Tracking the Digital Traces of Russian Trolls: Distinguishing the Roles and Strategy of Trolls On Twitter

Dongwoo Kim [1] and **Timothy Graham** [21] and **Zimin Wan** [1] and **Marian-Andrei Rizoiu** [31]
[1]The Australian National University, [2]Queensland University of Technology,
[3]University of Technology Sydney

Abstract

Online trolling has raised serious concerns about manipulating public opinion and exacerbating political divides among social media users. In this work, we analyse the role and behaviour of Russian trolls on Twitter through the lens of the social theory inspired by Tarde's ancient theory of monadology and its further development in Actor-Network Theory. Based on what the theory suggests, the social role or identity of individuals should be defined based on the traces they leave behind, which we define as a succession of timestamped items (e.g. tweet texts). To operationalise this, we develop a novel variant of text distance metric, *time-sensitive semantic edit distance*, accounting for temporal context across multiple traces. The novel metric allows us to classify roles of trolls based on their traces, in this case tweets, against a reference set. Through experiments aiming to identify the role of trolls into one of left-leaning, right-leaning, and news feed categories, we show the effectiveness of the proposed metric to measure differences between tweets in a temporal context. Through the qualitative analysis of tweet visualisation using the semantic edit distance, we find: (1) cooperation between different types of trolls despite having different and distinguishable roles; (2) complex and orchestrated interplay between left- and right-leaning trolls; (3) multiple agendas of news feed trolls to promote social disturbance and make Western democracy more fractious; and (4) a distinct shift of trolling strategy from before to after the 2016 US election.

1 Introduction

Over the past decade, social media platforms such as Twitter and Reddit have exploded in popularity, offering a public forum for millions of people to interact with each other. However, a novel problem also emerged: the increasing prevalence and influence of *trolls* and *social bots* in social media. Online trolls are predominantly human or hybrid (semi-automated) user accounts who behave in a deceptive, destructive, and/or disruptive manner in a social setting on the Internet (Buckels, Trapnell, and Paulhus 2014). Social bots are largely automated systems that pose as humans, and which seek to influence human communication and manipulate public opinion at scale. Bots have recently caused controversy during the 2016 U.S. presidential election, when it was found that they were not only highly prevalent but also highly influential and ideologically driven (Bessi and Ferrara 2016; Rizoiu et al. 2018a; Kollanyi, Howard, and Woolley 2016). The troll and bot account sets are not necessarily mutually exclusive, and recent studies uncovered that the Russian interference during the 2016 U.S. Election involved a combination of both types of accounts (Badawy, Ferrara, and Lerman 2018) to weaponise social media, to spread state-sponsored propaganda and to destabilise foreign politics (Broniatowski et al. 2018; Flores-Saviaga, Keegan, and Savage 2018).

There are several challenges that arise when studying the actions of malicious actors such as trolls and bots. The first challenge concerns distinguishing between their different types (or roles) in order to understand their strategy. Current state-of-the-art approaches usually build large amounts of features, and they use supervised machine learning methods to detect whether a Twitter account exhibits similarity to the known characteristics of social bots (Davis et al. 2016), and use text mining and supervised classification methods to identify online trolls (Mihaylov, Georgiev, and Nakov 2015). However, such approaches have several drawbacks, including requiring access to (often private) user features, and periodic retraining of models to maintain up-to-date information (Mihaylov, Georgiev, and Nakov 2015). The question is **can we circumvent this arms race of distinguishing the roles of online trolls? can we develop social theory-grounded approaches which do not over-rely on machine learning algorithms?** The second challenge lies in analysing and understanding the strategy of trolls. While prior work has focused on troll detection, recent studies show the existence of sub-types of trolls simulating multiple political ideologies (Boatwright, Linvill, and Warren 2018; Badawy, Ferrara, and Lerman 2018; Zannettou et al. 2018; Stewart, Arif, and Starbird 2018). This suggests a sophisticated and coordinated interplay between the different types of troll behaviour to manipulate public opinion effectively. The question is therefore **how to understand the behaviour and the strategy of trolls over time, and analyse the interplay between different troll sub-roles?**

This paper addresses the above challenges using a publicly available dataset of Russian troll activity on Twitter, published by Boatwright, Linvill, and Warren (2018), consisting of nearly 3 million tweets from 2,848 Twitter handles associated with the Internet Research Agency – a Russian "troll factory" – between February 2012 and May 2018.

To address the first challenge, we build upon recent developments in social theory, which combine Actor-Network Theory (ANT) with the 19^{th} century work of Gabriel Tarde (Latour et al. 2012; Latour 2002; Tarde 2011). In a nutshell, Latour et al. (2012) argue that actors in a network (human and non-human) are defined by the digital traces they leave behind, and that we can replace the idea of social role and identity in terms of individual attributes (e.g., ideology, gender, age, location, etc) with a definition of entities based on their traces. On Twitter, a trace of an individual is the sequence of their authored tweets. We tackle the problem of distinguishing the roles and identities of Russian trolls on Twitter by operationalizing ANT and Tarde's theory of monadology (Tarde 2011): we measure the similarities between the sequence of texts they generate over time against a reference set. To measure the similarity between two texts, we propose the *time-sensitive semantic edit distance* (t-SED), a novel variant of edit distance adapted to natural language by embedding two factors: *word similarity* and *time sensitivity*. Word similarity modulates the cost of the edit operations (i.e. deletion, insertion and substitution) according to the similarity of the words involved in the operation. The time sensitivity is highest when the two tweets are concomitant (i.e. they were authored within a short time interval of each other), and addresses issues such as changing discussion topics and concept drift. We show that for the task of distinguishing troll roles based on ground truth labels (e.g., left versus right trolls), our method outperforms a logistic regression baseline learner by more than 36% (macro F1).

We address the second challenge by constructing a two-dimensional visualisation, in which we embed the traces of Twitter trolls and their similarities (measured using t-SED) by using t-SNE (Maaten and Hinton 2008). In addition to observing that the tweets emitted by trolls with similar roles cluster together – which was expected from the theory of monadology – we make several new and important findings: (1) despite trolls having different and distinguishable roles, they worked together: the right trolls impersonate a homogeneous conservative identity, while the left trolls surround the conversation from all sides, with messages that simultaneously divide the Democrat voters on key issues and complement the destabilisation strategies of the right trolls; (2) we observe clusters of complex and orchestrated interplay between left and right trolls, both attempting to co-opt and strategically utilise ethnic identities and racial politics (e.g. *Black Lives Matter*), as well as religious beliefs; (3) the news feed trolls (i.e. impersonating news aggregators) are observed to have multiple agendas and sub-roles, such as clusters that disproportionately tweet about news of violence and civil unrest to create an atmosphere of fear, and other clusters that exhibit politically-biased reporting of federal politics; (4) there is an obvious shift of strategy between before and after the elections, with less distinguishable roles and strategies of trolls after the elections.

