Real-time Detection of Content Polluters in Partially Observable Twitter Networks

Mehwish Nasim
School of Mathematical Sciences
University of Adelaide
Adelaide, Australia
mehwish.nasim@adelaide.edu.au

Andrew Nguyen
School of Mathematical Sciences
University of Adelaide
Adelaide, Australia
andrew.nguyen03@adelaide.edu.au

Nick Lothian*
Tyto.ai
Adelaide, Australia
nick.lothian@gmail.com

Robert Cope
School of Mathematical Sciences
University of Adelaide
Adelaide, Australia
robert.cope@adelaide.edu.au

Lewis Mitchell
School of Mathematical Sciences
University of Adelaide
Adelaide, Australia
lewis.mitchell@adelaide.edu.au

ABSTRACT

Content polluters, or bots that hijack a conversation for political or advertising purposes are a known problem for event prediction, election forecasting and when distinguishing real news from fake news in social media data. Identifying this type of bot is particularly challenging, with state-of-the-art methods utilising large volumes of network data as features for machine learning models. Such datasets are generally not readily available in typical applications which stream social media data for real-time event prediction. In this work we develop a methodology to detect content polluters in social media datasets that are streamed in real-time. Applying our method to the problem of civil unrest event prediction in Australia, we identify content polluters from individual tweets, without collecting social network or historical data from individual accounts. We identify some peculiar characteristics of these bots in our dataset and propose metrics for identification of such accounts. We then pose some research questions around this type of bot detection, including: how good Twitter is at detecting content polluters and how well state-of-the-art methods perform in detecting bots in our dataset.

CCS CONCEPTS

• **Information systems** → **Social networking sites**; • **Security and privacy** → *Social network security and privacy*;

KEYWORDS

Civil unrest, Social bots, Content polluters, Missing links, Twitter

ACM Reference Format:
Mehwish Nasim, Andrew Nguyen, Nick Lothian, Robert Cope, and Lewis Mitchell. 2018. Real-time Detection of Content Polluters in Partially Observable Twitter Networks. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3184558.3191574

*Work undertaken while at Data to Decisions CRC.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191574

1 INTRODUCTION

1.1 Motivation

Bots and content polluters in online social media affect the socio-political state of the world, from meddling in elections [4, 13, 37] to influencing US veterans [15]. In late September 2017, Twitter admitted to Congress that it had found 200 Russian accounts that overlapped with Facebook accounts which were used to sway Americans and create divisions during the elections held in 2016 [37]. Of course, some bots are useful as well, for instance accounts that will tweet alerts to people about natural disasters. The problem arises when they try to influence people or spread misinformation. The importance of detecting bots in online social media has produced an active research area on this topic [9, 21].

State-of-the-art methods for bot detection use historical patterns of behaviour and a rich feature set including textual, temporal, and social network features, to distinguish automated bots from real human users [35]. However, for real-time application using large streamed datasets, such methods can be prohibitive due to the sheer volume, velocity, and incompleteness of data samples. In this work we develop a new method to detect one particular type of social bot – content polluters – in streamed microblog datasets such as Twitter. Content polluters are bots that attempt to subvert a genuine discussion by hijacking it for political or advertising purposes. As we will show, these bots are a major concern for applications such as real-time event prediction, such as social unrest, from social media datasets.

1.2 Problem context

Social unrest prediction is a growing concern for governments worldwide. This is evidenced by DARPA's Open Source Intelligence program, which produced numerous methods to predict the occurrence of future population-level events such as civil unrest, political crises, election outcomes and disease outbreaks [12, 25, 30, 32]. It has been observed that social events are either preceded or followed by changes in population-level communication behaviour, consumption and movement. A large fraction of population-level

changes are implicitly reflected in online data such as blogs, online social networks, financial markets, or search queries. Some of these data sources have been shown to effectively detect population-level events in real time. Methods have been developed for predicting such events by fusing publicly available data from multiple sources. There exists a plethora of research focused on social media-based forecasting models, suggesting that features from micro-blogs such as Twitter can predict and detect population-level events [30]. Once one develops a "gold standard" (ground truth) record of known events (e.g. election results, or protests occurring) models can be trained using open source data to make predictions. A significant challenge for such models is noise reduction through filtering "fake news", removing misclassified or irrelevant tweets, or mitigating the effects of missing data. This is of particular concern, as the changing limits on accessing social media data remains a major challenge for researchers [26]. Access to data through APIs and third parties can be inconsistent, incomplete, and corrupted by noise in the form of bots. Where bots are influencing people through fake social media accounts, they also act as *content polluters* on social media sites [33]. According to the Digital Forensics Research Lab (DFRL), "They can make a group of six people look like a group of 46,000 people."

The main goal of our work was finding out content polluters in a dataset comprising tweets related to Australian social unrest events in real time, without access to complete profile information of the users. Due to rate limits on the public API and the high cost of accessing data, we were restricted to using only streamed tweets satisfying certain criteria. While the actual event prediction algorithm is not the primary concern of this paper, further detail can be found in Osborne *et al.* [29].

