bloom in AI—models need data to combine with human labels—but brings unknown harms. Data collected for what 1 day appears good may be used later for what may not accord the same definition of good. For example, data to improve resource distribution to parolees were later used to create a model to gauge how likely offenders were to recommit crimes (Angwin and Larson, 2016). Datafication works at odds with user privacy as seen with consumer hacks, behavioral advertising, and government surveillance (Zuboff, 2019).

"AI for good" distracts from the larger world in which AI exists. Public visibility does not acknowledge the interdependent and exploitative nature of the technologies. Labeling them as "for the good" positions them as somehow intrinsically better than the social systems on which they depend. For example, tech companies implement systems they acquire from start-ups created from academic research. Most research papers come from graduate students whose long working hours are enabled by the labor of custodial staff and food service employees. In order to respond to questions on the appropriateness of a long short-term memory or a hidden Markov model one must not just understand their error rates, but also how to calculate derivatives, engage in basic math, and use language—skills learned through years of, for most, public schooling and from hundreds of teachers. AI models run on machines made thousands of miles away by people practitioners will never meet. These machines draw electricity produced by fossil fuel workers and which is distributed through a grid maintained by scores more. The startups themselves, or the tech companies that buy up startups to "scale" their systems, then farm out the process of data labeling to vast networks of invisible workers (Gray and Suri, 2019). To even have the capacity to build an AI system requires what Anderson describes of as "joint-production" (Anderson, 1999, p. 321). Those involved in AI systems are not just the visible actors of engineers, scientists, researchers, program managers, marketers, negotiators, lawyers, or end users. These terms too precisely assign agency, ability, and intentionality to what is best described as panning the sediment of streams of data.

The point is not to decry actors who lay claim to terms like "AI for the good" so much as to question how their actions reflect on their stated goals. Those who use such terms may even believe that they are saving the field of AI from "not good" domains, that their research areas are the more appropriate direction. Given the overheads and externalities of AI, it is not clear there is such a

need at all to focus on "not good" domains. Even with criteria to label AI systems as "good," the inherent interdependence raises questions about whether AI is inherently "bad" and whether any domain can redeem the system of production.

"AI for the good" is strategically vague. Left out by the use of "AI for the good" is the intensely political nature of any one of the areas associated with the term (as in domain definition). Recall the USC AI for Social Good project on policing which Green named as oppressive. Indeed, according to the definition of good from 1, the USC project would be bad—it does privilege community voices and reinforces forms of oppressive policing (which restrict peoples" capabilities).

Furthermore, non-profit organizations, which at least some AI practitioners associate with their use of "social good,"[18] might not even desire such technology. For example, for these non-profits, technical contributions might be better spent on upgrading old systems (like from Windows XP) rather than spending resources to get data in the "right" format for building AI systems.

Fourth, this all suggests that to be considered "good," projects must commit to a definition of social good and then show that, even after considering negative externalities, on the balance they still achieve good. On the whole, then, projects might better consider the degree to which they are "not bad."

CRITIQUES

In this section, I consider four critiques of my argument.

First, detractors might chafe at a focus on the words of AI. They might argue that focusing on words ignores the substance of technologies which would actually bring about "good." Of course the substance of the technologies is important, but in this paper I focus on the use of language, which, as I make the case for in the introduction, is also important.

Second, one might posit that even if "AI for the good" is vague, the use of such terms does no harm. While the claim of vagueness has been used to decry the difficulty of regulating AI technologies (Scherer, 2015), we use vague terms like energy or manufacturing and are able to operationalize them (Danaher, 2018). In this sense, the absence of a definition would be permissible so long as we "know it when we see it." This is not the case with claims of "social good." Such a response is strategically vague; it elides the externalities inherent to AI technologies and uses the weak criteria of the domain definition. Harm comes in allowing ourselves to feel good while perpetuating oppressive systems and when misallocating resources.

Third, a reader might say that "social good" is just marketing speak—not what practitioners say. That may be so, but the term appears from research to implementation: in governments, in funding agencies, in research papers, at conferences, in companies, and in public discourse. Even if the majority of the use of "AI for good" occurs externally to AI practitioners, it is through these routes that the notions of AI manifest. That is, practitioners must care about how their work is used and not just what it is.

[18]https://www.research.ibm.com/science-for-social-good/

Still, one might argue that, fourth, despite its flaws, "social good" is a relevant distinction. Even in the absence of a more robust criteria, there is a difference between machine learning researchers choosing to work on credit card companies being defrauded vs. those working on disease modeling. I suggest that there is a better approach than to ignore the ambiguity, the insufficient criteria, and the externalities of AI. Instead of banishing "AI for good," we might rather rename the field.

SUGGESTIONS

In this critique of the use of language, I also offer a suggestion. Namely, we should stop labeling projects as "for social good" and instead use the term "for not bad." The latter more accurately evokes the need to avoid the inherent bad traits of AI technologies without falling into the traps involved with vague claims to "social good."

Practitioners who would still like to use terms like "AI for the good" should read literature that studies the criteria for evaluation of social change projects and then apply those criteria. This includes work in the health sciences, social sciences, development studies, economics, and more. In the scope of technological changes for implementing theories of a just society, the literature in Information and Communication Technologies for Development provides some examples. Conferences in this space include the ACM Conference on Computing and Sustainable Societies (COMPASS) and the Workshop on Computing Within Limits[19]. The journal *Information Technology and International Development* focuses on the background theory of such work[20].

With such a background, practitioners may be better prepared to define and measure criteria of "good" to expand on my attempt above. More work to quantify the externalities of AI projects [building on examples like Maxmen (2019) and Strubell et al. (2019)] will then fill out such criteria. This might include comparable metrics on cost, energy usage, and potential for future misuse of data.

Sustained interaction with those in communities that are to be "innovated" will further concretize what constitutes "good." Tambe and Rice (2018) and Patton (2019) demonstrate how this can be done with social work. Action research, like as related to human computer interaction by Hayes (2011), provides another lens for community interaction in terms of accountability and shared credit for results. Many "social good" initiatives already discuss a focus on partnerships[21]—these should be expanded and made sure to recognize, if not attempt to address, the underlying structural issues.

"AI for not bad" avoids some of the problems of "AI for good." It more honestly describes the current vagueness and centers the externalities. Practitioners unwilling or unable to commit to explicit notions of good should consider adopting it.

[19]https://acmcompass.org/; http://computingwithinlimits.org

[20]https://itidjournal.org/index.php/itid

[21]For example, WadhwaniAI partners with India's Central Tuberculosis Division https://wadhwaniai.org

CONCLUSION

"AI for the good" is vague, lacks sufficient criteria, omits the externalities of AI, and elides the structural interdependence of AI projects. "AI for the good" may really be AI for flashy slide decks, AI for difficult-to-maintain and highly interdependent computational systems, AI for new statistical methods, or (at best) AI for public health analyses that may end up saving lives. In this paper, I raise concerns about the presentation of the "AI frontier" as beneficent. Following Green, I ask that the field "AI for the good" recognize that, as it is now, it really constitutes "AI for not bad." Practitioners would more honestly embrace this label or else do the work necessary to legitimately claim good.