The main contributions of this work include:

- We introduce a **sociologically-grounded classification framework to identify the role of trolls** through operationalising the theory of monadology based on the accumulation of social traces rather than user features;
- We develop a **time-sensitive semantic edit distance** to measure similarity between two elements that form part of traces (i.e. tweets), embedding two components of online natural language: word similarity and item concomitance;
- We propose **a visualisation based on the novel distance metric** to gain insight on the strategic behaviour of trolls before and after the 2016 U.S. presidential election.

2 Background and related work

We structure prior work into two parts. Section 2.1 briefly introduces the monad theoretical framework for measuring social roles using social media data, and Section 2.2 presents some related work on social media trolls and bots.

2.1 Quantifying social roles using online traces

A new (old) view of society. Gabriel Tarde's ancient theory of monadology (Tarde 2011) has recently been adapted into the body of social theory known as Actor-Network Theory (ANT). It promises a powerful framework for the study of identity and social change in heterogeneous networks (Latour et al. 2012). In the 19^{th} century, Tarde's ideas proved not only difficult to conceptualise but even more difficult to operationalise due to a lack of data. It is perhaps for this reason that his alternative approach to describing social processes was not empirically testable and subsequently relegated to a footnote in history.

However, Latour et al. (2012) argue that the onset of the information age and the availability of digital data sets make it possible to revisit Tarde's ideas and render them operational. By examining the digital traces left behind by actors in a network (human and non-human), Latour et al. (2012: 598) argue that we can 'slowly learn about what an entity "is" by adding more and more items to its profile'. The radical conclusion is that datasets 'allow entities to be individualised by the never-ending list of particulars that make them up' (Latour et al. 2012: 600). Hence, a *monad* is a 'point of view, or, more exactly, a type of navigation that composes an entity through other entities' (Latour et al. 2012: 600).

As an example of this form of analysis, Latour et al. (2012) use the example of looking up an academic named 'Herve C.' on the web to show how collecting information through various digital sources results in the assemblage of a network that defines an actor's identity. As the authors argue: 'The set of attributes - the network - may now be grasped as an envelope - the actor - that encapsulates its content in one shorthand notation' (Latour et al. 2012: 593). Instead of atomic nodes (micro) that somehow 'enter into' or 'end up forming' structures (macro), we have a very different view of identity: in order to know what something is and understand its role in society, we simply follow the traces that it leaves behind through *its relations to other entities*, or in other words we trace its monad:

> If for instance we look on the web for the curriculum vitae of a scholar we have never heard of before, we will stumble on a list of items that are at first vague. Let's say that we have been just told that 'Herve C.' is

> now 'professor of economics at Paris School of Management'. At the start of the search it is nothing more than a proper name. Then, we learn that he has a 'PhD from Penn University', 'has written on voting patterns among corporate stake holders','has demonstrated a theorem on the irrationality of aggregation', etc. If we go on through the list of attributes, the definition will expand until paradoxically it will narrow down to a more and more particular instance. Very quickly, just as in the kid game of Q and A, we will zero in on one name and one name only, for the unique solution: 'Herve C.'. Who is this actor? Answer: this network. What was at first a meaningless string of words with no content, a mere dot, now possesses a content, an interior, that is, a network summarised by one now fully specified proper name (Latour et al., 2012: 592).

2.2 Political trolls on social media

Bots and trolls in political discussions. In recent years online trolls and social bots have attracted considerable scholarly attention. Online trolls tend to be either human, 'sock puppets' controlled by humans (Kumar et al. 2017), or semi-automated accounts that provoke and draw people into arguments or simply occupy their attention (Herring et al. 2002) for amplify particular messages and manipulate discussions (Broniatowski et al. 2018; Badawy, Ferrara, and Lerman 2018). Recent studies have investigated the impact of trolls and bots in social media to influence political discussions (Bessi and Ferrara 2016), spread fake news (Shao et al. 2017), and affect the finance and stock market (Ferrara et al. 2016). Especially, in a political context, studies have shown that online trolls mobilised support for Donald Trump's 2016 U.S. Presidential campaign (Flores-Saviaga, Keegan, and Savage 2018), and, of particular interest to this paper, were weaponised as tools of foreign interference by Russia during and after the 2016 U.S. election (Boatwright, Linvill, and Warren 2018; Zannettou et al. 2018). It is the current understanding that Russian trolls successfully amplified a largely pro-Trump, conservative political agenda during the 2016 U.S. Election, and managed to attract both bots and predominantly conservative human Twitter users as 'spreaders' of their content (Badawy, Ferrara, and Lerman 2018; Stewart, Arif, and Starbird 2018).

Detection and role of trolls. While trolls and bots have become increasingly prevalent and influential, methods to detect and analyse their role in social networks has also received wide attention (Cook et al. 2014; Davis et al. 2016; Varol et al. 2017). On the other hand, identifying and differentiating specific sub-groups or types of trolls poses a difficult challenge, which has attracted relatively less attention. Different types of trolls can be employed by specific individuals or groups to achieve specialised goals, such as trolls employed by the Internet Research Agency (IRA) to influence the political discourse and public sentiment in the United States (Boatwright, Linvill, and Warren 2018; Stewart, Arif, and Starbird 2018). The Clemson researchers (Boatwright, Linvill, and Warren 2018) used advanced tracking software of social media to collect tweets from a large number of accounts that Twitter has acknowledged as being related with the IRA. Using qualitative methods, the researchers identified five types of trolls, namely right troll, left troll, news feed, hashtag gamer, and fearmonger (Boatwright, Linvill, and Warren 2018). They found that each type of troll exhibited vastly different behaviour in terms of tweet content, reacted differently to external events, and had different patterns of activity frequency and volume over time (Boatwright, Linvill, and Warren 2018, p.10-11).

Relation to our work. Therefore we observe a close connection between *semantics* (what trolls talk about), *temporality* (when they are active and how hashtags are deployed over time), and the particular *roles* and strategies that drive their behaviour (e.g. right versus left troll). Our interest in this paper is to develop and evaluate a framework that can accurately identify roles of users within a population (in this case Russian troll types) by clustering them based not only on semantics but also temporality, that is, the order in which activities occur over time. To our knowledge this has not been achieved in previous work, where temporal information is often ignored or disregarded in analysis, such as the use of cosine distance and Levenshtein edit distance to differentiate sockpuppet types (Kumar et al. 2017), network-based methods to infer the political ideology of troll accounts (Badawy, Ferrara, and Lerman 2018), or word embeddings and hashtag networks with cosine distance as edge weights (Zannettou et al. 2018). Where temporal analysis of troll activity has been undertaken, the focus has been on tweet volume and hashtag frequency at different time points (Zannettou et al. 2018) rather than how roles and strategies change over time.

