Detecting and analysing coordinated inauthentic behaviour on social media

Tim Graham
timothy.graham@qut.edu.au
@timothyjgraham (Twitter)
About the Centre

The QUT Digital Media Research Centre (DMRC) conducts world-leading communication, media, and law research for a flourishing digital society.

It is one of Australia’s top organisations for media and communication research, areas in which QUT has achieved the highest possible rankings in ERA, the national research quality assessment exercise. Our research programs investigate the digital transformation of media industries, the challenges of digital inclusion and governance, the growing role of AI and automation in the information environment, and the role of social media in public communication.

The DMRC has access to cutting-edge research infrastructure and capabilities in computational methods for the study of communication and society. We actively engage with industry and academic partners in Australia, Europe, Asia, the US, and South America; and we are especially proud of the dynamic and supportive research training environment we provide to our many local and international graduate students.

The DMRC is also a member of the global Network of Centers – a group of academic institutions with a focus on interdisciplinary research on the development, social impact, policy implications, and legal issues concerning the Internet.

The Queensland University of Technology (QUT) acknowledges the Turrbal and Yugara, as the First Nations owners of the lands where QUT now stands. We pay respect to their Elders, lores, customs and creation spirits. We recognise that these lands have always been places of teaching, research and learning. QUT acknowledges the important role Aboriginal and Torres Strait Islander people play within the QUT community.
Project collaborators

• Dr Tobias Keller

• Assoc. Prof. Daniel Angus

• Prof. Axel Bruns

• Mr Rio (Guangnan) Zhu
Structure

I. Background and motivation
II. Methods and algorithm
III. Preliminary results and evaluation
IV. Conclusion and future work
Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight

January 10, 2020 4.06pm AEDT

A targeted, coordinated online campaign has tried to mislead the public. While the myths have been debunked, the culpable parties remain unknown. SEAN DAVEY/AAP

In the first week of 2020, hashtag #ArsonEmergency became the focal point of a new online narrative surrounding the bushfire crisis.

The message: the cause is arson, not climate change.

Police and bushfire services (and some journalists) have contradicted this claim.

We studied about 300 Twitter accounts driving the #ArsonEmergency hashtag to identify inauthentic behaviour. We found 236 accounts with no connection to the crisis.

Authors

Timothy Graham
Senior Lecturer, Queensland University of Technology

Tobias R. Keller
Visiting Postdoc, Queensland University of Technology

Disclosure statement

Timothy Graham receives funding from the Australian Research Council.

Vaccines, 5G, Bill Gates: why are Australians gathering to spread coronavirus conspiracy theories?

Weekend anti-lockdown protests drew people with a range of grievances but are unlikely to grow as they have in the US

△ Any group with pre-existing conspiracy theories has projected their conspiracy theories onto the corona crisis,' digital media professor Alex Bruns says. Photograph: Scott Barbour/AAP

Anti-lockdown protests in Australia over the weekend mimicked some of the conspiracy theories and language of similar protests in the United States but were unlikely to grow in size or intensity. (https://www.theguardian.com/media/2020/may/12/vaccines-5g-bill-gates-why-are-australians-gathering-to-spread-coronavirus-conspiracy-theories)
Coordinated messaging

• “By co-retweeting the exact same third party message, astroturfing campaigns can boost the number of messages that fit their campaign goals” (Keller et al., 2019: 260)

• Accounts posting the same content, repeatedly and within a short time window (e.g. within 1 minute or 1 second)

• A similar method is used by Giglietto et al. (2020) to examine link posting coordination
  • CooRnet R package
Related work


2. Analysing user identity via time-sensitive semantic edit distance (t-SED): a case study of Russian trolls on Twitter
   • Kim, Graham et al. (2019)
   • Modified edit distance algorithm
   • Pairwise comparison between similarity of tweets authored by 3 types of Russian trolls
   • Reveals coordination and behavioural strategies

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).
II. Methods and algorithm
Definitions

• Let $G$ be an undirected, weighted graph where:
  • The nodes $v \in V$ are social media accounts on a given platform(s) (e.g. Facebook or Twitter).
  • The edges $e \in E$ represent the number of temporally coordinated messages (e.g. co-retweets, co-tweets, co-links) between pairs, $\{v_i, v_j\}$.

• To construct edges, we set a time threshold $\delta$ to find eligible pairs;
  • Previous studies have used $\delta \leq 60$ seconds [Keller et al. (2020)].

• For a co-retweet network, minimally there are four data points required:
  1. Tweet ID
  2. Tweet timestamp
  3. Account ID (i.e. the author of the tweet)
  4. Retweet ID (i.e. the ID of the original tweet)
Detecting coordination at scale

• The number of comparisons is equivalent to the number of edges in an undirected network: \( \frac{n(n-1)}{2} \).

• For small networks, not a big problem.
  • Naïve brute force [cf. Giglietto et al. (2020)].

• For most social media networks, we have a scaling problem \( \rightarrow \)
  • 1 million accounts = 500 billion comparisons.

• We implement a simple divide and conquer algorithm to help deal with this.
Sliding window co-retweet algorithm

1. Sort the messages (e.g. tweets) by time created.
2. Set a time slice $k$ for the windows, e.g. 60 seconds.
3. Set a $\delta$, e.g. 60 seconds.
4. Iterate through each time slice $k$, minimizing double counting of pairs:
   - e.g. 0:00 to 1:59 minutes; 1:00 to 2:59; 1:59 to 3:59; etc.
   - Use some tricks to reduce size of input within each time slice, e.g. subset the data where the frequency of the $\text{retweet\_id}$ is larger than 1.
   - Store discovered co-retweets in a list.
5. Combine all the results from each time slice and create co-retweet network $G$. 
Remarks on time complexity

• The function is $f\left(\frac{n(n-1)}{2}\right)$ where $n$ is the number of tweets and $k$ is the number of time slices/windows. Strictly speaking it is always going to be $O(n^2)$, in this case all tweets $n$ being in the same time slice (i.e. $k = 1$).

