2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Who's in the Gang? Revealing Coordinating Communities in Social Media

Derek Weber[1,2*] and Frank Neumann[1]
[1]School of Computer Science, University of Adelaide, Australia
[2]Defence Science and Technology Group, Adelaide, Australia
[*]derek.weber@{adelaide.edu.au,dst.defence.gov.au}

***Abstract*—Political astroturfing and organised trolling are online malicious behaviours with significant real-world effects. Common approaches examining these phenomena focus on broad campaigns rather than the small groups responsible. To reveal networks of cooperating accounts, we propose a novel temporal window approach that relies on account interactions and metadata alone. It detects groups of accounts engaging in behaviours that, in concert, execute different goal-based strategies, which we describe. Our approach is validated against two relevant datasets with ground truth data. See https://github.com/weberdc/find_hccs for code and data.**

I. Introduction

Modern information campaigns are participatory, exploiting online social networks (OSNs) by convincing users to become "unwitting agents" and relying on platform algorithms to spread desired narratives [1]. The use of *political* bots to influence the framing and discussion of issues in the mainstream media (MSM) remains prevalent [2]. Relevant research has focused on campaign detection [3], [4], identification of bots [5], and dissemination groups [6], but there is an increasing shift to look at how motivated actors coordinate their activities [7]–[9].

We present a new approach to detect groups engaging in potentially coordinated activities, revealed through anomalous levels of coincidental behaviour. Links in the groups are inferred from behaviours that, with intent, are used to execute a number of identifiable coordination strategies. We validate our new technique on various datasets and show it successfully identifies coordinating communities.

Our approach infers ties between accounts based on activity to construct *latent connection networks* (LCNs), in which *highly coordinating communities* (HCCs) are detected. We use a variant of *focal structures analysis* (FSA) [10] to do this, because of its focus on revealing influential sets of users rather than individuals. A window-based approach is used to enforce temporal constraints.

Comparison of two relevant datasets, including labeled ground truth, with a randomised dataset provides validation. These research questions guided our evaluation:

RQ1 How can HCCs be found in an LCN?
RQ2 How do the discovered communities differ?
RQ3 How consistent is the HCC messaging?
RQ4 Are the HCCs internally or externally focused?

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00

Further discussion of related work, the approach taken and analyses are available in [11].

II. Coordination Strategies

Online influence relies on two primary related mechanisms: *dissemination* and *engagement*. OSNs share a variety of features that enable these. **Dissemination** aims to maximise the reach of a message through repetition, targeting contentious social and political issues to cause outrage, or at least distract from other messaging. Using hashtags, reposting (e.g., Twitter retweets, Facebook shares, or Tumblr reposts), commenting on popular forums (e.g., Facebook Pages, subreddits, Twitter replies) or mentioning highly connected accounts are ways to disseminate a message. If the message is well crafted, it may entice others to spread it further, engaging with the content. **Engagement** is a subset of dissemination that explicitly solicits a response, targeting individuals or communities with rhetorical techniques that garner responses (e.g., making inflammatory comments, and sensationally presented biased or false reporting). For greatest effect teams of accounts will coordinate their actions strategically and avoid direct interaction, so their links must be inferred.

A number of online coordination strategies have been observed in the literature making use of both dissemination and engagement, including:

1) *Pollution*: flooding a community with repeated or objectionable content, causing the OSN to shut it down [7];
2) *Boost*: reposting content to make it trend [4], [6];
3) *Bully*: groups of individuals harassing another individual or community [12]; and
4) *Metadata Shuffling*: groups of accounts changing metadata to hide their identities [13].

A clarification of the present problem is, therefore:

To identify groups of accounts whose behaviour, though typical in nature, is anomalous in degree.

III. Methodology

The major OSNs share a number of features, primarily in how they permit users to interact. By focusing on these commonalities, it is possible to develop approaches that generalise across the OSNs that offer them.