3 Russian trolls dataset

This study uses a publicly available dataset of verified Russian troll activity on Twitter, published by Clemson University researchers (Boatwright, Linvill, and Warren 2018)[1]. The complete Russian troll tweets dataset consists of nearly 3 million tweets from 2,848 Twitter handles associated with the Internet Research Agency, a Russian "troll factory". It is considered to be the most comprehensive empirical record of Russian troll activity on social media to date. The tweets in this dataset were posted between February 2012 and May 2018, most between 2015 and 2017.

In this work, we aim to distinguish trolls based solely on their authored text. We focus on the tweet content, and we try to detect the author category for each tweet (i.e. what type of Russian troll it was authored by), rather than detect the author category for each handle. There are multiple types of Russian trolls categorised by the Clemson researchers, namely: right troll; news feed; left troll; hashtag gamer; and fearmonger, sorted in decreasing order by tweet frequency. In this research, we will focus on the top 3 most frequent trolls: right troll, news feed and left troll. In addition, in this study we only consider English-language tweets, although future work can easily generalise our methods to any language expressed as Unicode. This amounts to a subset of 1,737,210 tweets emitted by 733 accounts.

[1] Available at https://github.com/fivethirtyeight/russian-troll-tweets/

According to Boatwright, Linvill, and Warren (2018), right trolls behave like "MAGA[2] Americans", they mimic typical Trump supporters and they are highly political in their tweet activity. On the other hand, left trolls characteristically attempt to divide the Democratic Party against itself and contribute to lower voter turnout. They achieve this by posing as mimic Black Lives Matter activists, expressing support for Bernie Sanders[3], and acting derisively towards Hillary Clinton. While tweets posted by left and right trolls have a strong political inclination, news feeds trolls tend to present themselves as legitimate local news aggregators with the goal of contributing to, and magnifying, public panic and disorder.

4 Classifying the roles of trolls

In this section, we develop a new method for identifying roles of social media users. In Section 4.1, we operationalise the theory of monadology for online social media users. Next, in Section 4.2, we introduce a new time-sensitive distance measure for texts and a majority voting framework for detecting roles. Finally, in Section 4.3 we introduce a semantic version of the edit distance.

4.1 Operationalising Tarde's monadology

As discussed in Section 2.1, Latour et al. (2012) broadly suggest that we can replace the idea of social role and identity in terms of individual attributes (e.g., ideology, gender, age, location, etc), and instead define entities based on the traces they leave behind. Here we define a trace as a succession of timestamped items (i.e. here the tweets of an author). Therefore two users are similar when their traces contain a large number of similar items (e.g. the tweets have words and hashtags in common). While measuring similarity between items (using off-the-shelf similarity measures) is promising, it does not preserve the order in which items occur. However, the ordering of tweets provides crucial information to define and understand the roles that users have in the social space. Social roles are not static - user actions and identities change over time, and they define them differently from one epoch to another.

In line with the monadological theory (Latour et al. 2012; Tarde 2011), we aim to distinguish the social roles of Russian trolls via their trace data, which consists of tweets over time. To tackle this problem, we assume that each individual tweet can be labelled by one of the target labels. For example, in our dataset (see Section 3), a tweet can be classified into one of three categories or roles: left troll, right troll, and news feed troll. We further assume that if a tweet is written by an author whose label is already known, i.e., in a training set, then we label the tweet according to the label of the author. While each individual tweet may not correctly reflect the role of an account (for example, some right trolls may advocate for the other side to create more confusion from time to time), when they are aggregated it should reflect the correct role of the account (right versus left ideology, or news feed troll). Or from a social theory point of view, we might provocatively say that 'the whole is always smaller than its parts' (Latour et al. 2012). When we say 'right' troll, for example, what actually underpins this ideological category is a large and complex landscape of tweets, which, if we zoom in further, contains an even larger and more complex set of words, hashtags, urls, and so on. There are not multiple 'levels' of reality (e.g. micro and macro) – instead, we have a one-level standpoint whereby actors (e.g. Russian trolls) are defined by their network (i.e. tweets, elements of tweets). As (Latour et al. 2012, p. 593) write: "This network is not a second level added to that of the individual, but exactly the same level differently deployed".

We operationalise this problem in terms of *time-sensitive similarity*: tweets that share similar word/hashtag co-occurrence patterns *and* were authored closely in time should be clustered closely together. This provides a basis to evaluate whether, and to what extent, the claims of the social theory have empirical validity and predictive power. Importantly, it also provides the opportunity to reevaluate the roles that Russian trolls occupied on Twitter during and after the 2016 U.S. Election, and whether we can learn new insights about their roles and strategies by anchoring our analysis to the monadological theory. However, first we need to devise a similarity metric that can take into account both semantic and temporal information, which we undertake in the following two subsections.

4.2 Time-sensitive metric for trace classification

Given that each tweet is labelled using one target label, our goal aims to distinguish types of accounts based on their tweets. Unlike a general supervised classification problem, where pairs of an input and label are provided as an example, our classification algorithm needs to predict a label from a set of timestamped text snippets authored by a target account. We formulate the user account classification as a majority voting problem, where the majority label of tweets is the predicted label of the corresponding account. The open question is how to classify the label of an individual tweet given a set of training traces (i.e. sequences of tweets) in terms of the time-sensitive similarity.

We employ k-nearest neighbour algorithm (KNN) to classify labels of tweets. A KNN classifies a data point based on the majority labels of k-nearest neighbour based on *distances* to the other labelled data points. Let $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ be two sequences of tokens, constructed by tokenising tweets i and j; let t_i and t_j be the timestamps of tweets i and j, respectively. We propose a time-sensitive distance metric between tweets i and j formulated as

$$D(i,j) = \mathrm{dist}(\boldsymbol{s}_i, \boldsymbol{s}_j) \times \exp(\theta|t_i - t_j|), \quad \theta > 0 \qquad (1)$$

where $\mathrm{dist}(\boldsymbol{s}_i, \boldsymbol{s}_j)$ measures a distance between tokenised sequences $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and the exponential term penalises large timestamp differences between the two tweets. This is required because sometimes seemingly similar text snippets

[2] MAGA is an acronym that stands for Make America Great Again. It was the election slogan used by Donald Trump during his election campaign in 2016, and has subsequently become a central theme of his presidency.