In this work, I advocate for a more honest discipline. I ask those out there who interact with AI at any level—the new student wondering where to put her time, the executive of a company—to consider what their use of language ignores.

"AI for social good" speaks to the desire of many of practitioners to share what opportunities they have. It sounds nice. It imagines a world of lucrative careers optimized to better humanity. The world is not so simple. Perhaps it is enough that society, as bolstered by science, has tended toward longer lives, more food, and less violence (Pinker, 2019; Rosling et al., 2019), but extrapolation will not alone resolve problems. AI practitioners, like myself, are part of the prospecting of science from which we hope for gold, but in which will we likely find just sand—and perhaps leave in our tailings environmental damage and labor displacement. Lest that be so, we must be honest about what we are doing and what we might do better.

AUTHOR CONTRIBUTIONS

JM contributed everything to this paper and is accountable for the content of the work.

ACKNOWLEDGMENTS

I would like to thank Johan Michalove and Dallas Card for helping advise this paper into life. Many thanks to Momin Malik and Katja Mayer for organizing the workshop on critical data science and editing this paper. I would also like to thank the reviewers for providing me much needed feedback.

REFERENCES

Agre, P. E. (1997). "Towards a critical technical practice: Lessons learned from trying to reform AI," in *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*, eds G. C. Bowker, S. L. Star, W. Turner, and L. Gasser (Mahwah, NJ: Lawrence Erlbaum Associates), 131–158. Available online at: https://web.archive.org/web/20040203070641/http://polaris.gseis.ucla.edu/pagre/critical.html

Anderson, C. (2008, June). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.* Wired. Retrieved from: https://www.wired.com/2008/06/pb-theory/

Anderson, E. S. (1999). What is the point of equality? *Ethics* 109, 287–337. doi: 10.1086/233897

Angwin, J., and Larson, J. (2016). *Machine Bias.* ProPublica. Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Beer, D. (2017). The social power of algorithms. *Inform. Commun. Soc.* 20, 1–13. doi: 10.1080/1369118X.2016.1216147

Bringsjord, S., and Schimanski, B. (2003). "What is artificial intelligence? Psychometric AI as an answer," in *IJCAI'03 Proceedings of the 18th International Joint Conference on Artificial Intelligence* (San Francisco, CA: Morgan Kaufmann Publishers Inc.).

Collins, P. H. (2002). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. New York, NY: Routledge. doi: 10.4324/9780203900055

Dalton, C., and Thatcher, J. (2014). *What Does a Critical Data Studies Look Like, and Why Do We Care? Seven Points for a Critical Approach to 'big data*.' Society and Space, 29. Retrieved from: https://societyandspace.org/2014/05/12/what-does-a-critical-data-studies-look-like-and-why-do-we-care-craig-dalton-and-jim-thatcher/

Danaher, J. (2018, September 27). *Is Effective Regulation of AI Possible? Eight Potential Regulatory Problems*. Retrieved from Institute for Ethics and Emerging Technologies website: https://ieet.org/index.php/IEET2/more/Danaher20180927 (accessed June 6, 2019).

Dietterich, T. G., and Horvitz, E. J. (2015). Rise of concerns about AI: reflections and directions. *Commun. ACM* 58, 38–40. doi: 10.1145/2770869

D'Ignazio, C. (2018, September 2). *How Might Ethical Data Principles Borrow from Social Work?* Retrieved from Medium website: https://medium.com/@kanarinka/how-might-ethical-data-principles-borrow-from-social-work-3162f08f0353 (accessed July 1, 2019).

Easterday, J. (2019, February 8). *Open Letter to WFP Re: Palantir Agreement*. Retrieved from https://responsibledata.io/2019/02/08/open-letter-to-wfp-re-palantir-agreement/

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds Mach.* 28, 689–707. doi: 10.1007/s11023-018-9482-5

Ghani, R. (2018). "Data science for social good and public policy: examples, opportunities, and challenges," in *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY: ACM), 3–3. doi: 10.1145/3209978.3210231

Gray, M. L., and Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Boston, MA: Houghton Mifflin Harcourt.

Green, B. (2018). Data science as political action: grounding data science in a politics of justice. *ArXiv [preprint]* arxiv: 1811.03435 [Cs].

Greene, D., Hoffmann, A. L., and Stark, L. (2019). "Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning," in *Hawaii International Conference on System Sciences* (Maui, HI). doi: 10.24251/HICSS.2019.258

Harris, J. G., and Davenport, T. H. (2005). Automated decision making comes of age. *MIT Sloan Management Review* 2–10.

Hayes, G. R. (2011). The relationship of action research to human-computer interaction. *ACM Transac. Comput. Hum. Interact.* 18, 1–20. doi: 10.1145/1993060.1993065

Horvitz, E., and Mulligan, D. (2015). Data, privacy, and the greater good. *Science* 349, 253–255. doi: 10.1126/science.aac4520

Iliadis, A., and Russo, F. (2016). Critical data studies: an introduction. *Big Data Soc.* 3:2053951716674238. doi: 10.1177/2053951716674238

Johnstone, J. (2007). Technology as empowerment: a capability approach to computer ethics. *Ethics Inform. Tech.* 9, 73–87. doi: 10.1007/s10676-006-9127-x

Kleine, D. (2010). ICT4WHAT?—Using the choice framework to operationalise the capability approach to development. *J. Int. Dev.* 22, 674–692. doi: 10.1002/jid.1719

Latour, B., and Porter, C. (1996). *Aramis, or, The Love of Technology*, Vol. 1996. Cambridge, MA: Harvard University Press.

Lipton, Z. C. (2016). The mythos of model interpretability. *ArXiv [preprint]* arxiv:1606.03490 [Cs, Stat].

Maxmen, A. (2019). Can tracking people through phone-call data improve lives? *Nature* 569, 614–617. doi: 10.1038/d41586-019-01679-5

Mayer-Schönberger, V., and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, And Think*. Boston, MA: Houghton Mifflin Harcourt.

McMahan, B., and Ramage, D. (2017, April 6). *Federated Learning: Collaborative Machine Learning without Centralized Training Data.* Retrieved from Google AI Blog website: http://ai.googleblog.com/2017/04/federated-learning-collaborative.html (accessed June 22, 2019).

Niño, M., Zicari, R. V., Ivanov, T., Hee, K., Mushtaq, N., Rosselli, M., et al. (2017). Data projects for "social good": challenges and opportunities. *Int. J. Hum. Soc. Sci.* 11, 1094–1104. doi: 10.5281/zenodo.1130095

Nowotny, H. (2003). Democratising expertise and socially robust knowledge. *Sci. Pub. Policy* 30, 151–156. doi: 10.3152/147154303781780461

Patton, D. U. (2019, March 24). *Why AI Needs Social Workers and Non-Tech Folks*. Retrieved from Noteworthy website: https://blog.usejournal.com/why-ai-needs-social-workers-and-non-tech-folks-2b04ec458481 (accessed July 1, 2019).