1.3 Related Work

A social bot is a computer algorithm that automatically produces content and interacts with humans on social media, trying to emulate and possibly alter their behaviour [14]. Social bots inhabit social media platforms, and online social networks are inundated by millions of bots exhibiting increasingly sophisticated, human-like behaviour. In the coming years a proliferation of social media bots is expected as advertisers, criminals, politicians, governments, terrorists, and other organizations attempt to influence populations [34]. This introduces dimensions for social bots, including social network characteristics, temporal activity, diffusion patterns, and sentiment expression [14].

Ghost *et al.* [16] conducted an analysis on the follower/followee links acquired by over 40,000 spammer accounts suspended by Twitter. They showed that penalizing users for connecting to spammers can be effective because it would de-incentivize users from linking with other users in order to gain influence. Yang *et al.* [40] found that bot accounts in online social networks connect to each other by chance and integrate into the social network just like normal users. Network information along with content has been shown to detect spam in online social networks [20]. While researchers were proposing various bot-detection models, Lee *et al.* [24] identified and engaged strangers on social media to effectively propagate information/misinformation. They proposed a model to leverage peoples' social behaviour (online interactions) and users' wait times for retweeting.

Social bots evolve over time, making them resilient against standard bot detection approaches [9]. They are apt at changing discussion topics and posting activities [38]. Researchers have proposed complex models, such as those based on interaction graphs of suspicious accounts [19, 20, 22, 39]. An adversary often controls multiple social bots known as a *sybil*. One strategy to detect such accounts relies on investigating social graph structure, on the assumption that sybil accounts link to a small number of legitimate users [7]. Behavioural patterns and sentiments analysis have also been used for bot detection [11]. Such patterns can easily be encoded in features, thus machine learning techniques can be used to distinguish bot-like from human-like behaviour. Previous work uses network-based features or content analysis for bot detection, along with indicators such as temporal activity, retweets, and crowd sourcing [10, 36]. Such efforts require substantial network knowledge or the ability to quickly query an API for a complete history of social media postings by suspected bots. However, real-time applications, such as streaming messages based on keywords or geographic locations, render this impractical. A major challenge therefore is developing methodologies to detect and remove bots based on partial information, message histories, and network knowledge, in real time.

In this work we detect bots from individual tweets downloaded for predicting social unrest in Australian cities. Given filters on keywords and geographic location of events (such as protests, rallies, civil disturbances) collected in real time, it leaves a small but informative dataset for prediction. Predictions are generated in real time by analysing data from online social media platforms such as Twitter and validated against hand-labeled "Gold standard records" (*GSR*) [29]. The GSR is created by the news analysts; after going through a validation and cleaning process this data is ready to be used as the ground truth. If Twitter data is contaminated with social bots, it can greatly degrade prediction models. It is therefore imperative to develop techniques for detecting and removing social bots for real-time data streams.

Contributions: Our scientific contributions are as follows:

(1) We develop a method to identify social bots in data using only partial information about the user and their tweet history, in real time.
(2) We present a new dataset of hand-labelled bots and legitimate records, and use it to validate our method[1].
(3) We pose a set of research questions for evaluating whether Twitter users, Twitter, or existing state-of-the-art bot detection methods could detect bots in our dataset or not.

1.4 Dataset

Our dataset consists of timestamped tweets from 1 January 2015 till 31 December 2016 from 5 major capital cities in Australia. Tweets identify one of the following locations: 'Australia', 'Adelaide', 'Brisbane', 'Melbourne', 'Perth', or 'Sydney'. The data are targeted at studying civil unrest and intends to capture ways in which people express opinions and organize marches, rallies, peaceful/violent protests etc., within Australia. Such events aim to draw attention toward an issue e.g., infrastructure, taxes, immigration laws etc. Australia has a population of about 24.5 million people and, like

[1]Data can be accessed on http://maths.adelaide.edu.au/mehwish.nasim/

Table 1: Data statistics

Parameters	Adelaide	Brisbane	Melbourne	Perth	Sydney
Number of tweets	14087	5913	23720	8421	31568
Number of unique users	12039	3466	14611	6215	14515
Number of unique URLs	548	233	762	456	844
Average number of followers (in degree)	8812	9624	6733	5409	6052
Average number friends (out degree)	1223	1736	1517	1643	1860
Number of verified accounts	293	432	840	209	412

in many developed countries, predicting civil unrest events is of interest to law enforcement agencies, government bodies, media and academia. Notwithstanding this fact, the literature is devoid of exploratory studies conducted on this population for real-time prediction of civil unrest events. The basic statistics about protest-related tweets in our dataset are reported in Table 1.

Note that the dataset was devoid of information on the alters (followers/friends of egos), except for the total count of alters (numbers of followers and friends).