[3] Bernie Sanders was the alternative Democrat Presidential Nominee

may represent completely different concepts, because the meaning of the employed terms has evolved with time. For instance, the hashtag #MeToo had a general meaning prior to 15 October 2017, whereas afterwards the meaning of the same hashtag changed dramatically with the emergence of the #MeToo social movement on Twitter. By adopting the exponential penalisation inspired from point process theory (Leskovec et al. 2008; Rizoiu et al. 2018b), the KNN weights more towards temporally related tweets.

In general, the Euclidean distance metric is employed for the function $\text{dist}(\boldsymbol{s}_i, \boldsymbol{s}_j)$, when $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are defined in an Euclidean space. However, in our case, $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are sequences of word tokens for which the Euclidean distance is not defined. One may use a bag-of-words representation of tokens to map the sequence of tokens into a vector space, and then employ a text distance metric such as cosine distance. However, the bag-of-words representation loses the ordering of tokens, which may embed thematic concepts that cannot be understood from individual words. In the next section, we propose a new distance metric to compute the distance between two sequences of tokens while preserving the temporal ordering between tokens.

4.3 Semantic edit distance

We propose a new text distance metric $\text{dist}(\cdot,\cdot)$ to capture semantic distance between two sequence of symbols based on the edit distance (ED) metric. The ED is a method to quantify the distance between two symbolic sequences by calculating the minimum value of the required edit operations to transform one sequence into the other. There may be different sets of operations according to the definition of ED. The most common form of ED is known as Levenshtein distance (Navarro 2001). In Levenshtein's original definition, the edit operations include the insertion, deletion and substitution, and each of these operations has a unit cost. Therefore the original ED is equal to the minimum number of the required operations to transform one string into the other.

Formally, given two sequences $\boldsymbol{a} = a_1, a_2, ..., a_n$ and $\boldsymbol{b} = b_1, b_2, ..., b_m$, the edit distance is the minimum cost of editing operations required to transform $\boldsymbol{a}$ into $\boldsymbol{b}$ via three operations: (i) insert a single symbol into a string; (ii) delete a single symbol from a string and (iii) replace a single symbol of a string by another single symbol, associated with non-negative weight cost $w_{\text{ins}}(x)$, $w_{\text{del}}(x)$ and $w_{\text{sub}}(x, y)$, respectively. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be sequences of n and m symbols, respectively. The edit distance between $\boldsymbol{a}$ and $\boldsymbol{b}$ is given by $\text{ed}(n, m)$, defined recursively,

$$\begin{aligned} \text{ed}(i,0) &= \sum_{k=1}^{i} w_{\text{del}}(a_k) && \text{for } 1 \le i \le n \\ \text{ed}(0,j) &= \sum_{k=1}^{j} w_{\text{ins}}(b_k) && \text{for } 1 \le j \le m \end{aligned}$$

$$\text{ed}(i,j) = \begin{cases} \text{ed}(i-1, j-1) & \text{if } a_i = b_j \\ \min \begin{cases} \text{ed}(i-1,j) + w_{\text{del}}(a_i) \\ \text{ed}(i,j-1) + w_{\text{ins}}(b_j) \\ \text{ed}(i-1,j-1) + w_{\text{sub}}(a_i, b_j) \end{cases} \\ & \text{otherwise.} \end{cases}$$

In the original ED, each operation is often assumed to have a unit cost. The unit cost assumption is convenient to measure the distances, however, it does not reflect the similarity between different symbols. For example, given sentences $\boldsymbol{s}_1 =$ "I like music", $\boldsymbol{s}_2 =$ "I love music" and $\boldsymbol{s}_3 =$ "I hate music", $\text{ed}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \text{ed}(\boldsymbol{s}_2, \boldsymbol{s}_3) = 1$ if we assume word level edit distance where each symbol corresponds to a word token in a vocabulary. However, "love" and "like" have more similar meaning than "love" and "hate" or "like" and "hate". In this situation, we expect that the distance between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ should be less than the distance between $\boldsymbol{s}_2$ and $\boldsymbol{s}_3$.

As it turns out from the previous example, it is essential to understand and measure similarity between word symbols to further measure the distance between sentences. No canonical way exists to measure similarities between words, but it is often assumed that if words are used frequently in similar context, then the words play a similar role in sentences. To capture the context of words, we compute the co-occurrence statistics between words, which shows how often a certain word is used together with the other words. We first construct a co-occurrence matrix based on the number of times a pair of words is used in the same tweet. Then, to measure how frequently a pair of words are used in similar context, we compute the cosine similarity between two co-occurrence vector corresponding to the rows from the co-occurrence matrix. Therefore, the more two words are used frequently in a similar context provided by co-occurring words, the more similar they are based on the cosine similarity. From now on, we denote $\text{sim}(x, y)$ as the cosine similarity of word x and y from the co-occurrence matrix. Many previous studies have shown that the co-occurrence patterns can capture meaningful properties of words including synonymousness (Mikolov et al. 2013; Pennington, Socher, and Manning 2014).

We propose an edit distance endowed with novel cost functions of the three edit operations, named semantic edit distance (SED), using the word similarity. The more similar two sentences are, the fewer the editing operation should cost. In other words, the cost of operation equals to the dissimilarity of two words. Based on this intuition, we propose three cost functions for edit operations as follows:

$$w_{\text{del}}(a_i) = 1 - \text{sim}(a_i, a_{i-1}) \tag{2}$$

$$w_{\text{ins}}(b_i) = 1 - \text{sim}(b_i, b_{i-1}) \tag{3}$$

$$w_{\text{sub}}(a_i, b_j) = 1 - \text{sim}(a_i, b_j) \tag{4}$$

The intuitions behind each of cost function are

- For the deletion in Eq. (2) (or insertion in Eq. (3), resp.), if two consecutive symbols are similar, deleting (inserting) the latter one would not cost much, and the deletion (insertion) operation would have little influence on distance between two strings.

	Train	Validation	Test
Left troll	99	53	79
Right troll	188	106	155
News feed	20	11	22

Table 1: Number of accounts for each category used for training, validation, and testing. 50 tweets are sampled for each account in the classification experiments.

- For the substitution in Eq. (4), if the symbol a_i of sequence $\boldsymbol{a}$ is similar to the symbol b_j of sequence $\boldsymbol{b}$, the substitution should not cost much.