Pinker, S. (2019). *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. London: Penguin Books.

Polanyi, K. (2001). *The Great Transformation: The Political and Economic Origins of Our Time, Vol. 45*. Boston, MA: Beacon Press.

Powles, J. (2018, December 7). *The Seductive Diversion of 'Solving' Bias in Artificial Intelligence*. Retrieved from Medium website: https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53 (accessed March 20, 2019).

Rosling, H., Rosling, O., and Rosling Rönnlund, A. (2019). *Factfulness*. London: Sceptre.

Rouvroy, A., Berns, T., and Libbrecht, E. (2013). Algorithmic governmentality and prospects of emancipation. *Réseaux* 177, 163–196. doi: 10.3917/res.177.0163

Schank, R. C. (1987). What is AI, anyway? *AI Magazine* 8, 59–59.

Scherer, M. U. (2015). Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harvard J. Law Tech.* 29, 353–398. doi: 10.2139/ssrn.2609777

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT*'19* (Atlanta, GA). doi: 10.1145/3287560.3287598

Stark, L., and Hoffmann, A. L. (2019). Data is the new what? *Popular* metaphors and professional ethics in emerging data culture. *J. Cult. Analyt.* doi: 10.22148/16.036. [Epub ahead of print].

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *ArXiv [preprint]* arxiv:1906.02243 [Cs].

Tambe, M., and Rice, E. (2018). *Artificial Intelligence and Social Work*. Cambridge: University Press.

Varshney, K. R., and Mojsilovic, A. (2019). Open platforms for artificial intelligence for social good: common patterns as a pathway to true impact. *ArXiv [preprint]* arxiv:1905.11519 [Cs].

Walch, K. (2018, November 1). *Artificial Intelligence Is Not A Technology*. Forbes. Retrieved from: https://www.forbes.com/sites/cognitiveworld/2018/11/01/artificial-intelligence-is-not-a-technology/

World Food Program (2019). *Palantir and WFP Partner to Help Transform Global Humanitarian Delivery*. Retrieved from: https://www1.wfp.org/news/palantir-and-wfp-partner-help-transform-global-humanitarian-delivery

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. New York, NY: Public Affairs.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor and reviewer HL declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

frontiers in Big Data

PERSPECTIVE
published: 01 October 2019
doi: 10.3389/fdata.2019.00035

Experimenting With Algorithms and Memory-Making: Lived Experience and Future-Oriented Ethics in Critical Data Science

Annette N. Markham and Gabriel Pereira**

Department of Digital Design and Information Studies, Aarhus University, Aarhus, Denmark

In this paper, we focus on one specific participatory installation developed for an exhibition in Aarhus (Denmark) by the Museum of Random Memory, a series of arts-based, public-facing workshops and interventions. The multichannel video installation experimented with how one memory (Trine's) can be represented in three very different ways, through algorithmic processes. We describe how this experiment troubles the everyday (mistaken) assumptions that digital archiving naturally includes the necessary codecs for future decoding of digital artifacts. We discuss what's at stake in critical (theory) discussions of data practices. Through this case, we offer an argument that from an ethical as well as epistemological perspective critical data studies can't be separated from an understanding of data as lived experience.

Keywords: algorithms, critical data science, arts based research, memory-making, ethics, lived experience

OPEN ACCESS

Edited by:
Katja Mayer,
University of Vienna, Austria

Reviewed by:
Catherine D'Ignazio,
Emerson College, United States
Nick Seaver,
Tufts University, United States

****Correspondence:***
Gabriel Pereira
gpereira@cc.au.dk
Annette Markham
amarkham@gmail.com

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *24 April 2019*
Accepted: *13 September 2019*
Published: *01 October 2019*

Citation:
Markham AN and Pereira G (2019) Experimenting With Algorithms and Memory-Making: Lived Experience and Future-Oriented Ethics in Critical Data Science. Front. Big Data 2:35. doi: 10.3389/fdata.2019.00035

INTRODUCTION

In recent years, Google Photos and Apple Memories made headlines by promising to cut through the clutter of people's big data by automatically curating our most meaningful photos and videos. These services rely on machine learning and algorithmic processing of data. Far from neutral, these algorithmic services play a key role in how people enact and make sense of their everyday lives. Whether we use Helen Kennedy's phrase to describe this phenomenon as a form of "new data relations" (Kennedy, 2016) or Cheney-Lippold's "algorithmic identities" (Cheney-Lippold, 2011), algorithms are woven into everyday life at the most intimate levels (Gregg, 2011). As Markham (2015) puts it, this intimacy is one that we can see through the lens of a personal relationship, since algorithmic systems function as interpersonal "participants in a continual symbolic interaction process whereby our understandings of self, other, and our social worlds are co-constituted" (p. 5).

We agree with other critical data studies scholars (e.g., Kitchin and Lauriault, 2014; Iliadis and Russo, 2016) that laying out the granularity of how data is generated or represented is important because data analytic processes wield significant and often hidden power in shaping future knowledge, historical legacies, and social formations. As citizens go about their everyday lives and also reflect on various aspects of their lived experience, the power of data analytics presents a "seductive allure" of being "speedy, accessible, revealing, panoramic, prophetic, and smart" (Beer, 2019). As participatory action researchers, we are bridging the academic and public spheres to facilitate general users' knowledge around the idea, not uncommon among critical data scholars, that these "assemblages of data" are co-creators of future imaginaries, acting with moral agency to, as Martin (2018) notes, "silently structure our lives" (p. 2). Within this ecology, as Markham et al. (2018) emphasize, "The locus of responsibility and accountability for ethical design, behavior, and

outcomes is difficult to ascertain" (p. 1). We use the example of an artistic video installation we built called *Memory Glitch* to highlight this difficulty. Through three algorithmic transformations of an elderly woman's interview about her experiences in the second World War, we consider how future memories are impacted by algorithmic rewriting of the codecs, or formulas for encoding and decoding data formats. When and where this happens will of course vary: imagining the long future, it could be caused by data loss as physical memory storage devices decay; in the more immediate future, it could be within the automated memory management processes of organizing, prioritizing, and otherwise "curating" a file. It is by now a familiar criticism of algorithmic processes that multiple stakeholders and agents, human, and nonhuman, operate in these systems.

To this ongoing conversation we add the suggestion that focusing critical ethical attention on the algorithmic management of memory and meaning in unexpected ways can enhance the practices of critical data science. We do this partly by foregrounding the fragility of a person's recorded lived experience as it is algorithmically filtered, morphed, transformed, or otherwise remixed. But we seek to go beyond current scholarly refrains that digital archives are precarious, data modeling is flawed, or algorithms are biased. Instead, we build a case for using arts-based and personalized interventions as a way of enabling end users to better "apprehend (theorize, imagine)," in the words of Magalhães (2018, p. 3), the implications and moral agency of algorithmic processes in their everyday lives.