Traditional social network analysis relies on long-standing relationships between actors. On OSNs these are typically friend/follower relations. These are expensive to collect and

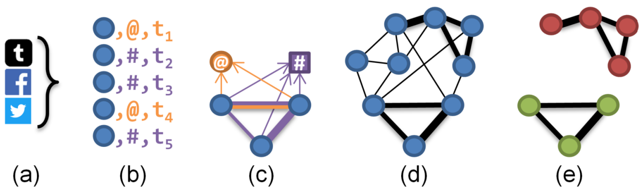


Fig. 1. Conceptual LCN / HCC pipeline (circles are accounts).

Algorithm 1 Extract HCCs (FSA_V)

Input: L=(V, E): An LCN, θ: HCC threshold
Output: H: Highly coordinating communities

$E' \leftarrow$ MergeMultiEdges(E)
$g\_mean \leftarrow$ MeanWeight(E')
$louvain\_communities \leftarrow$ ApplyLouvain(L)
Create new list, H
for $l \in louvain\_communities$ **do**
  Create new community candidate, $h = (V_h, E_h)$
  Add heaviest edge $e \in l$ to h
  $growing \leftarrow$ **true**
  while $growing$ **do**
    Find heaviest edge $\vec{e} \in l$ connected to h not in h
    $old\_mean \leftarrow$ MeanWeight(E_h)
    $new\_mean \leftarrow$ MeanWeight(Concatenate($E_h, \vec{e}$))
    if $new\_mean < g\_mean$ **or** $new\_mean < (old\_mean \times \theta)$ **then**
      $growing \leftarrow$ **false**
    else
      Add $\vec{e}$ to h
  if MeanWeight(E_h) $> g\_mean$ **then**
    Add h to H

quickly degrade in meaning if not followed with frequent activity. By focusing on active interactions, it is possible to understand not just who is interacting with whom, but to what degree. This provides a basis for constructing (or inferring) social networks, acknowledging they may be transitory.

LCNs are built from inferred links between accounts. Supporting criteria include retweeting the same tweet (*co-retweet*), using the same hashtags (*co-hashtag*) or URLs (*co-URL*), mentioning the same accounts (*co-mention*), or joining the same 'conversation' (a tree of *reply* chains with a common root tweet) (*co-conv*).

A. The LCN / HCC Pipeline

The key steps to extract HCCs from raw social media data are shown in Figure 1.

Step 1. Convert social media posts P to common interaction primitives, I_{all}. This step removes extraneous data and provides an opportunity for the fusion of sources.

Step 2. From I_{all}, filter the interactions, I_C, relevant to the set C=$\{c_1, c_2, ..., c_q\}$ of criteria (e.g., co-mentions and co-hashtags). Illustrated in Figure 1b are the filtered mentions (in orange) and hashtag uses (in purple), ordered by timestamp.

Step 3. Infer links between accounts given C, ensuring links are typed by criterion (e.g., links between accounts in Figure 1c). The result, M, is a collection of inferred pairings. We use $\beta^c_{\{u,v\}}$ for the number of inferred links between accounts u and v due to criterion $c \in C$.

Step 4. Construct an LCN, L, from the pairings in M. This network L=(V, E) is a set of vertices V representing accounts connected by weighted edges E of inferred links. These edges represent evidence of different criteria linking the adjacent vertices. The weight of each edge $e_{\{u,v\}} \in E$ between vertices representing u and v for each criterion c is $w^c_{\{u,v\}}$ (i.e., $\beta^c_{\{u,v\}}$).

To create a simple weighted network for easier processing, the multi-edges must be collapsed, however the edge weights are incomparable (e.g., retweeting the same tweet is not equivalent to using the same hashtag). For practical purposes, the edge weights can be combined with simple or weighted summation for simple community extraction, but implementations may maintain information of this step (e.g., the contribution of evidence for each criteria) for later analysis.

Some criteria may result in highly connected LCNs, even if its members never directly interact. The final step filters out these coincidental connections.