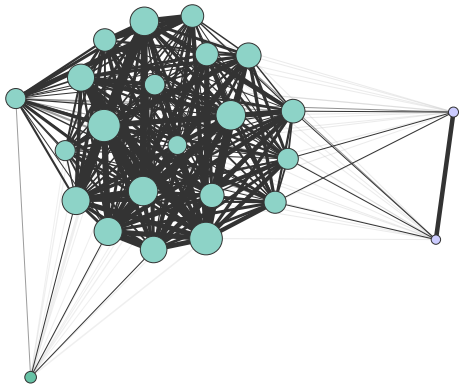
2 DETECTING CONTENT POLLUTERS

We investigate two characteristics of tweets i.e., temporal information and message diversity in a tweet.

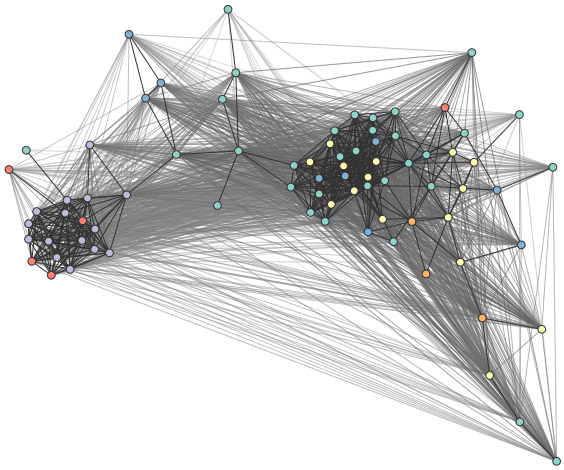
Temporal Patterns: In the first step we were interested in 1). users who tweet frequently, 2). pairs of users who tweet on the same day using the desired keywords. Since no information about the network of individual users is available, we cannot construct a follower-friend network graph. Instead, we construct a two mode user-event network. For all the events in the data we connect two users if they have tweeted on the same event day. We represent this problem in graph theoretic terms as follows:

Let G be a bipartite graph of users and events. Let U be the set of users and let V be the set of events. Let $u, v \in U$ and let $i, j \in V$. For any $i \in V$ if $N(u) \cap N(v) \neq \{\}$ then $(u, v) \in E$ in the one-mode projection of the bipartite graph. The *neighbourhood* $N(v)$ of a vertex $v \in U$ is the set of vertices that are adjacent to v. The resulting projection is an undirected loopless multigraph. If the edge set E contains the same edge several times, then E is a multiset. If an edge occurs several times in E, the copies of that edge are called parallel edges. Graphs that have parallel edges are also called multigraphs.

Similar to other social networks such as friendship networks, event networks are a result of complex sociological processes with a multitude of relations. When such relations are conflated into a dense network, the visualization often resembles a "hairball". Various approaches to declutter drawings of such networks exist in the literature. We use the recent *backbone layout* approach for network visualization [28], which accounts for strong ties (or multiplicity of edges) and uses the union of all maximum spanning trees as a sparsifier to ensure a connected subgraph. In Figure 1b, the thickness of edges represents how often a pair of nodes tweet on the same 'event day'[2] whereas, the size of the nodes indicates the individual frequency of tweets by a user[3]. We noticed that bots



(a) Two purple nodes at the right side that are loosely connected to the core, are bots. They have tweeted together frequently and their individual frequency to tweet is low as compared to other nodes in the graph, however the dyadic (pairwise) frequency is higher.



(b) Two densely connected components in the tweets graph.

Figure 1: Graphs containing bots and legitimate users from the Melbourne events network.

tweeted together frequently. Their individual frequency to tweet is low as compared to other nodes in the graph, however the dyadic (pairwise) frequency is higher. For instance, the two purple nodes on the right have tweeted together frequently, in Figure 1a. Their individual frequency to tweet is low as compared to other nodes

[2]Event day was confirmed from the *GSR*.
[3]Networks visualizations are created in *visone* (http://www.visone.info/).

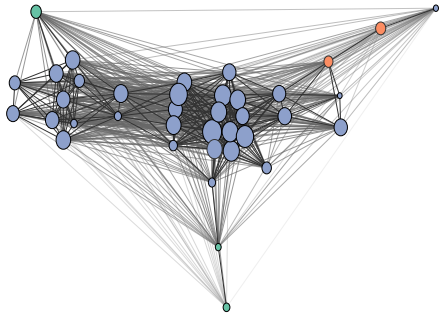


Figure 2: Graph containing bots and legitimate users from the Melbourne events network.

in the graph, however the dyadic (pairwise) frequency is higher. These two nodes are weakly connected to the core. Upon checking their complete profiles, the users were found to be political bots. This motivated us to further explore the tweets-graph.

The core of the network (green nodes) were found to be news channels and popular blogs in Australia, such as *MelbLiveNews*, *newsonaust*, *7NewsMelbourne* and *LoversMelbourne* to name a few. Media accounts are likely to report population-level events on the day of the events, thus they form a strongly-connected core of the events network graph.