We have been studying these issues through the *Museum of Random Memory* (MoRM), a series of arts-based, public-facing experiments. Over 3 years we have conducted eight workshop/exhibitions in five countries to help people investigate how automated data-related processes might be influencing their own personal and cultural memories. This becomes a study of complex entanglements of lived experience, digitalization of memories, and algorithmic logics. MoRM is an interventionist action, involving an international group of artists, data scientists, filmmakers, computer scientists, scholars, activists, museum curators, lawyers, and university administrators. The eight experiments performed by MoRM have taken different paths of inquiry: some have focused on showing citizens how their digital traces are tracked as they search for things using a browser; others focus on complicating where and how memory is located in everyday analog/digital/data objects.[1]

A large part of the MoRM goal is to combine critical (theory) data studies with a future making orientation and to add examples that illustrate the importance of an ethic of care[2] in data science practices. The larger project critiques and imagines alternatives to normative ways of working and thinking through data. We believe that there is a troubled and important set of relationships to explore between humans, their data, digital platforms, machine learning trends, and multiple external stakeholders with political and economic interests. What scholar-activist roles can we take to intervene in these often taken-for-granted datascapes?

In what follows, we focus on *Memory Glitch*, a specific MoRM installation developed for exhibition at the *Affects, Interfaces, Events* conference, August 28–30, 2018 in Aarhus, Denmark. The multi-channel video installation experimented with how the memory of one person, Trine, can be decoded and rendered in three very different ways, through algorithmic processes. We describe how this experiment highlights visually and evocatively the everyday (mistaken) assumptions that digital archiving naturally includes the necessary formulas for future decoding of digital artifacts. We conclude by discussing what's at stake in critical (theory) discussions of data preservation practices.

MEETING THE "DATA"

It started as a conversation. One morning, as Trine was returning a book to the library, she walked by our MoRM exhibit and heard the MoRM researchers ask passersby to "donate a memory, a random memory, something you want to remember or forget." She went home, collected her artifact, and returned later that afternoon to donate her memory. The physical artifact she brought was a photocopy of some newspaper clippings where she, alongside some others, was featured as a jazz singer. The memory she wanted to donate, however, was quite different:

> *I want to donate the memory of the Germans occupying my home town in Northern Jutland when I was a little girl.*

As with other participants, we invited Trine to spend some time with a MoRM researcher to talk about her memory. Sitting with her in a cozy space, one of us asked Trine why she felt this memory was important, as well as how she thought digital preservation might influence what future archeologists might find if they dug up artifacts from 2017. As the conversation was being filmed, the researcher wrote a few sketchy notes on what Trine was saying:

> *Growing up during German occupation in Northern Denmark. 'People helped each other'. 'And we're losing that'. 'Poor, rich, didn't matter'. 'We oldies talk a lot about it when we meet at the bus stop'. It's boring to ride the bus (esp. 4-6 pm), and 'they never get up for you—even if you have a limp'.*

Trine reminded us repeatedly that it was crucial to make people remember this time period of Danish history. She expressed concern that "digital media make it more difficult for people to have conversations about the old days," and how "nobody really talks to each other anymore because they're busy on their phones."

Like many other participants at this exhibition, Trine spent far longer than we anticipated: 3 h. With Trine's approval, we made an audiovisual recording of her conversation. Her memories of

[1]For more details, see Rehder and Ostrowski (2017), Bratton et al. (2016), Markham (2019), or the project website (*https://futuremaking.space/morm/*).

[2]This stance, as articulated by Luka and Millette (2018), emphasizes "the integration of feminist and intersectional values into considerations of data analyses, including big data" (p. 4). In critical data studies or critical data science, if we follow the work of Hoffmann (2016, 2018), this means not only centralizing ethics but also considering how data construction, data modeling, and data processing might conduct violence on people, symbolically, culturally, or physically.

post-WWII Denmark became video files, stored in the project's hard drives.

Fast forward 1 year. The conversation becomes a meta conversation among the research team. We are combing through the archives of this event, searching for snippets to showcase at two academic conferences: *Data Justice* and *Affects, Interfaces, Events*. Trine's video has been a topic of much interest in our ongoing conversations. She is an engaged citizen, telling a poignant story, which makes her video an affecting piece. But how much should we edit this piece? Because her conversation wanders off point frequently and the interview lasted 3 h, we know we need to cut it in many ways to reshape it for the new exhibition. We also discuss how we might remix the video to highlight only certain points. These are natural decisions any journalist, filmmaker, or artist might make. For us, it raised serious questions from an ethical perspective.

First, what is our justification for remixing or altering someone's memory after they've donated it to us for safekeeping? Second, should we show people's memories in a different context than the one in which they made the original donation? What is our responsibility toward the people we've encountered and the data we've collected? Trine believed her contribution would be saved, archived as part of a larger digital preservation project. She believed her story would remain whole. She believed it would be accessible in the future. Of course she signed a consent form and agreed to future transformations, but to what extent should curators and archivists take responsibility for developing the public's understandings of digital preservation? An ethic of care means more than just meeting needs or expectations, but, as characterized in design disciplines, "doing so in a manner that is attentive, responsive, and respectful to the individuals in need of care." (see also Edwards and Mauthner, 2002; Engster, 2005; Luka and Millette, 2018). Avram et al. (2019) suggest this both complicates and requires "fundamentally dialogic and adaptive tinkering that defies a factual evaluation or judgement of practice."

After much debate, we agreed that even with these ethical troubles, we should still show pieces of this video conversation. Remixing Trine's memories into a montage of sound and images, through glitch art techniques, would highlight the illusion of data as an obdurate or secure object. Our goal was to address the myth that massive-scale data collection yields accessible data or usable archives. Trine's case could help us trouble the concept of data itself, the limits of digital preservation, and the precarious future of memory and heritage in a world of continually changing data storage and decoding formats.

Methodologically, the following weeks involved editing the narrative considerably, to find a few minutes in the video that we believed represented the heart of her story. We also played with various statements in Trine's narrative that were completely (or seemingly) unrelated to her memory of the German occupation of northern Denmark, to highlight the challenge of identifying relevance, not for viewers but in terms of the context of the lived experience of events in the 1940s and the later lived experience of recording a memory for future digital preservation.

None of our ideas included showing the video in a straightforward way. Although we had her consent, we considered that showing it in that way could not do justice to her story. We kept this ethical question on the table, iteratively discussing the impact of altering and retelling her story for our own ends—that is, presenting her face and voice to elicit an affective response from people in an entirely different context than her original contribution. Part of this discussion involved flipping the ethics discussion to the other side, whereby we acknowledged the potential positive impact of glitching Trine's memory. After all, our experiment was intended as a critical commentary for the public to see how "accurate" or "complete" data preservation is impossible, for many reasons potentially beyond the control of any single stakeholder.