Step 5. Identify the highest coordinating communities, H, in L (Figure 1e), using a community detection algorithm appropriate to the scale of the data [14]. We propose FSA_V (Algorithm 1), a variant of FSA [10], which augments FSA's focus on revealing influential sets of accounts with edge weights without the computational cost of recursion or a stitching phase, resulting in a simpler algorithm. FSA_V divides L into communities using the Louvain algorithm [15] and builds candidate HCCs within each, starting with the 'heaviest' (i.e., highest weight) edge (representing the most evidence of coordination). It then attaches the next heaviest edge until the candidate's mean edge weight (MEW) is no less than θ ($0 < \theta \leq 1$) of the previous candidate's MEW, or is less than L's overall MEW. In testing, edge weights appeared to follow a power law, so θ was introduced to identify the point at which the edge weight drops significantly; θ requires tuning. A final filter ensures no HCC with a MEW less than L's is returned.

This algorithm prioritises edge weights while maintaining an awareness of the network topology by examining adjacent edges, something ignored by simple edge weight filtering. Our goal is to find sets of strongly coordinating users, so it is appropriate to prioritise strongly tied communities while still acknowledging coordination can also be achieved with weak ties (e.g., 100 accounts paid to retweet a single tweet).

Using FSA_V, the complexity of the entire pipeline is low order polynomial, O(n^2), due primarily to the pairwise comparison of accounts to infer links in Step 3, which we constrain with temporal window size.

B. Addressing the Temporal Aspect

Temporal information is a key element of coordination, and thus is critical for effective coordination detection. Frequent posts within a short period may represent genuine discussion or deliberate attempts to game trend algorithms [8]. We treat the post stream as a series of discrete windows to constrain detection periods. An LCN is constructed from each window

(Step 4), and these are then aggregated and mined for HCCs (Step 5). As we assume posts arrive in order, their timestamp metadata can be used to sort and assign them to windows.

IV. Evaluation and Validation

Our approach was evaluated by searching for *Boost* by co-retweet and other strategies in two datasets, while varying window sizes (γ). FSA_V was compared against two other community detection algorithms when applied to the LCNs built in Step 4 (aggregated). We then validated the resulting HCCs through content, temporal and network analysis.

TABLE I
Dataset statistics

	Tweets (T)	Retweets (RT)		Accounts	T / Account / Day	RT / Account / Day
DS1	115,913	63,164	(54.5%)	20,563	0.31	0.17
- GT	4,193	2,505	(59.7%)	134	1.74	1.04
DS2	1,571,245	729,937	(56.5%)	1,381	3.12	1.45

TABLE II
HCCs by coordination strategy

			GT			DS1			DS2		
	Strategy	γ	Nodes	Edges	Comp.	Nodes	Edges	Comp.	Nodes	Edges	Comp.
LCN	Boost	15	44	112	5	8,855	80,702	419	855	23,022	14
	Pollute	15	51	154	2	13,831	1,281,134	73	1,203	65,949	5
	Bully	60	70	482	1	16,519	1,925,487	222	1,103	37,368	5
FSA_V	Boost	15	9	6	3	633	753	167	113	758	19
	Pollute	15	9	5	4	135	93	50	24	15	9
	Bully	60	11	7	4	338	280	119	109	1,123	16
kNN	Boost	15	9	21	1	1,041	33,621	1	675	22,494	1
	Pollute	15	11	37	1	724	153,424	1	1,040	65,280	1
	Bully	60	18	135	1	1,713	663,413	1	692	35,136	1
Threshold	Boost	15	11	16	3	85	68	31	8	10	2
	Pollute	15	24	26	3	44	37	10	6	13	1
	Bully	60	15	19	3	25	23	8	10	10	3

TABLE III
HCCs by window size γ (Boost, FSA_V)

		Graph Attributes			HCC Sizes		Nodes in common			
	γ	Nodes	Edges	HCCs	Min.	Max.	γ=15	γ=60	γ=360	γ=1440
DS1	15	633	753	167	2	18	633	218	93	100
	60	619	1,293	151	2	13	-	619	208	193
	360	503	1,119	127	2	19	-	-	503	350
	1440	815	2,019	141	2	110	-	-	-	815
DS2	15	113	758	19	2	65	113	34	29	25
	60	77	394	18	2	27	-	77	62	54
	360	98	775	15	2	32	-	-	98	56
	1440	69	380	15	2	27	-	-	-	69

The two real-world Twitter datasets selected (Table I) represent two primary collection techniques: filtering a stream of posts using keywords direct from the OSN (DS1) and collecting the posts of specific accounts (DS2):

DS1 Tweets relating to a regional Australian election in March 2018, including a ground truth subset (GT); and

DS2 A large subset of the Internet Research Agency (IRA) dataset published by Twitter in October 2018[1].