We denote $\mathrm{sed}(\cdot, \cdot)$ as $\mathrm{ed}(\cdot, \cdot)$ endowed with the above three operation costs. Finally, applying $\mathrm{sed}(\cdot, \cdot)$ to $\mathrm{dist}(\cdot, \cdot)$ in Eq. (1) results in a time-sensitive semantic edit distance, denoted as t-SED in the rest of this paper.

5 Evaluation of role prediction

In this section, we measure the classification performance of our proposed method on the Russian troll dataset.

Experimental settings. We use a subset of the Russian troll dataset consisting of users labelled as right trolls, left trolls and news feed, as described in Section 3. We split the accounts into 50% train, 20% validation, and 30% test datasets. The detail statistics of each dataset is described in Table 1. For each account, we randomly sample 50 tweets for ease of computation. We tokenise the text of tweets using a tweet pre-processing toolkit[4] and remove infrequent words which occur less than 3 times in the corpus. The co-occurrence matrix used to compute the word similarity is constructed from the entire dataset.

We test the proposed KNN approach with three distance measure (i.e. cosine distance[5], ED and SED), and their time-sensitive counterparts as defined in Eq. (1) (denoted as t-Cosine, t-ED and t-SED, respectively). Note that the SED ranges from zero to the maximum length of sentences, therefore SED depends on the lengths of sequences; short sentences are likely to have small SED. To investigate the effect of sequence length, we additionally propose and test two normalised SEDs: SED/Max[6] and SED/ED[7]. As a baseline, we also test a logistic regression classifier with and without temporal information (denoted as LR and t-LR). We train the LR models with bag-of-words features to classify the label of an individual tweet and predict account label based on majority vote. To add temporal information into the logistic regression, we compute the normalised timestamp of tweets and we add it into the feature set. Finally, the classification performance is measured by macro and micro F1. Note that the dataset shows highly skewed distribution toward the right troll accounts. Table 2 summarises all the tested approaches and their performances.

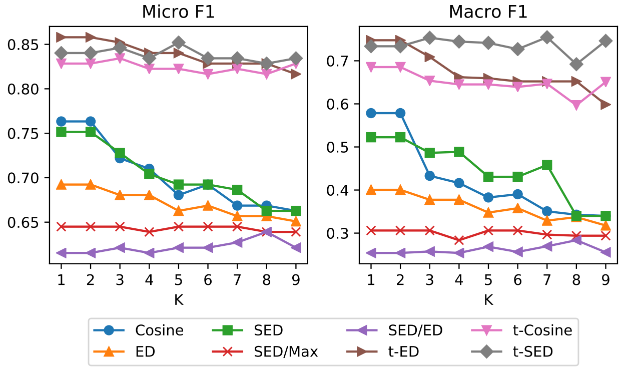


Figure 1: Macro and micro F1 scores on validation set with KNN. Cosine, ED, and SED performs the best when k is 1. t-SED shows relatively consistent performances over varying numbers of neighbours.

		Micro F1		Macro F1	
	Method	K	F1	K	F1
Baseline	LR	-	0.75	-	0.55
	ED	1	0.72	1	0.46
	Cosine	1	0.75	1	0.54
Semantic	SED	1	0.78	1	0.62
	SED/Max	1	0.65	1	0.35
	SED/ED	8	0.62	8	0.29
Temporal	t-LR	-	0.79	-	0.61
	t-ED	1	0.83	1	**0.75**
	t-Cosine	3	0.81	1	0.61
	t-SED	5	**0.84**	7	**0.75**

Table 2: Micro and macro F1 scores on test set along with the number of neighbours (K) for KNN. Although SED outperforms all baseline models, t-SED significantly outperform their non time-sensitive counterparts implying the importance of incorporating the temporal dimension.

Results. Fig. 1 shows the classification performances of different metrics on the validation dataset with varying number of neighbourhood size in KNN. Note that t-SED has additional parameter θ, which controls the exponential rate. We perform a grid search on θ to find the best configuration and report the best validation performance in Fig. 1. One interesting pattern from the validation set is that all other metrics except time sensitive metrics suffer from having a large number of neighbours, whereas the time sensitive metrics retain stable performances across the different number of neighbours. We conjecture that including more neighbours include more tweets from different time ranges, as done with the non time-sensitive metrics, eventually hurts the classification of individual tweets.

The results shown in Table 2 are the best performances of each approach over the range of k on the validation set. Overall, the t-SED outperforms all other metrics for both macro and micro F1 scores. We interpret the results from four perspectives: **1) The importance of word similarity** by comparing ED and SED. By adding the word similar-

[4] Available at https://github.com/s/preprocessor

[5] The bag-of-words is used to map a sequence to vector

[6] Length normalisation: $\mathrm{sed}(\boldsymbol{a}, \boldsymbol{b}) / \max(|\boldsymbol{a}|, |\boldsymbol{b}|)$

[7] Ratio normalisation: $\mathrm{sed}(\boldsymbol{a}, \boldsymbol{b}) / \mathrm{ed}(\boldsymbol{a}, \boldsymbol{b})$

ity into the cost function, SED can significantly outperform ED in the role prediction. **2) The importance of preserving order between tokens in tweets.** Although the naive ED performs worse than the cosine distance, SED outperforms cosine outlines the importance of preserving order between tokens, alongside with accounting for word similarity. **3) The importance of accounting for temporal information.** All time-sensitive measures outperform their non-temporal counterparts (e.g. t-LR vs. LR, or t-SED vs. SED). This result implies significant amounts of concept drifting in the dataset. We also note that t-ED and t-SED have similar performances, despite the performance difference between ED and SED. This seems to suggest that the temporal information is more important for the successful classification than the word similarity. However, the best value for k obtained on the validation set shows that t-SED can use a larger number of neighbours than t-ED, which implies the potential robustness of t-SED against t-ED by accounting wider context through the word similarity. **4) Circumvention of trainable models.** Given the most accessible features, the trace of text, t-SED outperforms training based models (i.e. LR and t-LR). This implies that distance-based methods can outperform training models in the most restricted scenario where only the authored text is available (without user attributes). It is also important to note that the distance-based methods do not need to be retrained, whereas the training based models, such as the logistic regression, require periodical retraining to model temporal changes in a corpus.

Note that the two normalised metrics, SED/Max and SED/ED, do not help increase the performance. We conjecture that a normalisation is unnecessary due to the limited number of characters can be written at a time[8].

6 Results and findings: the strategy of trolls

To understand the behaviour and the strategy of trolls over time, we visualise the tweets using t-SNE (Maaten and Hinton 2008), a technique originally designed to visualise high-dimensional data through their pairwise distances. Fig. 2 shows the visualisation of tweets from two different time ranges using SED and t-SED. Each dot in the figure represents a single tweet embedded into a two dimensional space constructed by t-SNE. To emphasise different behaviour of trolls over time, we plot tweets from two different time ranges: one before the presidential election (September 2016) in Fig. 2a and Fig. 2c, and another after the election (April 2017) in Fig. 2b and Fig. 2d.

6.1 Re-examination of Russian trolls

Together they troll. Although the Clemson University researchers argued that "with the exception of the fearmonger category, handles were consist and did not switch between categories" (Boatwright, Linvill, and Warren 2018, p. 6), the data-driven visualisation presented in Fig. 2 reveals a somewhat different picture. A cursory examination of the visualisation shows several instances where the positioning of tweets from both right trolls and left trolls overlap, suggesting that they tweet about similar or related topics. Similarly, Fig. 2a and Fig. 2c reveal sub-clusters of news feed trolls that are positioned within the predominantly right troll cluster. Therefore, we cannot observe a clear separation in authored tweets based their categories through the distance metrics. Nonetheless, many tweets are locally clustered in line with their categories, which ultimately helps us to correctly classify their type, as shown in Section 5.

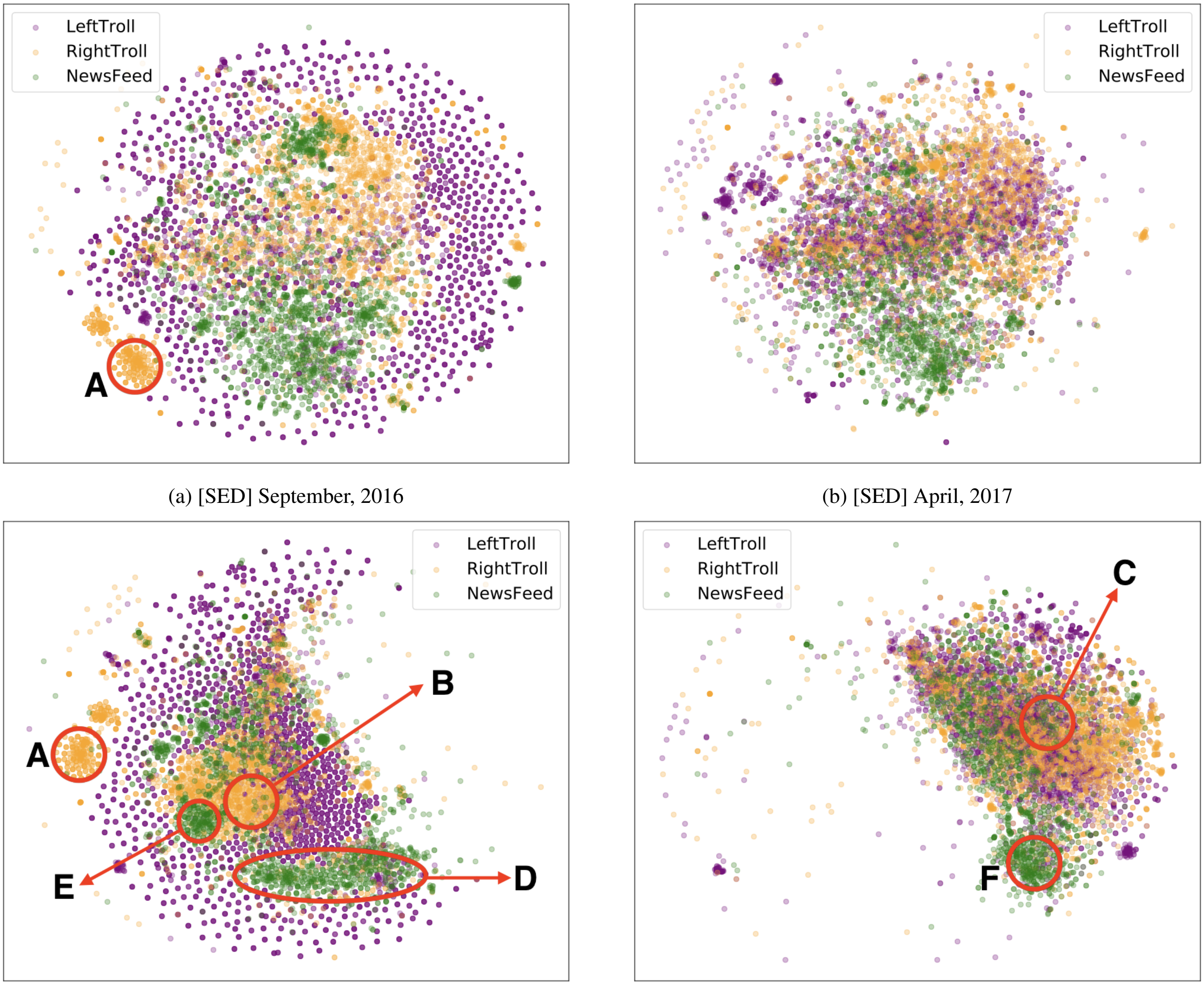
Right vs. left strategy. A notable pattern from Fig. 2 is that the tweets authored by the left trolls are spread across all regions whereas those of the right trolls are more focused and relatively well clustered. As discussed previously, we know that generally the right trolls were focused on supporting Trump and criticising mainstream Republicanism and moderate Republican politicians. Compared to left trolls, right trolls have a more singular or homogeneous identity, and employ common hashtags used by similar real Twitter users, including #tcot, #ccot, and #RedNationRising (Boatwright, Linvill, and Warren 2018, p. 7). On the other hand, left trolls have a more complex discursive strategy. As (Boatwright, Linvill, and Warren 2018) argue, these trolls send socially liberal messages, with an overwhelming focus on cultural identity. Accordingly, the position of the left trolls on the visualisation provides a stark picture of their complementary and supporting role in driving the IRA's agenda building campaign on Twitter. Left trolls are literally surrounding the conversation on all sides. In some areas they are attacking Hillary Clinton and mainstream Democratic politicians, in others they are mimicking activists from the Black Lives Matter movement, and in others discussing religious identity and Christian moralism. Left troll tweets are certainly distinguishable on the visualisation in terms of their position, but we can see how they play into, and strategically function alongside, the news feed and right trolls. To examine these observations in more detail we can zoom in to specific regions of interest within the visualisation to analyse the tweet content.

6.2 Left and right trolls worked together

Leveraging racial politics. Most of the tweets from the notable right troll cluster in Fig. 2a and Fig. 2c (**tag A** in the figures) contain the hashtag #ThingsMoreTrustedThanHillary, which shows the strategic behaviour of certain right trolls to make the Democratic candidate distrustful. This strategy is also part of the over-arching agenda of the left trolls, who not only undermined the trust in Hillary Clinton, but co-opted the Black Lives Matter movement and related topics to negatively impact her campaign. As (Boatwright, Linvill, and Warren 2018, p. 8) show, left trolls authored tweets such as "NO LIVES MATTER TO HILLARY CLINTON. ONLY VOTES MATTER TO HILLARY CLINTON" (@Blacktivists, October 31, 2016). Furthermore, the central cluster of right trolls (**tag B**) in Fig. 2c contains tweets that show the support of Trump from black people, in addition to tweets from typical Trump supporters. The following are example tweets from this cluster of right trolls:

- "Join Black Americans For Trump, "Trump is the best choice for ALL Americans!" Join Today at https://t.co/NJBoTamxDi #Blacks4Trump" (@marissaimstrong);

[8] Twitter has a 140-character limitation before Nov. 2017



(a) [SED] September, 2016

(b) [SED] April, 2017

(c) [t-SED] September, 2016

(d) [t-SED] April, 2017

Figure 2: Tweets from two different time ranges (left vs right column) are embedded into two-dimensional spaces, via t-SNE, with two variants of edit distances (SED for the top row, and t-SED for the bottom row). The locations are computed using the distances between pairwise distances between all tweets from the two time ranges – i.e. one space for SED and one for t-SED. We plot the tweets separately based on the period they were written. For SED, there is a relatively clear distinction between the different categories before the election, but they are relatively indistinguishable after the election. For t-SED, the gap between 2016 and 2017 is wider than for SED since the distance between tweets increases exponentially with their time difference.

- "Why I Support Donald Trump https://t.co/U0oT8odMOB #BlacksForTrump #Blacks4Trump #BlackLivesMatter #ImWithHer #DemExit #MAGA" (@hyddrox).

We therefore observe a complex interplay between left and right trolls, whereby both attempt to co-opt and strategically utilise racial politics and ethnic identity (Phinney 2000), even though they use different approaches. This resonates with and provides new insights on recent analysis of Russian troll communication on Twitter, where trolls were found to make a "calculated entry into domestic issues with the intent to polarise and destabilize" (Stewart, Arif, and Starbird 2018, p. 4).

Utilising religious beliefs. From the central part of Fig. 2d, we observe tweets from both ideologies and we find a cluster of conversations (**tag C**) related to personal religious beliefs, such as:

- "Just wait & #God will make things great" (@acejinev);
- "Each of us is a Masterpiece of God's Creation. Respect Life! #Prolife" (@j0hnlarsen).

Although the hashtags used in these religious tweets are often different for the left and right trolls, their similarity is captured by the metric. This reveals an interesting strategy whereby trolls from both left and right pretend to be ordinary American citizens who, although ideologically different, are united by shared religious beliefs. This indicates that not all trolls in a category acted unitarily, and the tweets they emitted cluster into groups corresponding to their different subroles and strategies. Using the proposed t-SED measure and visualisation, one can zoom in and gain richer insights into the strategies and identities of these user accounts (down to individual actions).

6.3 The multiple agendas of news feed trolls

When studying news feed trolls, we observe how different clusters promote specific themes of news articles. The slender cluster of news feed trolls (**tag D**) in Fig. 2c often contain the hashtag #news, and report incidences of violence and civil unrest. For example, tweets from this cluster include:

- "Warplanes hit Aleppo in heaviest attack in months, defy U.S. #news" (@specialaffair);
- "Pedestrian hit, killed by train in Mansfield https://t.co/kvmFEgf8Ps #news https://t.co/TXsol3YjgA" (@todaybostonma);
- "One person killed during violent Charlotte protest; officer hurt https://t.co/IYyg0xmf0L https://t.co/UbzzAeW3zR" (@baltimore0nline).

On the other hand, the small left-most cluster of news feed trolls (**tag E**) in Fig. 2c focus on the hashtag #politics, and have a focus on federal political issues and politicians as well as policy and regulation. Tweets from this cluster include, for example:

- "Obama Trade Setbacks Undercut Progress in Southeast Asian Ties #politics" (@newspeakdaily);
- "Is federal government trying to take down for-profit colleges? #politics" (@batonrougevoice);
- "Clinton takes aim at Trump supporters https://t.co/fxEox7N74Z #politics" (@kansasdailynews).

This suggests that the IRA strategy for news feed trolls comprised multiple agendas – they are not simply a homogeneous group. We observe that the clusters help to illuminate the within-group variation for this troll category, and we might speculate that the clusters correspond to, or at least highlight, the agenda-setting strategies that news feed trolls carried out, as well as their relationship to other types of trolls (i.e., by analysing their proximities in t-SNE space).

6.4 The shifting sands of troll identity over time

By analysing two different time ranges, we can notice that there is less separation between the tweets belonging to the different roles (see Fig. 2b and Fig. 2d). The clustering structure around September 2016 is comparatively clearer than the structure around April 2017. The difference suggests that, for some reason, the strategic behaviour of trolls changed markedly before and after the election. We are not aware of any previous research that has identified this, let alone that offers an explanation why. Interestingly, one strategy that appears to continue after the elections is seeding fear: the t-SED visualisation in Fig. 2d reveals a cluster of news feed trolls (**tag F**) with the particular focus on reporting negative news about shootings and gun violence, crime and fatalities. Example tweets from this cluster include:

- "Gunman shoots woman at Topeka Dollar General https://t.co/gQgQy8B0Hh" (@kansasdailynews);
- "Police: Father accidentally shoots son while fighting off intruder https://t.co/kAD9shfp7t" (@kansasdailynews);
- "Police: Suspect dead after woman, child abducted in Homewood https://t.co/TcTNmc5oFu" (@todaypittsburgh).

Although we do not have scope in this paper to undertake further analysis of the clustering after the election, it is obvious that t-SED (Fig. 2d) offers a different view of Russian troll strategies as compared to SED (Fig. 2b). Zooming out to the largest scale, we see generally that taking into account temporal information is important because it outlines the drift in topics of discussion.

7 Discussion

The troll is always smaller than its parts.

We return to the social theory discussed in Section 2.1 which underpins the framework set out in this paper. As we have shown, the visualisation and analysis presented in Section 6 affords a nuanced analysis of the *gradation* and heterogeneity of Russian troll identities, which are not as disjoint or homogeneous as previous work suggested. However, this is not to say that prevailing analytic categories of Russian trolls are insufficient or invalid – on the contrary, what we offer here builds upon and extends existing scholarship by contributing new empirical insights, methods, and theory.

The visualisation suggests a complex intermingling of discursive troll strategies that make up, and do the work of, the over-arching agenda of the IRA. Together, the Russian troll tweets form an actor-network that has been variously described as 'state-sponsored agenda building', 'Russian interference', 'election hacking', and 'political astro-turfing', among others. The titular claim of Latour et al. (2012) that 'the whole is always smaller than its parts' finds compelling empirical validation in our analysis. Thus one might ask, what is Russian interference on Twitter? Answer: *this* network of Russian troll roles (right, left, news feed, etc). Who/what are these Russian troll roles? *This* network of tweets. Who/what is this particular Russian troll tweet? *This* cluster of semantically and temporally similar tweets.

Hence, each time that we wish to pinpoint the identity of a Russian troll, we must look to its parts (in this case tweets, but also location, meta-data, etc); each time we want to pinpoint the meaning or identity of a tweet, we again looks to the parts (words, hashtags, author, etc) and to other tweets that have similar temporal-semantic elements. The traditional notion of micro versus macro, or individual component versus aggregate structure, can be largely bypassed. What matters are the similarities that pass between actors (in this case tweets by Russian trolls) from one time point to another *on the same level*. We can form an understanding of the social phenomenon through navigating and exploring the 2D plane in which the elements (in this case tweets) are arranged and visualised. Specifically we have examined this through quantifying the identities of Russian trolls by following and mapping similarities and differences in their trace data over time.

8 Conclusion

In this study, we address a new challenging problem: how to characterise online trolls and understand their tactics based on their roles. We focus on the different types of trolls identified in recent studies to picture a more detailed landscape of how Russian trolls attempted to manipulate public opinion and set the agenda during the 2016 U.S. Presidential Election. In order to do so, we propose a novel operationalisation of recently revisited Tarde's ancient social theory, which posits that individuals are defined by the traces of their actions over time. We define a novel text distance metric, called *time-sensitive semantic edit distance*, and we show the effectiveness of the new metric through the classification of Russian trolls according to ground-truth labels (left-leaning, right-leaning, and news feed trolls). The metric is then used to construct a novel visualisation for qualitative analysis of troll roles and strategies. We discover intriguing patterns in the similarities of traces that Russian trolls left behind via their tweets, providing unprecedented insights into Russian troll activity during and after the election.

Assumptions, limitations and future work. This work makes a number of simplifying assumptions, some of which can be addressed in future work. First, we assume that each tweet is assigned exactly one label, selected from the same set as the user labels. Future work will relax this assumption, allowing tweets to have multiple labels, possibly from a set disjoint from the user labels. Second, we measure the similarity between the traces of two users by measuring the similarity between tweets and performing a majority vote. Future work will introduce trace similarity metrics directly working on a trace level instead of using an aggregated approach. Third, we had to characterise the social theory in order to operationalise it as a formal method. In future work we will attend more closely to the nuances of Tarde's monadology and its development within ANT. Finally, we aim to construct and publish an interactive version of the visualisation in Fig. 2.

References

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. *arXiv preprint arXiv:1802.04291*.

Bessi, A., and Ferrara, E. 2016. Social bots distort the 2016 u.s. presidential election online discussion. *First Monday* 21(11).

Boatwright, B. C.; Linvill, D. L.; and Warren, P. L. 2018. Troll factories: The internet research agency and state-sponsored agenda building. *Resource Centre on Media Freedom in Europe*.

Broniatowski, D. A.; Jamison, A. M.; Qi, S.; AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S. C.; and Dredze, M. 2018. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health* 108(10).

Buckels, E. E.; Trapnell, P. D.; and Paulhus, D. L. 2014. Trolls just want to have fun. *Personality and Individual Differences* 67:97 – 102. The Dark Triad of Personality.

Cook, D. M.; Waugh, B.; Abdipanah, M.; Hashemi, O.; and Rahman, S. A. 2014. Twitter deception and influence: Issues of identity, slacktivism, and puppetry. *Journal of Information Warfare* 13(1):58–71.

Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. Botornot: A system to evaluate social bots. In *WWW*.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM* 59(7):96–104.

Flores-Saviaga, C.; Keegan, B.; and Savage, S. 2018. Mobilizing the trump train: Understanding collective action in a political trolling community. In *ICWSM'18*.

Herring, S.; Job-Sluder, K.; Scheckler, R.; and Barab, S. 2002. Searching for safety online: Managing "trolling" in a feminist forum. *The information society* 18(5):371–384.

Kollanyi, B.; Howard, P. N.; and Woolley, S. C. 2016. Bots and automation over Twitter during the first U.S. presidential debate: Comprop data memo 2016.1. Oxford, UK.

Kumar, S.; Cheng, J.; Leskovec, J.; and Subrahmanian, V. 2017. An army of me: Sockpuppets in online discussion communities. In *WWW*, 857–866.

Latour, B.; Jensen, P.; Venturini, T.; Grauwin, S.; and Boullier, D. 2012. 'the whole is always smaller than its parts'–a digital test of gabriel tardes' monads. *The British journal of sociology* 63(4):590–615.

Latour, B. 2002. Gabriel Tarde and the end of the social.

Leskovec, J.; Backstrom, L.; Kumar, R.; and Tomkins, A. 2008. Microscopic evolution of social networks. In *SIGKDD*, 462–470.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *JMLR* 9(Nov):2579–2605.

Mihaylov, T.; Georgiev, G.; and Nakov, P. 2015. Finding opinion manipulation trolls in news community forums. In *CoNLL*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Navarro, G. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)* 33(1):31–88.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Phinney, J. S. 2000. Ethnic and racial identity: Ethnic identity.

Rizoiu, M.-A.; Graham, T.; Shang, R.; Zhang, Y.; Ackland, R.; Xie, L.; et al. 2018a. # debatenight: The role and influence of socialbots on twitter during the 1st 2016 us presidential debate. In *ICWSM'18*.

Rizoiu, M.-A.; Lee, Y.; Mishra, S.; and Xie, L. 2018b. Hawkes processes for events in social media. In Chang, S.-F., ed., *Frontiers of Multimedia Research*. 191–218.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592* 96–104.

Stewart, L. G.; Arif, A.; and Starbird, K. 2018. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web*.

Tarde, G. 2011. Monadology and sociology, re. press. *Melbourne (originally published in 1893)*.

Varol, O.; Ferrara, E.; Davis, C. A.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2018. Who let the trolls out? towards understanding state-sponsored trolls. *CoRR* abs/1811.03130.