A few weeks and conversations later, one of the authors contacted Trine and discussed our interest in her story. She was open and interested in the questions and curious about what our next step would be. We met with her two more times and, with her consent, started developing an art installation that would experiment with what algorithms had to say about her memory.

MEMORY GLITCH: EXPERIMENTING WITH ALGORITHMIC MEMORY-MAKING

The installation, entitled *Memory Glitch*, included three flat-panel displays, which were placed sequentially in the corridor of a public cultural center in Aarhus, as part of the *Affects, Interfaces, Events* conference. The screens present (retell, remix) Trine's story as seen from an algorithmic perspective. We show some still images below from the sequence of screens: *Memory Glitch 1* (**Figure 1**), *Memory Glitch 2* (**Figure 2**), and *Memory Glitch 3* (**Figure 3**). All images are reproduced here with Trine's written and verbal informed consent.

The first screen uses an automated transcription algorithm from Google's machine learning API. Voice is recognized through a series of mathematical operations through a series of deep-learning neural network algorithms. A continuous sampling of sound waves and comparison to thousands of other wave forms produces words, displayed if they meet the defined confidence value. Trine's story thus becomes text, synced to her voice, including all mistranslations and errors.

As viewers listened to her voice and watched the live transcription on the screen, they could begin to see how the text was not a seamless transcription of the audio. It was, at times, difficult to comprehend. The transcribed words were enlarged and flashed on the screen in a sequence that sometimes—but not always—matched her spoken words. Trine's story thus acquired different scales.

Memory Glitch 2 was an interactive screen. We used a Kinect infrared camera to calculate presence and movements in the corridor around the video. As the viewer moved closer to the video, the pixelation of the image increased. Thus, if viewers wanted to get closer to the video to see the picture or hear the sound more clearly, the fidelity of the visual information would be lost. We conceived this piece to demonstrate the inherently fraught experience of working through any digital archive, where one would encounter the impossibility of truly grasping a full picture. Indeed, as one of the onsite curators noted, as the viewers

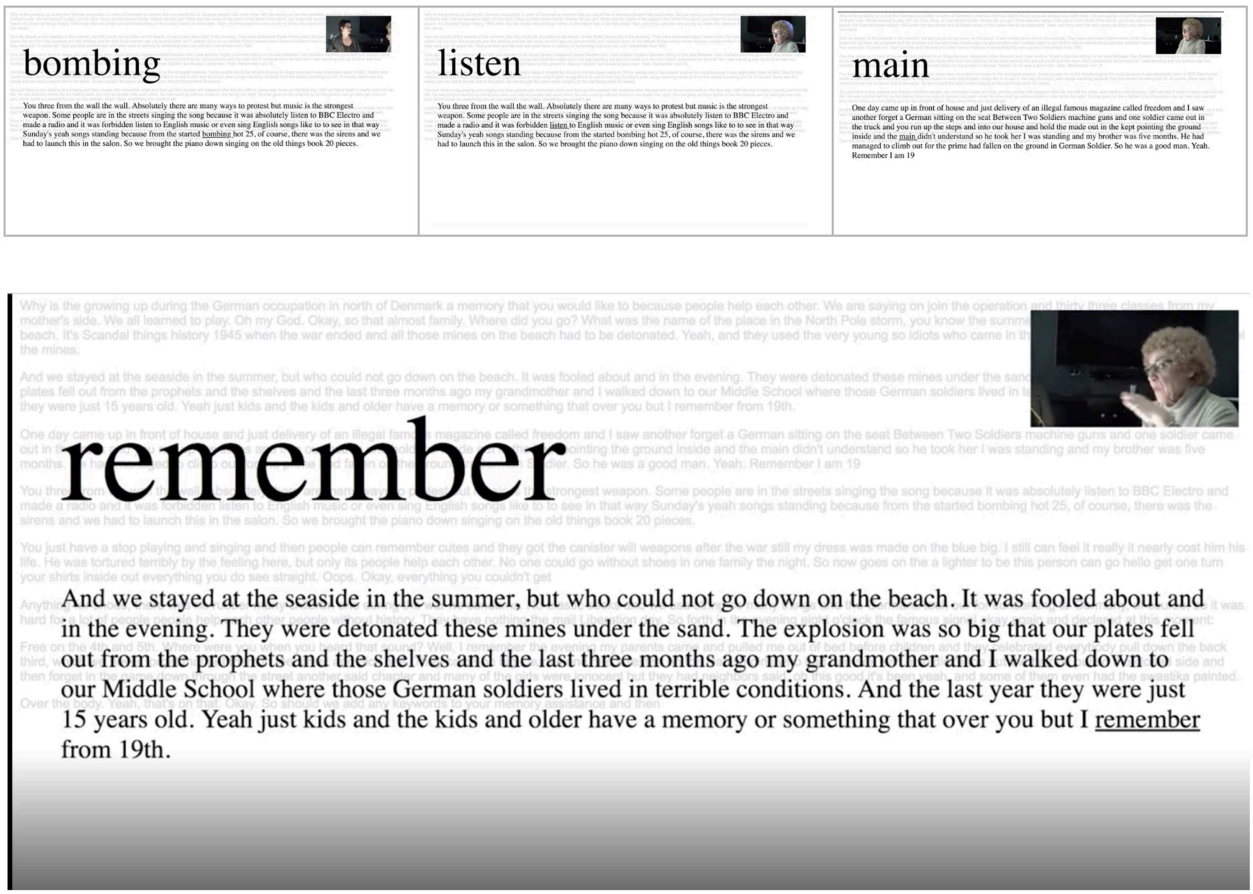

FIGURE 1 | Still shots from video demonstrate output of Google Cloud Speech, v1p1beta1, extended video model. Transcript is aligned with Gentle and displayed with active word and paragraph in sync with video. Fourth image shows closer view.

FIGURE 2 | Four progressive snapshot images of the video as the viewer walks closer to the screen, which is suspended from the wall in the exhibit space. A Kinect infrared camera is used to transform the image through pixilation as the viewer moves toward or away from the screen.

move closer, the image shifts from a representation of Trine to a representation of the viewers themselves. The camera's infrared sensors, pointed at the viewers, use their body heat to glitch the video.

The final screen in the series, *Memory Glitch 3*, was produced by a Machine Learning algorithm (OpenPose) that detects body keypoints in the image. We used the algorithm to mark the hand movements in the video, thus visualizing gestures. As Trine talks,

FIGURE 3 | Still images of three variations of code to produce video clip that focuses only on hand movements of participants in the interview. Dots represent presence of a hand in the visual field. Predictive analytics anticipates where the hand will move and produces clusters of dots that lighten in color as other motion is detected, illustrating motion (version 1) or, as the target of the predictive algorithm is refined, discrete lines that approximate fingers (version 3).

her hands express, emphasize, and capture an oblique perspective of her narrative. The coding we play with in this piece alternates from dots to lines, which have different visual effects. Both are responses to the predictive algorithm's analysis of where her hands will go.