DS1 consists of tweets collected during a regional Australian election over an 18 days period in early 2018. Nine hashtags and 134 candidate and party Twitter handles were used as filter terms and reply and quote chains were retrieved post-collection. The candidates' activity over the same period was collected separately, forming the ground truth. DS2 consists of the IRA tweets posted in 2016, the year of the US Presidential election, where evidence of coordination was expected to be greatest. The data were collected, held and analysed in accordance with University of Adelaide ethics protocol H-2018-045.

Window size γ was set at $\{15, 60, 360, 1440\}$ (in minutes) and the three community detection methods used on the aggregated LCNs were:

- FSA_V (θ=0.3);
- *k nearest neighbour* (*kNN*) with k=$ln(|V|)$ (*cf.* [4]);
- a simple threshold retaining the heaviest 90% of edges.

Values for θ and the threshold were based on experimenting with values in $[0.1, 0.9]$, maximising the MEW to HCC size ratio, using the γ=$\{15, 1440\}$ DS1 and DS2 LCNs.

A. Results

1) HCC Detection (RQ1): Detecting different strategies. The three detection methods all detected HCCs when searching for *Boost* (co-retweets), *Pollute* (co-hashtags), and *Bully* (co-mentions) (Table II). Notably, *kNN* consistently builds a single large HCC, highlighting the need to filter the network prior to applying it (*cf.* [4]). The *kNN* HCC is also consistently nearly as large as the original LCN for DS2, perhaps due to the low number of accounts and the fact that every edge of the retained vertices is retained, regardless of weight. It is not clear, then, that *kNN* is producing meaningful results.

Varying window size. Different strategies may be executed over different time periods, based on their aims. *Boost*ing a message to game trending algorithms requires the messages to appear close in time, whereas some forms of *Bully*ing exhibit only consistency and low variation (mentioning the same account repeatedly). Polluting a user's timeline on Twitter can also be achieved by frequently joining their conversations over a sustained period. Varying γ searching for *Boost*, we found different accounts were prominent over different timeframes (Table III); the overlap in the accounts detected in each timeframe differed considerably even though the number of HCCs stayed relatively similar. HCC sizes seemed to follow a power law; most were very small but a few were large.

HCC detection methods. Similarly, HCCs discovered by the three community extraction methods (Table IV) exhibit large discrepancies, suggesting that whichever method is used, tuning is required to produce interpretable results. This is evident in the literature, e.g., Cao *et al.* conducted significant pre-processing when identifying URL sharing campaigns [4]. Here we present the variation in results while controlling methods and other variables and keeping the coordination strategy constant, as our focus is the method's effectiveness.

2) HCC Differentiation and Consistency (RQ2/RQ3): The HCC detection methods used relied on network information; in contrast we examine content, metadata and temporal information to validate the results. We contrast DS1 and DS2 results with GT (*cf.* [9]) and a RANDOM dataset (*cf.* [4]), constructed by randomly assigning non-HCC accounts from

[1] https://about.twitter.com/en_us/values/elections-integrity.html

TABLE IV
HCCs BY DETECTION METHOD (BOOST, γ=15)