We then clustered all tweets in a similar manner to construct a graph where two users have an edge between them if they have tweeted on the same day, irrespective of whether there was an event that day or not. We used the Louvain Method for clustering the network [5], based on the concept of *modularity*. Optimizing the modularity results in the best possible grouping of nodes in a given network. We then found two strongly-connected components in the graph: 1. News channels, and 2. Bots. We analysed the strongly-connected vertex-induced subgraphs from the network. One such component for the city of *Melbourne* is shown in Figure 2, which is a strongly-connected component from Figure 1b. Bots are the purple nodes (validated by manual inspection of profiles). Green nodes represent false positives. Orange nodes are not bots but are also not relevant for predictions, since these users were not geographically located in Australia and were tweeting about Victoria in the UK.

Message diversity: We computed the diversity in the tweets based upon mentions of URLs and hashtags. We selected the top most tweeted URLs, $\{K\}$ ($|K| = 20$), and then filtered out the users ($\bar{U} \subseteq U$) who mentioned those URLs. The motivation for this approach is that an event prediction model should be resilient against bot-URLs that are infrequently mentioned in the tweets, so these will not greatly impact the prediction accuracy. We then computed the following three measures for each of the remaining users: i). total number of tweets containing any URL(s), u_i^{all}, ii). number of tweets mentioning URL $k \in K$, u_i^k and iii). diversity score i.e., the difference between the two measures, $u_i^d = u_i^{all} - u_i^k$.

We then plot the diversity score distribution for every $u^k \in \bar{U}$, for every URL $k \in K$. This immediately provides some relevant insights about the behaviour of content polluters: Figure 3a shows a legitimate URL (i.e., linked to by legitimate users), whereas, Figures 3b and 3c show bot-URLs (i.e., URLs linked to by bots). Users who tweet these URLs are classified as *potential bots*. The figures show that the diversity of users linking to legitimate URLs is generally far greater than those linking to bot-URLs. The temporal patterns of bot-URL mentions and those which are being tweeted at regular intervals indicated that these users were indeed bots.

We measure the extent of diversity in two ways:

(1) *The Gini* coefficient ($G \in \mathcal{R}$, G=[0,1]):

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |u_i^d - u_j^d|}{2n \sum_{i=1}^{n} u_i^d}, \tag{1}$$

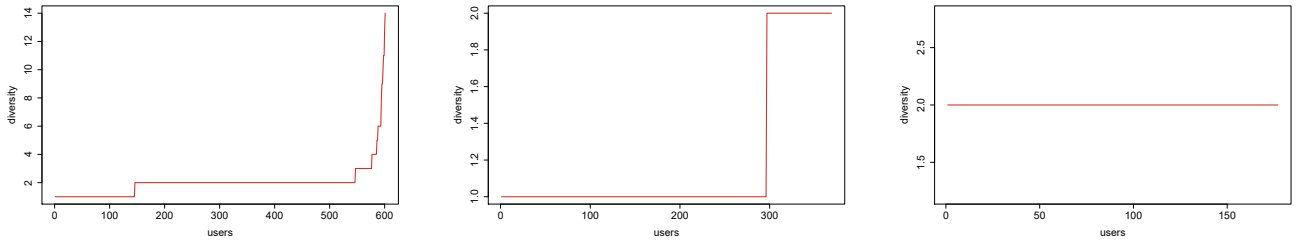
where n is the number of users tweeting a particular URL. The Gini coefficient G describes the relative amount of inequality in the distribution of diversity: $G = 0$ indicates complete equality while $G = 1$ indicates complete inequality. A high G suggests coordination among the observations. The Gini coefficient does not measure absolute inequality and the interpretation can vary from situation to situation. Legitimate accounts such as news channels, newspapers, and famous activists are likely to tweet legitimate and diverse URLs, thus the Gini coefficient for legitimate URLs is high as compared to illegitimate URLs. The Gini coefficient for a sample of ten URLs is shown in Figure 4.

(2) *Rank-size Rule:* We observed that only a fraction of URLs are mentioned very frequently in the tweets and very large number of URLs barely find their way in more than a single tweet. It is interesting to note that cities and their rank also follow a similar distribution; this pattern is generally known as the *rank-size rule* [31]. This has also been observed in various studies on calling behaviour of users [2][3] [27].

We fit a curve on every user versus URL-diversity graph and measure the coefficient of determination R^2. Values close to zero indicate that the model explains little of the variability of the response data around its mean. For legitimate URLs, we obtained values close to 1 (Figure 3).

Recently, Gilani *et al.* [18] evaluated the characteristics of automated versus human accounts by looking at complete tweet histories. They initially hypothesized that bots tweet a number of different URLs, however in the actual data they found that humans may also post a number of URLs. Conversely, in this work we looked at most frequently posted URLs and then for each URL we analysed how diverse the users' tweets are who are tweeting that URL.

We detected 849 bots in the data using message diversity on URLs, which we call *content polluters*. These content polluters contributed about 7% of tweets in the data. We computed some statistics on content polluters versus legitimate users, shown in Figure 5. In [14], authors argued that social bots tend to have recent accounts with long names. However, we did not find a significant difference in our data between content polluters and regular users. The average account age of content polluters accounts was 2.9 years as compared to legitimate users which was 4.2 years. This difference was significant ($p < 0.01$). This suggests that these particular type of bot accounts are relatively old and have remained (potentially) undetected by Twitter. The length of Twitter names for bots had on average 11 characters as compared to non-bots that had 12 characters. None of the bots had *verified* Twitter accounts. A total of 109 political bot accounts were created on 20 February 2014 with



(a) Legitimate (Gini = 0.8, $R^2 = 0.98$) **(b) Bots (Gini = 0.32, $R^2 = 0$)** **(c) Bots (Gini = 0, $R^2 = 0$)**

Figure 3: Message diversity measured through 3 URLs for bots and genuine users.

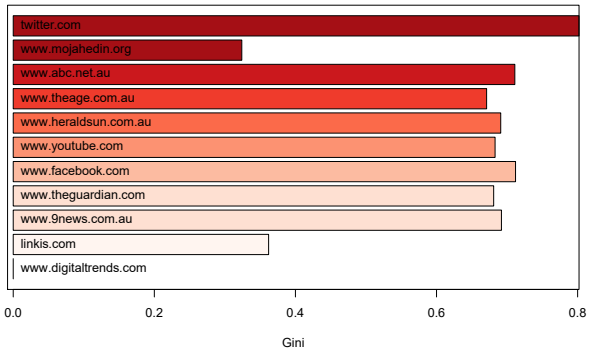


Figure 4: Gini score for ten URLs. High Gini coefficient indicates a legitimate URL. The three URLs with the lowest Gini coefficients were being tweeted by content-polluting bots.

only 12 unique names, a strong indication of being a bot network. We also found several digital media bot accounts. Such accounts aim at becoming famous by attracting followers [6]. A set of such accounts was created on 30 March 2016. This set consisted of 8 accounts with an average friend count of 4099 and follower count of 1112.

We also explored the dataset from [23] using our algorithm. The dataset contains more than 600k tweets. The Gini coefficient for each dataset (bots and non-bots) was around 0.5, hence we remain inconclusive. The data set from Gilani *et al.* [18] only consisted of the number of URLs each user mentioned, therefore it was not possible to check the relative frequency of any particular URL. We argue that the nature of content polluting bots makes them difficult to distinguish in traditional bot-detection datasets. This motivates our research questions below and the creation of a new human-validated content-pollution dataset in the next section.

3 CREATING A CONTENT-POLLUTING BOT DATASET

Given the peculiarities in the bot accounts that we found in our analysis, we move on to some pertinent research questions.

3.1 Do humans succeed in detecting content polluters?

We conducted a user study to hand-label a set of Twitter accounts that contained equal number of content polluters (from our list obtained in the previous section) and legitimate accounts. We asked three independent hand-labellers to create the dataset. Users were first shown several examples of content polluters as well as of legitimate accounts. All three participants were well versed with using Twitter. All participants found it very difficult to assess non-English accounts even with automatic translation.

The participants recorded the following comments:

Participant 1
Domain Knowledge: Advance Twitter User
Comments: *"What I'm struggling with is that, the user doesn't actually initiates a suspicious tweet. He simply retweets a whole bunch of content polluting tweets".*

Strategy:

- If user has tweeted or retweeted from well known news spam sites then mark as bot.
- Otherwise look through pattern of tweets, if very spamy tweet behaviour, for example highly consistent frequency of tweeting behaviour and tweets are from a single source then mark as bot.
- See if they regular mention and interact with other twitter users which indicates a good sign for a regular account.
- Look at profile details and follower and followees ratio to distinguish if it appears like a regular account or a bot.

Participant 2
Domain Knowledge: Twitter User/Domain Expert
Comments: *"This was a really hard task. It is not at all clear what the difference is between a bot and a human. This is much slower than labelling individual tweets."*
Strategy:

- Look at the twitter account. If the user has tweeted well-known news spam URLs/services (@convoy, dv.it, 360WISE, mojahedin.org) then mark it as a bot.
- If not, scroll through the account. If I can find some original content (e.g., comments on a retweet) then mark it as a legitimate account.
- If it always retweets from a single or only a few accounts then I mark it as a bot.
- Otherwise it comes down to judgement. This includes things like looking at avatar icons - if all the followers seem strangely similar (all Anime figures for example, or all faded pictures), or if it always uses the same non-twitter link shortener then I mark it as a bot.
- Then there were a set of accounts who all posted the same content. I only noticed this after a while, so I probably only caught some of them.
- There are a set of Markov-chain-like accounts (e.g. 1240541203). It can be difficult to distinguish them because of the messy, non-standard language typical of many Twitter users, and the limited amount of text to work from.

Participant 3
Domain Knowledge: Everyday Twitter User
Comments: *" I think there are lots of accounts that are part automatic and potentially part human (e.g., with the annoying "I've gained/lost n followers" tweets), which seems like a challenge. I tried to work out if these were actually human."*
Strategy:

- In each case, I would skim through the tweets. The presence of coherent original content without a URL suggested that the account was likely human.
- Tweets from recognisable spam sources (e.g. 360WISE) suggested the account was a bot.
- Overwhelming consistency across tweets suggested a bot (i.e., every tweet with exactly the same text formatting, or using the same URL shortener), except when this was associated with what appeared to be a carefully curated account for a business.
- For the rest, I looked at profile information, follower ratios, and the broader content of the tweets and made the best judgement I could.

We constructed a labelled dataset from this collection by recording where 2 out of the 3 hand-labellers agreed on a classification. For our content polluter algorithm, we observed that the proportion of observed correct prediction (for both the classes) on this hand-labelled dataset was 0.57.

We test the following hypothesis: $\mathbf{H1_a}$: Our method was able to find bot/non-bot accounts with greater than 50% accuracy. Hence, the null hypothesis is: $\mathbf{H_0}$: Our method randomly labelled the bot/non-bot accounts.

After applying a t test we reject the null hypothesis at a significance level of $\alpha = 0.05$ ($p = 0.00029$).

3.2 How efficient is Twitter in detecting social bots?

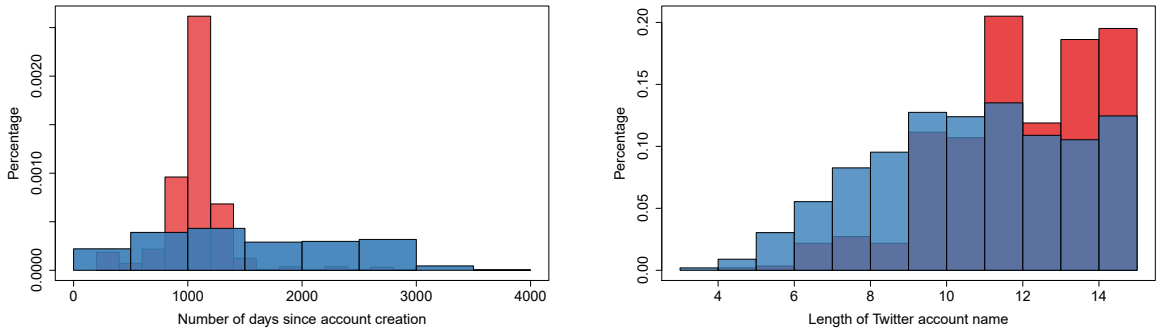
Twitter continuously searches for suspicious accounts, and accounts that are found to be malicious may be deleted. In this experiment we studied how many bot accounts Twitter suspended from our detected bots list. The dataset that we analysed is from 2015/2016 but we conducted this experiment in April 2017, giving us a comprehensive set of accounts determined by Twitter to be bots. We used the Twitter API for this experiment. Given a query for a specific account, the Twitter API returns an error message if the account is suspended by Twitter or deleted by the user. It the error code 63 is returned then it means Twitter has suspended the account, whereas, error code 50 means the user deleted the account. For active accounts, metadata information about the account is returned. Upon querying the Twitter API, we found that Twitter had suspended 153 accounts out of the 849 content polluters that we have detected.

3.3 How efficient other methods are for bot detection?

We also tested the performance of state-of-the-art bot detection system called *Truthy* [10], also known as BotOrNot? [35]. It is a publicly available API service developed by the Indiana University at Bloomington in May 2014 to evaluate the similarity of a Twitter account with the known characteristics of social bots. It uses the complete profile of a user to determine how likely the user account is to be a bot. BotOrNot? employs a supervised machine-learning classifier that exploits more than 1000 features from the Twitter account under investigation. Features are derived based on network information and tweeting behaviour. The authors state that despite the fact that the service is specifically designed for detection of bots, the performance against evolved spambots might be worse than that was reported in the paper. We queried this service against our genuine set of bot accounts. Truthy displays scores against each account. Higher scores mean more bot-like accounts. Figure 6 shows the overall performance of Truthy against our list of bots. The mean score was 0.55 with a standard error of 0.14. Table 2 shows the performance summary. We remark that for the task of detecting content polluters our method performs comparably to Truthy, using only the information of URL diversity at the sampled-tweet level. We reiterate that we detected content polluter accounts using message diversity since we did not have access to complete account information, whereas Truthy exploited features obtained from the complete user profile and network. However, the aim of Truthy is much different from what we are trying to achieve. We utilize **single tweets** with users' metadata to filter out bots in real-time for event prediction.

4 DISCUSSION

In this work we discovered social bots in protest-related tweets, using a stream of sampled Twitter data. Unlike previous bot-detection



(a) Account creation time **(b) Length of Twitter account name**

Figure 5: Characteristics of users. Bots(red) versus Legitimate users(blue)

Table 2: Performance summary of our method versus Truthy. According to Truthy, 65% of the true positives in the user study were likely to be a bot, whereas, 21% of false positives also had a greater than 0.5 probability to be a bot.

	Truthy ($pr >= 0.5$))	Truthy (μ)	Truthy (σ)
True Positives reported by user study	65%	0.556	0.159
False Positives reported by user study	21%	0.392	0.131

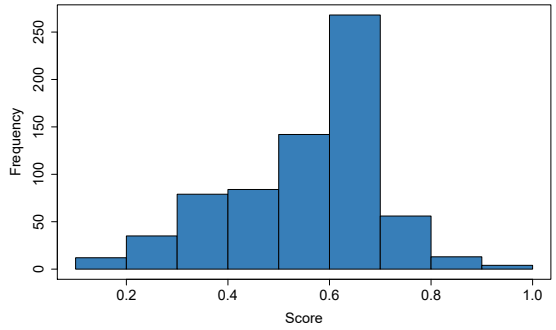


Figure 6: Results obtained from Truthy on the complete list of bots. Please note that the results were obtained after removing the 153 accounts that were suspended by Twitter.

studies, our dataset was devoid of network information and detailed tweeting history. We showed the efficacy of a start-of-the-art Twitter-bot detection technique using complete profile and network information on our dataset. Further, we analysed the capabilities of Twitter and naive users in distinguishing bots from legitimate users. Twitter continuously looks for malicious accounts, which are deleted. However, this process may be very slow and a number of accounts remain undetected [1]. It is difficult to track users who delete their tweets, since Twitter does not provide access to deleted tweets, or tweets older than 30 days. We show that even in the presence of complete network information, existing methods are not apt at detecting content-polluting bots. We argue that for real-time Twitter streams where it is difficult to obtain detailed profile information because of constraints on time and scalability, a cost-effective way is to compute the message diversity of tweets for each user. A low diversity might indicate suspicious accounts. The most challenging aspect of this work is to validate results since user perceptions are not always correct, and standard bot detection methods are very much prone to misclassification despite using complete twitter account information [9, 17, 18]. Results from our user study indicate that our method agrees with the participants on most accounts that are legitimate. However, there is some difference in opinion for the bot accounts since it largely involves human perception of what a bot or a content polluter could be. For instance, participant-3 indicated that they did not consider an account to be a bot when this was associated with what appeared to be a carefully curated account for a business. However, when we looked into the original tweets, certain users used hashtags such as '*#melbourne*', while promoting their business that had nothing to do with the city Melbourne. Participants also indicated that some accounts seemed to be part automated and potentially part human. Even advanced Twitter users found distinguishing between bots and legitimate users a challenging task. Deletion of tweets is a major issue for traditional bot detection methods. In the case of US elections, a recent news article says, "Twitter is either unable or unwilling to retrieve a substantial amount of tweets from bots and fake users spreading disinformation. Those users, which have been tied to Russia, have since deleted those tweets" [8]. In the absence of malicious tweets any validation method is prone to failure.

Performance Improvement: In February 2017 we used our content-polluters detection methodology in order to improve the performance of a social media-based predictive model [29]. Users from a large Australian law enforcement agency provided positive feedback on improvements in the predictions. We noted that the model was no longer erroneously predicting events related to 'escorts', which improved model performance noticeably. Further, removal of bots also removed 18 non-interesting events in the month of February, related to lottery ticket sales.

5 CONCLUSION AND FUTURE WORK

We found that content polluters in this dataset often timed their tweets together. By analysing the temporal patterns one could infer the presence of bot accounts. However, we also noticed that tweets from news channels were also temporally correlated. Using only temporal methods could lead to misclassification of legitimate accounts. We also found that bots used a small set of URLs in their tweets, therefore by finding out the most frequently-used URLs and computing their relative usage in tweets from the all the unique users in the dataset, one could successfully detect content polluters. Our analysis leads us to believe that conventional machine learning methods may require a number of features and may not be apt at correctly identifying bots. The bots that we detected in our dataset helped to remove noise in the data and significantly improved the performance of prediction models. In future we aim to:

(1) Analyse non-protest related tweets for detecting bots, and utilise other available relations such as user-event relations, temporal relations, social interactions etc.
(2) Characterize bots into various categories and explore whether some bots could even be useful for civil unrest prediction.
(3) Conduct more user studies with a larger number of participants, in order to further understand the characteristics of content polluters.

6 ACKNOWLEDGEMENTS

The authors acknowledge financial support from Data to Decisions CRC. MN and LM also acknowledge support from the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

REFERENCES

[1] Marco T Bastos and Dan Mercea. 2017. The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review* (2017), 0894439317734157.
[2] Frank Bentley and Ying-Yu Chen. 2015. The Composition and Use of Modern Mobile Phonebooks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2749–2758.
[3] Ofer Bergman, Andreas Komninos, Dimitrios Liarokapis, and James Clarke. 2012. You never call: Demoting unused contacts on mobile phones using DMTR. *Personal and Ubiquitous Computing* 16, 6 (2012), 757–766.
[4] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. (2016).
[5] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
[6] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th annual computer security applications conference*. ACM, 93–102.
[7] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. 2012. Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 15–15.
[8] Monica Chin. 2017. Report: Twitter deleted tweets related to the Russian investigation. (2017). http://mashable.com/2017/10/13/twitter-deleted-russian-tweets/#CIbGh7BglkqS
[9] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 963–972.
[10] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 273–274.
[11] John P Dickerson, Vadim Kagan, and VS Subrahmanian. 2014. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 620–627.
[12] Andy Doyle, Graham Katz, Kristen Summers, Chris Ackermann, Ilya Zavorin, Zunsik Lim, Sathappan Muthiah, Patrick Butler, Nathan Self, Liang Zhao, et al. 2014. Forecasting significant societal events using the Embers streaming predictive analytics system. *Big data* 2, 4 (2014), 185–195.
[13] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. (2017).
[14] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.
[15] Vlad Howard Philip N. Gallacher, John D. Barash and John Kelly. [n. d.]. Junk News on Military Affairs and National Security: Social Media Disinformation Campaigns Against US Military Personnel and Veterans. ([n. d.]). http://comprop.oii.ox.ac.uk/publishing/working-papers/vetops/
[16] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. 2012. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 61–70.
[17] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. 2017. Of Bots and Humans (on Twitter). In *Proceedings of the 9th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'17). https://doi. org/10.1145/3110025.3110090*.
[18] Zafar Gilani, Ekaterina Kochmar, and Jon Crowcroft. 2017. Classification of Twitter Accounts into Automated Agents and Human Users. In *Proceedings of the international conference on Advances in Social Network Analysis and Mining ASONAM*.
[19] Jin Seop Han and Byung Joon Park. 2013. Efficient detection of content polluters in social networks. In *IT Convergence and Security 2012*. Springer, 991–996.
[20] Xia Hu, Jiliang Tang, and Huan Liu. 2014. Online Social Spammer Detection.. In *AAAI*. 59–65.
[21] Imrul Kayes and Adriana Iamnitchi. 2017. Privacy and security in online social networks: A survey. *Online Social Networks and Media* 3 (2017), 1–21.
[22] Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2017. How to Manipulate Social Media: Analyzing Political Astroturfing Using Ground Truth Data from South Korea.. In *ICWSM*. 564–567.
[23] Kyumin Lee, Brian David Eoff, and James Caverlee. 2011. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter.. In *ICWSM*.
[24] Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou, and Jeffrey Nichols. 2014. Who will retweet this?: Automatically identifying and engaging strangers on twitter to spread information. In *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 247–256.
[25] Sathappan Muthiah, Bert Huang, Jaime Arredondo, David Mares, Lise Getoor, Graham Katz, and Naren Ramakrishnan. 2015. Planned Protest Modeling in News and Social Media.. In *AAAI*. 3920–3927.
[26] Mehwish Nasim, Raphaël Charbey, Christophe Prieur, and Ulrik Brandes. 2016. Investigating Link Inference in Partially Observable Networks: Friendship Ties and Interaction. *IEEE Transactions on Computational Social Systems* 3, 3 (2016), 113–119.
[27] Mehwish Nasim, Aimal Rextin, Numair Khan, and Muhammad Muddassir Malik. 2016. Understanding Call Logs of Smartphone Users for Making Future Calls. In *18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM.
[28] Arlind Nocaj, Mark Ortmann, and Ulrik Brandes. 2014. Untangling hairballs: From 3 to 14 degrees of separation. In *22nd International Symposium, Graph Drawing 2014*. 101–112.
[29] Grant Osborne, Nick Lothian, Grant Neale, Terry Moscou, Andrew Nguyen, Jie Chen, Wei Kang, and Brenton Cooper. 2017. The beat the news system: Forecasting social disruption via modelling of online behaviours. *Journal of the Australian Institute of Professional Intelligence Officers* (2017).
[30] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, et al. 2014. 'Beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1799–1808.

[31] Kenneth T Rosen and Mitchel Resnick. 1980. The size distribution of cities: an examination of the Pareto law and primacy. *Journal of Urban Economics* 8, 2 (1980), 165–186.
[32] Parang Saraf and Naren Ramakrishnan. 2016. EMBERS autogsr: Automated coding of civil unrest events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 599–608.
[33] Pablo Suárez-Serrato, Margaret E Roberts, Clayton Davis, and Filippo Menczer. 2016. On the influence of social bots in online protests. In *International Conference on Social Informatics*. Springer, 269–278.
[34] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. The DARPA Twitter bot challenge. *Computer* 49, 6 (2016), 38–46.
[35] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions : Detection , Estimation , and Characterization. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017) Online*. 280–289.
[36] Bo Wang, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2015. Making the most of tweet-inherent features for social spam detection on twitter. *arXiv preprint arXiv:1503.07405* (2015).
[37] Charlie Warzel and Emma Loop. [n. d.]. Twitter Tells Congress It Found 200 Russian Accounts That Overlapped With Facebook. ([n. d.]). https://www.buzzfeed.com/charliewarzel/twitter-russian-accounts?utm_term=.immV81Pgd#.siA4vLxop
[38] Liang Wu, Xia Hu, Fred Morstatter, and Huan Liu. 2017. Detecting Camouflaged Content Polluters.. In *ICWSM*. 696–699.
[39] Chao Yang, Robert Harkreader, and Guofei Gu. 2013. Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security* 8, 8 (2013), 1280–1293.
[40] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y Zhao, and Yafei Dai. 2014. Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8, 1 (2014), 2.