As viewers watched the screen and listened to the conversation, dots and lines played across the screen, appearing and disappearing in seemingly random ways. These were her hands, moving across the screen. The vocal became gestural. But for most viewers, as we hoped, it was challenging to know what

was happening on this screen. Because the algorithm analyzed every frame of the video separately, Trine's gestures were combined at times with her interlocutor. Many viewers asked the researchers, functioning as museum curators at the exhibition, "What is this video supposed to be showing?" For them, the content was obscured. This engenders a double poignancy for us as designers and data scientists who created this rendering. On the one hand, we could feel both the loss of fidelity to the original video story and the emerging beauty of the dancing lines and dots. On the other, we knew that for Trine and possibly others trying to access this as a "clear" representation of her story, it was impenetrable.

Showing three versions of the same video—none of which gives the viewer a logical tale, effectively challenges any simple notion of both the memory and the process of remembering or forgetting. In these videos, the aesthetics, the context, as well as the algorithms transform the original data—already itself an abstraction from the lived experience—into something different. Once memories are put into the world, much like data, they're at risk of being lost, because they have been transformed. To push this reflection into considerations about the datafication process behind the reconfiguring of memory, we offered additional questions in written curator notes alongside the exhibit: What is the relation of the person to the algorithm, vis a vis their personal memory or memory making? What do our personal archives look like when they become data? How do automated processes influence and govern not only what we remember, but what we eventually will see when we try to access a digital memory? What are the ethics of these transformative processes of data science?

CRITICAL REFLECTIONS ON THE ALGORITHMIC RENDERING OF MEANING AND MEMORY

This *Memory Glitch* experiment examines how data are not just made but will continually transform throughout their lifespans. The starting point for algorithmic processing is the creation of the data object. As critical algorithm scholars focus on the complications of the algorithmic in machine learning processes, we cannot leave aside or forget the matter of where the data itself originates and how its transformation from lived experience to a computational form is an alteration from the untidiness of everyday life (Cheney-Lippold, 2017) into a measurable unit of cultural information, "flattened and equalized" (Markham, 2013) so it can be made comparable with units of cultural information from other instances and contexts.

How might we learn more about what is being condensed or flattened by reversing this process? How might we conceptualize data as lived experience (as well as within lived experience)? The expression 'lived experience' has been of particular interest to ethnographers and phenomenologists (cf. van Manen, 1990). Here, we use the term as it has become more colloquially understood, to refer to the whole of sensory and experiential being-in-the-world. In terms of digital and social media use, or the use of platforms to engage in communication and interactions or build/maintain social relations, lived experience also references how this is accomplished with and through digital media on an everyday level, which complicates how we might think about various sensory/physical, affective/emotional, and cognitive processes and modalities. While we don't rehearse the longstanding theoretical discussions around this complication (which has been covered too extensively to even summarize here), we do want to emphasize that anything we might call "memory" is only and always embedded in, created by, and experienced through lived experience.[3]

Our exploration of Trine's video required us to understand how memory was being reduced, simplified into units that would be recombined later, a process we now simply call datafication. We start with the classic idea, made again popular in the edited collection by Gitelman (2013), that data is always already cooked, meaning both that it has been generated according to human values and decisions and also that it only exists because it has been abstracted—or artificially severed from—the context in which it originated. Once objectified, the data is compiled with other units of cultural information, which enables us to do certain things with it, or think certain things because of a larger scale analysis.

Reconnecting the data to the person was an essential step for us to recognize what the original disconnect may have done (or may be doing) to the lived experience that led to the construction of the data form itself. What decontextualization occurs and with what possible consequence?

Looking at Trine's story being transcribed in *Memory Glitch 1*, for example, we start to see—especially through the transcription errors—the importance of her accent, the inflection of her words, among other nuances that Google's transcription services fail to notice. Likewise, in *Memory Glitch 3*, as Trine's hand gestures are highlighted by foregrounding them as data points flowing across the screen, other elements of the situation are blurred. If we focus only on the verbal content of her story, the emphasis and urgency of her telling is erased. The cultural, affective, lived experience of Trine becomes visible through those transformations because they never completely represent what we would expect.

Shifting this point slightly, once we reconnect the data object to the body, story, and person of Trine, we begin to see the flaws in both the data form and the code used to decipher and re-present it. This becomes particularly poignant in *Memory Glitch 2*, where the presence and movements of the viewer directly changes the way the data is decoded in visible form. The observer, archivist, or data archeologist can watch how their body heat functions as an algorithmic layer, overlaying new instructions, effectively obscuring previous instructions, generating a visual that changes as the viewer's body changes. The memory Trine imagines she preserved in digital form morphs again and again into a funhouse mirror image of the body literally viewing it.

In all three video glitches, the boundary we may at one point in time draw to demarcate what counts as the relevant data object

[3] At the same time, *Memory Glitch 1, 2, 3* compels us to flip this idea to consider that anything we call data is also memory. Literally, when data occupies space in a computer, it is called memory. But the computational concept of memory is that it is useful information for performing, and more importantly, recalling and repeating certain operations.

might be redrawn entirely differently at some unknown later point in time, when some other aspect of the recording becomes salient. Multiple elements are plausible markers of relevance—words spoken, geotags in the metadata, hand gestures, or the interviewer's critical remarks about the current political party. The missing element will (arguably always) be the meaning in the moment of the retelling.

In these three video renderings, we illustrate only parts of what is presumed to be a whole. And through this partiality, we both recognize and emphasize that a full memory could never be actualized. At such a point, a different formula would be applied to the record to draw a boundary around a different object to call it "data."

In this analysis, data objects—when severed from their contexts with all the associated affective connections—add (yet) another level of abstraction from the lived experience, even as they represent essential elements of the lived experience. This not only reiterates Boyd and Crawford's (2012) point that "taken out of context, Big Data loses its meaning," but also goes a step further in identifying how this process takes place, and how it happens when it happens. "Contexting" is the term used by Asdal and Moser (2012) to discuss how humans construct contexts continuously and experimentally, by which certain things are taken as explanatory contexts for others, and these processes are quite variable and political. Certainly this is what we are emphasizing when we foreground the context originating in the datafication and simultaneously remove or relegate to the backstage the multiple contexts preceding this datafication—those involving Trine's lived experience, followed by her donating her story as a memory we should not forget, followed by our repeated viewing and discussion about this video in our research team, and so forth. This analytical move is useful in that it juxtaposes different contexts, as well as different possible futures, confronting the contemporary "taken-for-grantedness" of data, which presents an imaginary of data analysis as impersonal, apolitical, and, because it is—or claims to be—aggregated and anonymized, separated from its origins and effects on human bodies.

At the same time, this exercise helps us see how any human or algorithmic codec will reconstruct a memory based on a particular set of constraints. This is not only a computational but a distinctly human issue, whereby facts are always after the fact, a matter of retrospective sensemaking (cf. Weick, 1969). In this double hermeneutic loop, we recognize how all forms of algorithmic sensemaking involve manipulation of data and transformation of meaning.