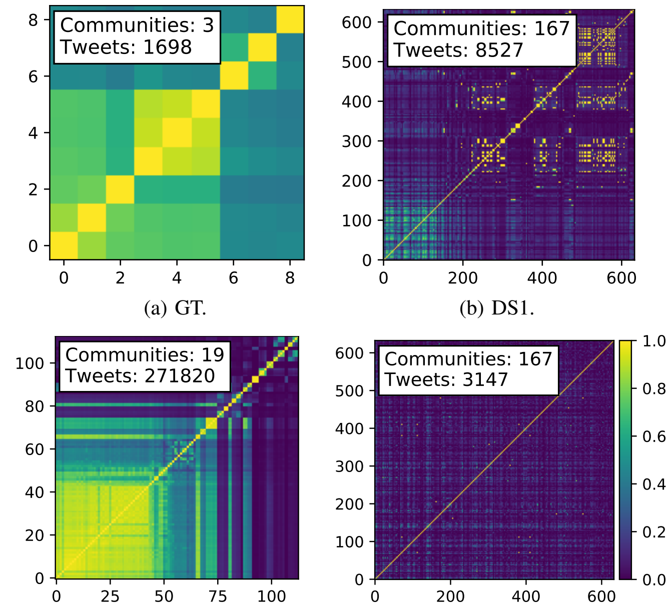
		Graph Attributes			HCC Sizes		Nodes in common		
		Nodes	Edges	HCCs	Min.	Max.	FSA_V	kNN	Threshold
DS1	FSA_V	633	753	167	2	18	633	56	36
	kNN	1,041	33,621	1	1,041	1,041	-	1,041	44
	Threshold	85	68	31	2	14	-	-	85
DS2	FSA_V	113	758	19	2	65	113	88	4
	kNN	675	22,494	1	675	675	-	675	8
	Threshold	8	10	2	2	6	-	-	8

DS1 to groups matching the distribution of its HCCs (FSA_V, γ=15). As DS2 consisted entirely of bad actors, it was felt non-HCC accounts from DS1 would be more representative of non-coordinating 'normal' accounts.

Internal consistency. If HCCs are boosting a message, it is reasonable to assume the content of HCCs members will be more similar internally that when compared externally, to the content of non-members. Treating each HCC member's tweets as a single document, we created a doc-term matrix using 5 character n-grams for terms, and then compared the members' document vectors using cosine similarity. This approach was chosen for its performance with non-English corpora [16], and because using individual tweets as documents produced too sparse a matrix. Visualising the similarities between accounts, grouping them by HCC (Figure 2), the HCCs are discernible as being internally similar. This method ignores the number of tweets HCCs post, so we can draw no conclusions about connections between HCC size and the internal similarity of their content, though more active HCCs (i.e., with more tweets) are more likely to be similar, through co-occurrence of n-grams. The RANDOM groupings demonstrated little to no similarity, internal or external, as expected, while the DS2 HCCs demonstrated high internal similarity, as expected of organised accounts over an extended period.

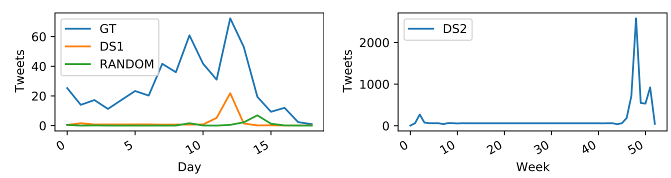
Messaging. In GT and DS2, where the ratio of tweets to accounts is high, similarity of content among HCC members is clear: the large yellow square in the bottom left of Figure 2c shows an HCC where each member's contribution is highly similar to that of the others. The effect is less clear for DS1, given the high number HCCs found by FSA_V, but the speckled yellow areas indicate high similarity between some HCCs - these could be candidates for merging.

Temporal patterns. Campaign types exhibit different temporal patterns [3]. We used the same temporal averaging technique as Lee *et al.* [3] to compare the daily activities of the HCCs found in GT, DS1 and RANDOM (Figure 3a) and weekly activities in DS2 (Figure 3b). The GT accounts were clearly most active at two points prior to the election (around day 15), during the last leaders' debate and just prior to the mandatory electoral advertising blackout. DS1 and RANDOM HCCs were only consistently active at different times: around the day 3 leaders' debate and on election day, respectively. Inter-HCC variation may have dragged the mean activity value down, as many small HCCs were inactive each day. Reintroducing FSA's stitching element to FSA_V may



(a) GT. (b) DS1. (c) DS2. (d) RANDOM.

Fig. 2. Similarity matrices of content posted by HCC accounts (γ=15, FSA_V). Each axis has an entry for each account, grouped by HCC. Each cell represents the similarity between the two corresponding accounts' content, calculated using cosine similarity (yellow = high similarity). Each account's content is represented as a vector of 5 character n-grams of all their tweets.



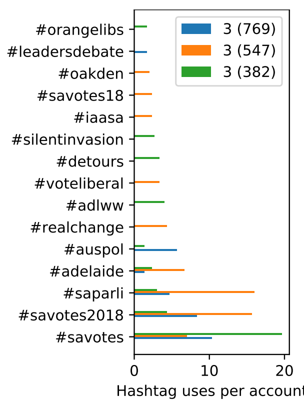
(a) GT, DS1 and RANDOM. (b) DS2.

Fig. 3. Averaged temporal graphs of HCC activities (γ=15, FSA_V).

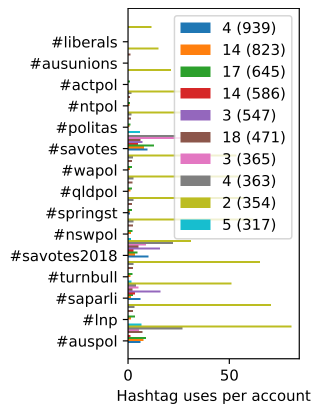
avoid this. In DS2, HCC activity increased in the second half of 2016, culminating in a peak around the election, inflated by two very active HCCs, both of which used many predominantly benign hashtags over the year.

Hashtag use. The most frequent hashtags in the most active HCCs revealed the most in GT. It is possible to assign some HCCs to political parties via the partisan hashtags (e.g., `#voteliberals`), although the hashtags of contemporaneous cultural events are also prominent (Figure 4a). DS1 hashtags are all politically relevant, but are dominated by a single small HCC which used many hashtags very often (Figure 4b). These accounts clearly attempted to disseminate their tweets through using $1,621$ hashtags in 354 tweets. Similarly, DS2 hashtags were dominated by a single HCC (using $41,317$ relatively general hashtags in $40,992$ tweets) and one issue-motivated HCC (Figure 4c). Given DS2 covers an entire year, it is unsurprising that the largest HCCs use such a variety of hashtags that their hashtags do not appear on the chart.

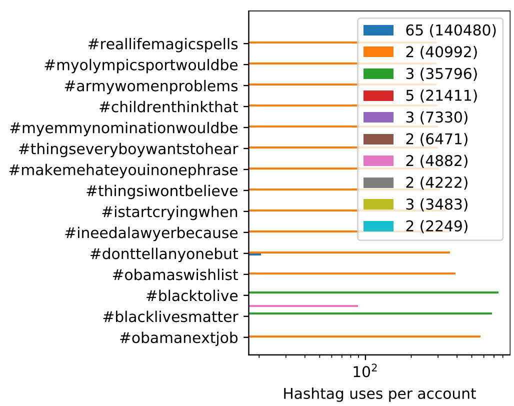
Analysing co-occurring hashtags can help further explore the HCC discussions to determine if HCCs are truly single groups or merged ones. Applied to GT HCC activities (Fig-



(a) GT.



(b) DS1.



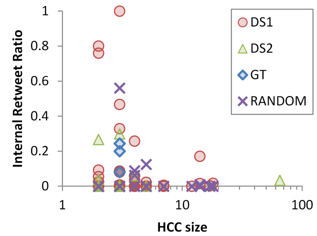
(c) DS2.

Fig. 4. Most used hashtags (per account) of the most active HCCs (γ=15, FSA_V). The labels indicate member and tweet counts. Not all HCCs used a hashtag often enough to be visible.

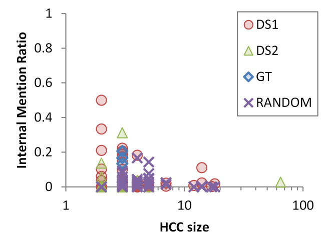
ure 4a), it was possible to delineate subsets of hashtags in use: e.g., one HCC promoted a political narrative in some tweets with `#orangelibs` and discussed cultural events in others with `#adlww`, but was definitely one group.