• At the same time though the best case is $O(n)$ when the number of time slices equals the number of windows ($k = n$).

• This is similar to the quicksort algorithm which is technically also $O(n^2)$, but most implementations tend to run as a usual case in $O(n \times \log(n))$.

• Thus, a likely complexity (but one that would require testing) is that we are closer to $O(n \times \log(n))$ in most cases.
Remarks on network filtering

1. Removing random chance or ‘serendipitous’ co-retweets;
   • Solution $\rightarrow$ minimum edge weight of 2.
   • This solution finds ‘loosely’ coordinated clusters

2. Empirically, we find that a minimum edge weight $\geq 5$ effectively filters out non-coordinated accounts (true negatives), but the trade-off is that we miss a lot of true positives.
   • This solution finds ‘strongly’ coordinated clusters

3. How to reveal highly automated accounts, or coordinated bot activity;
   • Solution $\rightarrow$ reduce $\delta$ to 1 second (instead of 60).

• Ultimately, the parameter values and filtering techniques are qualitative decisions that depend on the domain and research questions – the optimal approach is a question for future empirical evaluation.
III. Preliminary results and evaluation:
Three case studies
Debate Night 2016

- Dataset of 6.5 million tweets; 1.45 million accounts
- See: Rizoiu, Graham et al. (2018)

- **Nodes**: Twitter accounts.
- **Edges**: Accounts that co-retweeted *within 1 second* of each other, *at least twice*.
- 25,242 nodes
- 89,708 edges

- Account status colour codes:
  - Red: Suspended
  - Yellow: Deleted
  - Blue: Active
COVID-19 coordinated bot activity

- Tracking hashtags:
  - #coronavirus
  - #coronacrisis
  - #covid19
  - #COVID—19
  - #COVID-19
  - #covid19aus
- Over 25 million retweets (of 2.6 million original tweets).
- Modified co-retweet method with $\delta = 1$ (second).
- Co-retweet the same tweet within 1 second or less.
- Also use TweetBotOrNot2 bot detection model - [https://github.com/mkearney/tweetbotornot2](https://github.com/mkearney/tweetbotornot2)
COVID-19 highly automated co-retweet network

- Human (blue; score less than 0.2)
- Bot (red; score greater than 0.8)
- Indeterminate (yellow; score between 0.2 and 0.8).

- Minimum edge weight 2
- 5,752 nodes
- 6,559 edges
- Report in press – Graham et al. (2020)
Cluster #1 – Turkish political bots

Coordinated bots magnify political tensions in Turkey and sow confusion and anxiety about the spread of the coronavirus and its effective response from different (and competing) levels of government in the country.
5G conspiracy co-link network

- **Nodes**: Public Facebook Pages.
- **Edges**: Pages that sent the same link *within 60 seconds* of each other, *at least twice*.
- Nodes sized by degree (or ‘activity level’).
- Nodes coloured by community cluster:

<table>
<thead>
<tr>
<th>Modularity Class</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td></td>
<td>19.12%</td>
</tr>
<tr>
<td>277</td>
<td></td>
<td>11.45%</td>
</tr>
<tr>
<td>308</td>
<td></td>
<td>3.26%</td>
</tr>
<tr>
<td>101</td>
<td></td>
<td>2.21%</td>
</tr>
<tr>
<td>264</td>
<td></td>
<td>0.94%</td>
</tr>
<tr>
<td>156</td>
<td></td>
<td>0.82%</td>
</tr>
</tbody>
</table>
IV. Conclusion & future work
Conclusion

• Coordinated message analysis provides a whole new perspective on social media data.
• Coordination is an inalienable feature of inauthentic coordinated activity – difficult to hide!
• This approach offers an alternative to flawed ML based approaches to bot detection.
• We developed a reasonably scalable algorithm and are generalising this analytical framework.
• Preliminary evaluation suggests that the methods accurately detect coordinated activity, including inauthentic behaviour such as botnets and troll campaigns.
• Twitter and other platforms must act quickly to take up these approaches alongside ML and human based.
Future work and current challenges

1. Refining and **improving** the co-retweet algorithm; what other approaches could we take?
   - Topological data analysis
2. Investigating sampling approaches.
3. Generalising the algorithm to handle any form of coordinated messaging.
4. **Further empirical testing** of how the choice of algorithm parameters (especially $\delta$) affects the results.
5. Further empirical testing of network filters (especially edge weight).
6. Applying the method to other **interesting case studies**.

Figure 3: The nerve of a cover of a set of sampled points in the plane.

(Sourced from: Chazal and Mitchell (2017)).
Acknowledgments

This research is funded in part by the Centre for Responsible Technology, and by the Australian Research Council projects DP200101317 Evaluating the Challenge of ‘Fake News’ and Other Malinformation and FT130100703 Understanding Intermedia Information Flows in the Australian Online Public Sphere. Computational resources and services used in this work were provided by the eResearch Office, Division of Research and Innovation, Queensland University of Technology, Brisbane, Australia.

Facebook data are provided courtesy of CrowdTangle.
References