One way to specify the calculus used to make decisions at the level of encoding as well as decoding is to separate the algorithm from the algorithmic. An algorithm is generally considered machinic (vs. human) and in computer science traditions is an "abstract, formalized descriptions of a computational procedure" (Dourish, 2016). More broadly, as Cheney-Lippold (2011) notes, algorithms function as inference systems. In the latter conceptualization, what an algorithm does is more important than what it is, a point well-articulated by Gillespie's (2014) idea that algorithms generate or facilitate particular "knowledge logics." This emphasizes the work algorithms do. As Gillespie adds in 2016, "What makes something algorithmic is that it is produced by or related to an information system committed (both functionally and ideologically) to the computational generation of knowledge or decisions" (p. 25–26). The algorithmic intervenes in terms of step by step procedures. These procedures are formalized and automated. In computational settings, this automation helps the algorithm work "instantly, repetitively, and across many contexts, away from the guiding hand of its implementers" (Gillespie, 2016, p. 26). The process, which involves many stakeholders and systems beyond just the algorithm, builds possibilities for particular futures while simultaneously limiting other options. To return to the point made earlier about the difficulty of identifying agency in this process, Markham et al. (2018) conclude that "We can call this complication of locating moral agency and responsibility a wicked problem. There are no straightforward boundaries, definitions, or answers. Rather, there are only questions to be continually addressed" (p. 6). What our analysis helps us see is that this difficulty stems from our understanding that whatever functions algorithmically is not embedded in a location or element, but in relations (Magalhães, 2018). It is not an object or thing, but a set of process with/in contexts (Seaver, 2015; see also Dourish, 2004).

MEMORY, ETHICS, AND FUTURE-MAKING

In *Memory Glitch,* we link the algorithmic to the process of making data. These decisions are quite often hidden within the features and affordances of digital services themselves. Apple Memories and Google Photos are powerful tools, helping us store and organize, remember or forget. The problem is that for users, as well as these companies, "remembering" takes center stage, rather than the "forgetting," what is left out, or what will be omitted in future renderings. In *Memory Glitch,* we used three different predictive data models to classify, in different ways, Trine's experience. As the algorithms used their own pre-made (limited) categories, her experience was flattened (and/or expanded)—retrofitted into the logic outlined by the data models. Rouvroy (2013) would go as far as to say that "the subjective singularities of individuals, their personal psychological motivations or intentions do not matter. What matters is the possibility to link any trivial information or data left behind or voluntarily disclosed by individuals with other data gathered in heterogeneous contexts and establish statistically meaningful correlations" (p. 11–12). Trine's embodied presence and memory is replaced by her "statistical body," which ultimately functions as "de-territorialized signals, inducing reflex responses in computer systems, rather than as signs carrying meanings and requiring interpretation" (Rouvroy, 2013, p. 4).

Memory does not exist unproblematically (if at all) in the data traces we leave. Of course, even as we say this, we recognize that these traces of data carry the potentialities of remembering. We're not arguing that there is no value in these different renderings of memory, and the different futures they produce. We're suggesting, instead, that memory can't be contained by an

artifact because it is always in the relations, in the connections, in the process. And because the memory is always different than the object of the memory.

This is a different approach toward data ethics than the one taken by Metcalf and Crawford (2016), who analyze research practices in data science and make the argument that researchers often "represent themselves as dealing with systems and math, not people—human data is treated as a substrate for testing systems, not the object of interest in itself." (p. 3) Metcalf and Crawford help researchers think about the origins of data by positing "data are people," which may help protect the persons who (often unwittingly) participate in big data experiments. Our questions turn in a different direction. Through the video installation we are trying to direct attention to a different level of impact, whereby we're not as focused on the typical ethics question of whether or not we are harming people through various forms of data collection or analysis, but rather on: *what possible futures are being enabled or disabled*?

We're also not asking what ethical or moral principle is being used in different moments or by various stakeholders in the data science processes of data archiving and digital preservation, but rather: What sort of ethic is being *produced*? Markham (2015) reminds us that any creation of a data object constitutes a choice about what counts as data and what is discarded as non-relevant. In this action, we're building the ethics of the future. When the creation of a data object generates or attends to only certain elements of experience, to what extent has this already manipulated lived experience? Or are we simply manipulating the representation of lived experience: its memory, future, etc?

In Trine's case, she wanted her experiences of WWII to be remembered so these memories could create a better world, where people remember the atrocities of the war and respect and help each other. But once this memory is datafied, her desire about what this data means, or how it should be interpreted by future viewers/listeners/readers, is separated from the objects that are retained. Once the decoder ring—the sensemaking logic—is detached, meaning becomes a floating signifier, up for grabs. To draw on Theresa Senft's (2008, p. 46) apt turn of phrase, the notion of "the grab" is evocative because it emphasizes how anything we take to be real—in a world of digital/data objects and endless copy/paste possibilities–is the outcome, not of gazing, but grabbing. As she says:

> To grab means to grasp, to seize for a moment, to capture (an object, attention), and perhaps most significant: to leave open for interpretation, as in the saying "up for grabs." What is grabbed, like a screenshot, is just that, a moment frozen in time for inspection. The material, affective, embodied, lived part of this is never singular or just a 3D version of the screenshot. What is seen indicates what is not seen. Accidental or intentional, the grab still has impact. And has an ethic (Senft, 2018).

In *Memory Glitch 1, 2, 3,* a confluence of entities, processes, and decisions create a momentary stillness. To be sure, the case of Trine's memory being transformed or reconfigured is common. It depicts the almost by now banal disconnect between what people expect their digital archive to be and what actually is available and rendered over time. Yet when the exact same dataset is presented in multiple transmogrified forms that each tell a different story, this set of videos creates a moment for reflection. Viewers and developers alike can consider the potential violence (Hoffmann, 2016) of automated machinic processes on people whose memories are impacted. On the flipside, they can also imagine their role as an interactant with the algorithm as an active, if mysterious partner, which Magalhães (2018) contends can lead to greater, not less ethical agency for everyday users.

This is a matter of impact. And a question about what kind of analysis and models do we want to produce, to generate a better set of future ethics? The models we construct through data analytics cannot be separated from the futures they build. Focusing critical ethical attention on future practices and technologies that may render historical meaning in unexpected ways can help data scientists, consumers, and companies understand the impossibility of mapping data to memory in a one to one fashion and identify various algorithmic agents in the process of digital memory making. Creative and artistic play with algorithmic possibilities, for everyday users, can build more nuanced considerations of what a future holds when we have interpersonal, intimate relationships with autonomous nonhuman entities that function on our behalf. What do these relations entail? And if, after understanding the impossibility of preserving memory as data, we still want to preserve memories in ways that give us a sense greater fidelity to the original lived experience, what sort of "digital decoder rings" should be included to help future viewers (try to and likely fail to) understand our contexts?