Examining the Ground Truth. The GT HCCs identified consist only of members within the same political party, across all values of γ. Some accounts appeared in each window size. HCCs of six major parties were identified. Examination of these HCCs' content confirmed they were genuine.

3) Focus of connectivity (RQ4): Groups that retweet or mention themselves create direct connections; therefore to be covert, it would be sensible to have a low *internal retweet* and *mention ratios* (IRR and IMR, respectively). Figure 5 shows IRR and IMR for the datasets. The larger the HCC size, the greater the likelihood of retweeting or mentioning internally, so it is notable that DS2's largest HCC has IRR and IMR's of around 0, though even the smaller HCCs have low ratios. Ratios for the smallest HCCs seem largest, possibly due to low numbers of posts, many of which may be retweets or include a mention, inflating the ratios. The hypothesis that political accounts would retweet and mention themselves frequently is not confirmed by these results, possibly because they are retweeting and mentioning other official or party accounts.



(a) Retweets.



(b) Mentions.

Fig. 5. The proportions of each HCCs retweets and mentions referring to accounts within the HCC (γ=15, FSA_V).

V. Conclusion

We have presented four coordination strategies and a pipeline-based approach to finding groups engaging them in "orchestrated activities" [8], and successfully evaluated them on real-world datasets. Further HCC validation methods, strategy evolution, and real-time applications are to be explored.

References

[1] K. Starbird, A. Arif, and T. Wilson, "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations," *PACMHCI*, vol. 3, no. CSCW, pp. 127:1–127:26, 2019.

[2] A. Bessi and E. Ferrara, "Social bots distort the 2016 U.S. Presidential election online discussion," *First Monday*, vol. 21, no. 11, 2016.

[3] K. Lee, J. Caverlee, Z. Cheng, and D. Z. Sui, "Campaign extraction from social media," *ACM TIST*, vol. 5, no. 1, pp. 9:1–9:28, 2013.

[4] C. Cao, J. Caverlee, K. Lee, H. Ge, and J. Chung, "Organic or organized?: Exploring URL sharing behavior," in *CIKM*. ACM, 2015, pp. 513–522.

[5] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[6] N. Vo, K. Lee, C. Cao, T. Tran, and H. Choi, "Revealing and detecting malicious retweeter groups," in *ASONAM*. ACM, 2017, pp. 363–368.

[7] G. E. Hine, J. Onaolapo, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn, "Kek, cucks, and God Emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the Web," in *ICWSM*. AAAI Press, 2017, pp. 92–101.

[8] C. Grimme, D. Assenmacher, and L. Adam, "Changing perspectives: Is it sufficient to detect social bots?" in *HCI (13)*, ser. Lecture Notes in Computer Science, vol. 10913. Springer, 2018, pp. 445–461.

[9] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea," in *ICWSM*. AAAI Press, 2017, pp. 564–567.

[10] F. Şen, R. T. Wigand, N. Agarwal, S. T. Yuce, and R. Kasprzyk, "Focal structures analysis: Identifying influential sets of individuals in a social network," *Soc. Net. Anal. Min.*, vol. 6, no. 1, pp. 17:1–17:22, 2016.

[11] D. Weber and F. Neumann, "Who's in the gang? Revealing coordinating communities in social media," *CoRR*, vol. abs/2010.08180, 2020.

[12] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Community interaction and conflict on the web," in *WWW*. ACM, 2018, pp. 933–943.

[13] E. Mariconti, J. Onaolapo, S. S. Ahmad, N. Nikiforou, M. Egele, N. Nikiforakis, and G. Stringhini, "What's in a name?: Understanding profile name reuse on Twitter," in *WWW*. ACM, 2017, pp. 1161–1170.

[14] Y. Fang, X. Huang, L. Qin, Y. Zhang, W. Zhang, R. Cheng, and X. Lin, "A survey of community search over big graphs," *The VLDB Journal*, vol. 29, no. 1, pp. 353–392, jul 2019.

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[16] M. Damashek, "Gauging similarity with n-grams: Language-independent categorization of text," *Science*, vol. 267, no. 5199, pp. 843–848, 1995.