A critical data science, we argue, can use its strengths at building creative algorithmic processes to create interventions like ours that help reveal the potentiality for generating new meaning as memories are manipulated through automated systems. This can have both enabling and constraining potentiality.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

Memory Glitch 1 and Memory Glitch 3 were coded by computational artist and MoRM researcher Robert Ochshorn, San Francisco. Memory Glitch 2 was coded by MoRM researcher GP and Digital Design Masters candidate Ann Derring, Aarhus. The team mentioned in this report included Justin Lacko, Andrew Sempere, Kseniia Kalugina, Ramona Dremljuga, AM, Ann Light, Morna O'Connor, GP, Anu Harju, Nathalia Novais, Sarah Schorr, Dalida Benfield, and Christopher Bratton. The video installation was designed and curated by AM, Morna O'Connor, Ann Light, Justin Lacko, and GP. This MoRM project is a part of the AUFF funded project *Creating Future Memories* at the Institute for Communication & Culture, Aarhus University.

REFERENCES

Asdal, K., and Moser, I. (2012). Experiments in context and contexting. *Sci. Technol. Hum. Values* 37, 291–306. doi: 10.1177/0162243912449749

Avram, G., Choi, J. H.-J., De Paoli, S., Light, A., Lyle, P., and Teli, M. (2019). Repositioning CoDesign in the age of platform capitalism: from sharing to caring. *CoDesign* 15, 185–191. doi: 10.1080/15710882.2019.1638063

Beer, D. (2019). *The Data Gaze: Capitalism, Power and Perception*. London: SAGE.

Boyd, D., and Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inform. Commun. Soc.* 15, 662–679. doi: 10.1080/1369118X.2012.678878

Bratton, C., Holford, E., Lacko, J., Sempere, A., Dremljuga, R., Benfield, D., et al. (2016). Creating future memories: a dialogue on process. *Disrup. J. Media Pract*. Available online at: http://journal.disruptivemedia.org.uk/creating-future-memories/

Cheney-Lippold, J. (2011). A new algorithmic identity. *Theory Cult. Soc.* 28, 164–181. doi: 10.1177/0263276411424420

Cheney-Lippold, J. (2017). *We Are Data: Algorithms and the Making of Our Digital Selves*. New York, NY: NYU Press.

Dourish, P. (2004). What we talk about when we talk about context. *Pers. Ubiquitous Comput.* 8, 19–30. doi: 10.1007/s00779-003-0253-8

Dourish, P. (2016). Algorithms and their others: algorithmic culture in context. *Big Data Soc.* 3, 1–11. doi: 10.1177/2053951716665128

Edwards, R., and Mauthner, M. (2002). "Ethics and feminist research: theory and practice," in *Ethics in Qualitative Research*, eds M. Mauthner, M. Birch, J. Jessop, and T. Miller (London: SAGE), 14–31.

Engster, D. (2005). Rethinking care theory: the practice of caring and the obligation to care. *Hypatia* 20, 50–74. doi: 10.1111/j.1527-2001.2005.tb00486.x

Gillespie, T. (2014). "The relevance of algorithms," in *Media Technologies: Essays on Communication, Materiality, and Society*, eds T. Gillespie, P. J. Boczkowski, and K. A. Foot (Cambridge: MIT Press), 167–194.

Gillespie, T. (2016). "Algorithm," in *Digital Keywords: A Vocabulary of Information Society and Culture*, ed B. Peters (Princeton, NJ: Princeton University Press), 18–30.

Gitelman, L. (ed.). (2013). *Raw Data Is an Oxymoron*. Boston: MIT Press.

Gregg, M. (2011). *Work's Intimacy*. Cambridge: Polity Press.

Hoffmann, A. L. (2016). Toward a conception of data violence. *Presented at Terms of Privacy: Intimacies, Exposures, Exceptions*. McGill University, Montreal, QC, Canada.

Hoffmann, A. L. (2018). "Data, technology, and gender: thinking about (and from) trans lives," in *Spaces for the Future: A Companion to Philosophy of Technology*, eds J. C. Pitt and A. Shew (New York, NY: Routledge), 3–28.

Iliadis, A., and Russo, F. (2016). Critical data studies: an introduction. *Big Data Soc.* 3, 1–7. doi: 10.1177/2053951716674238

Kennedy, H. (2016). *Post, Mine, Repeat: Social Media Data Mining Becomes Ordinary*. London: Palgrave Macmillan.

Kitchin, R., and Lauriault, T. P. (2014). *Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112 doi: 10.2307/j.ctt21h4z6m.6

Luka, M. E., and Millette, M. (2018). (Re)framing big data: activating situated knowledges and a feminist ethics of care in social media research. *Soc. Media Soc.* 4, 1–10. doi: 10.1177/2056305118768297

Magalhães, J. C. (2018). Do algorithms shape character? Considering algorithmic ethical subjectivation. *Soc. Media Soc.* 4, 1–10. doi: 10.1177/2056305118768301

Markham, A. N. (2013). Undermining 'data': a critical examination of a core term in scientific inquiry. First *Monday* 18. doi: 10.5210/fm.v18i10.4868

Markham, A. N. (2015). "Produsing ethics [for the digital near-future]," in *Produsing Theory in a Digital World 2.0*, ed R. A. Lind (New York, NY: Peter Lang), 247–256.

Markham, A. N. (2019). Taking data literacy to the streets: critical pedagogy in the public sphere. *Qualitative Inquiry* 1–11. doi: 10.1177/1077800419859024

Markham, A. N., Tiidenberg, K., and Herman, A. (2018). Ethics as methods: doing ethics in the era of big data research—introduction. *Soc. Media Soc.* 4, 1–9. doi: 10.1177/2056305118784502

Martin, K. (2018). Ethical implications and accountability of algorithms. *J. Bus. Ethics* 1–16. doi: 10.1007/s10551-018-3921-3

Metcalf, J., and Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data Soc.* 3, 1–14. doi: 10.1177/2053951716650211

Rehder, M. M., and Ostrowski, K. (2017). "MoRM and future memories: collaboratively investigating a contrapuntal museum as a participatory research approach," in *The Power of Play: Voices From the Play Community*, ed M. Poulsen (Aarhus: Counterplay), 144–155.

Rouvroy, A. (2013). "The end (s) of critique: data behaviourism versus due process," in *Privacy, Due Process and the Computational Turn*, eds M. Hildebrandt and K. de Vries (New York, NY: Routledge), 157–182.

Seaver, N. (2015). The nice thing about context is that everyone has it. *Media Cult. Soc.* 37, 1101–1109. doi: 10.1177/0163443715594102

Senft, T. M. (2008). *Camgirls: Celebrity and Community in the Age of Social Networks*. New York, NY: Peter Lang.

Senft, T. M. (2018). Personal conversation with Annette Markham. Skype.

van Manen, M. (1990). *Researching Lived Experience*. London: SUNY Press.

Weick, K. E. (1969). *The Social Psychology of Organizing*. New York, NY: McGraw-Hill

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership