Measuring, Characterizing, and Detecting Facebook Like Farms

13

MUHAMMAD IKRAM, UNSW and Data61-CSIRO
LUCKY ONWUZURIKE, University College London
SHEHROZE FAROOQI, University of Iowa
EMILIANO DE CRISTOFARO, University College London
ARIK FRIEDMAN, Atlassian
GUILLAUME JOURJON and MOHAMMED ALI KAAFAR, Data61-CSIRO
M. ZUBAIR SHAFIQ, University of Iowa

Online social networks offer convenient ways to reach out to large audiences. In particular, Facebook pages are increasingly used by businesses, brands, and organizations to connect with multitudes of users worldwide. As the number of likes of a page has become a de-facto measure of its popularity and profitability, an underground market of services artificially inflating page likes *("like farms")* has emerged alongside Facebook's official targeted advertising platform. Nonetheless, besides a few media reports, there is little work that systematically analyzes Facebook pages' promotion methods. Aiming to fill this gap, we present a honeypot-based comparative measurement study of page likes garnered via Facebook advertising and from popular like farms. First, we analyze likes based on demographic, temporal, and social characteristics and find that some farms seem to be operated by bots and do not really try to hide the nature of their operations, while others follow a stealthier approach, mimicking regular users' behavior. Next, we look at fraud detection algorithms currently deployed by Facebook and show that they do not work well to detect stealthy farms that spread likes over longer timespans and like popular pages to mimic regular users. To overcome their limitations, we investigate the feasibility of timeline-based detection of like farm accounts, focusing on characterizing content generated by Facebook accounts on their timelines as an indicator of genuine versus fake social activity. We analyze a wide range of features extracted from timeline posts, which we group into two main categories: lexical and non-lexical. We find that like farm accounts tend to re-share content more often, use fewer words and poorer vocabulary, and more often generate duplicate comments and likes compared to normal users. Using relevant lexical and non-lexical features, we build a classifier to detect like farms accounts that achieves a precision higher than 99% and a 93% recall.

This work was done when the author was with Data61-CSIRO.
The first two authors contributed equally. A preliminary version of this article, titled "Paying for Likes? Understanding Facebook Like Fraud Using Honeypots," appeared in *Proceedings of the 2014 ACM Internet Measurement Conference* (IMC'14). See Section 6 for a summary of the new results presented in this article.
Authors' addresses: M. Ikram, G. Jourjon, and M. A. Kaafar, Australian Technology Park Research Lab, Sydney, Australia, Level 5, 13 Garden Street, Eveleigh, NSW 2015; emails: {Muhammad.Ikam, Guillaume.Jourjon, Dali.Kaafar}@data61csiro.au; A. Friedman, 341 George St, Sydney NSW 2000, Australia; email: arik.friedman@gmail.com; L. Onwuzurike and E. De Cristofaro, University College London, Department of Computer Science, Gower Street, London WC1E 6BT, United Kingdom; emails: {lucky.onwuzurike.13, e.decristofaro}@ucl.ac.uk; S. Farooqi and M. Z. Shafiq, Department of Computer Science, The University of Iowa, Iowa City, IA 52242-1419, Office 317 MacLean Hall, US; emails: {shehroze-farooqi, zubair-shafiq}@uiowa.edu.

https://doi.org/10.1145/3121134

CCS Concepts: • **General and reference** → **Empirical studies**; *Measurement*; • **Information systems** → **Social networks**; **Spam detection**; • **Security and privacy** → **Social network security and privacy**;

Additional Key Words and Phrases: Measurement, like farm detection, social networks, machine learning

ACM Reference format:
Muhammad Ikram, Lucky Onwuzurike, Shehroze Farooqi, Emiliano De Cristofaro, Arik Friedman, Guillaume Jourjon, Mohammed Ali Kaafar, and M. Zubair Shafiq. 2017. Measuring, Characterizing, and Detecting Facebook Like Farms. *ACM Trans. Priv. Secur.* 20, 4, Article 13 (September 2017), 28 pages.
https://doi.org/10.1145/3121134

1 INTRODUCTION

Online social networks provide organizations and public figures with a range of tools to reach out to, as well as broaden, their audience. Among these, *Facebook pages* make it easy to broadcast updates, publicize products and events, and get in touch with customers and fans. Facebook allows page owners to promote their pages via targeted advertisement, that is, pages can be "suggested" to users from specific age or location groups or with certain interests. Page ads constitute one of the primary sources of revenue for Facebook, as its advertising platform overall is reportedly used by 2 million small businesses of the 40 million that have active pages (Snyder 2015).

At the same time, as the number of likes on a Facebook page is considered a measure of its popularity (Carter 2013), an ecosystem of *"like farms"* has emerged that offers paid services to artificially inflate the number of likes on Facebook pages. These farms rely on fake and compromised accounts as well as incentivized collusion networks where users are paid for actions from their account (Viswanath et al. 2014). Popular media reports (Arthur 2013; Cellan-Jones 2012; Metzger 2012; Muller. 2014; Schneider 2014) have speculated that Facebook ad campaigns may also garner significant amounts of fake likes, due to farm accounts' attempt to diversify liking activities and avoid Facebook's fraud detection algorithms. With the price charged for 1,000 likes varying from $14.99–$70 for worldwide users to $59.95–$190 for USA users, it is not far fetched to assume that selling likes may yield significant profits for fraudsters. This also creates potential problems for providers like Facebook, as they lose potential ad revenues while possibly disenfranchising page owners who receive likes from users who do not engage with their page. However, even though the understanding of fake likes is crucial to improve fraud mitigation in social networks, there has been little work to systematically analyze and compare Facebook page promotion methods. With this motivation in mind, we set to shed light on the like farming ecosystem with the aim of characterizing features and behaviors that can be useful to effectively detect them. In the process, we review the fraud detection tools currently deployed by Facebook and assess their efficacy for more sophisticated like farms.

Specifically, our article makes three main contributions: (1) We present a first-of-its-kind honeypot-based comparative measurement study of page likes garnered via Facebook ads and like farms and analyze likes based on demographic, temporal, and social characteristics; (2) we perform an empirical analysis of graph-based fraud detection tools used by Facebook and highlight their shortcomings against more sophisticated farms; and (3) we propose and evaluate timeline-based detection of like farm accounts, focusing on characterizing content as an indicator of genuine versus fake social activity, and build a classifier, based on lexical and non-lexical features, that detects like farm accounts with at least 99% precision and 93% recall.

1.1 Roadmap

Honeypot-based measurement of like farms. Aiming to study fake likes garnered from like farms and, potentially, from Facebook advertising, we have create 13 Facebook *honeypot* pages with

the description: "This is not a real page, so please do not like it" and intentionally kept them empty (i.e., no posts or pictures). We have promoted 8 of them using four like farms (i.e., targeting users in the USA and worldwide for each, as farms mostly offer user targeting for only this two locations) and 5 using Facebook ad campaigns (with two targeting users in the USA and worldwide as the like farms). The other three target one developed and two developing countries, as Facebook reports that "false" accounts are less prevalent in developed markets and more in developing markets.[1] After monitoring likes garnered by the pages, and collecting information about the likers (e.g., gender, age, location, friend list, etc.), we perform a comparative analysis based on demographic, temporal, and social characteristics.

We identify two main *modi operandi* for the like farms: (1) Some seem to be operated by bots and do not really try to hide their activities, delivering likes in bursts and forming disconnected social sub-graphs, while (2) others follow a stealthier approach, mimicking regular users' behavior, and rely on a large and well-connected network structure to gradually deliver likes while keeping a small count of likes per user. The first strategy reflects a "quick and dirty" approach where likes from fake users are delivered rapidly, as opposed to the second one, which exhibits a stealthier approach that leverages the underlying social graph, where accounts (possibly operated by real users) slowly deliver likes. We also highlight a few more interesting findings. When targeting Facebook users worldwide, we obtain likes from only a few countries and that likers' profiles seem skewed toward males. Moreover, we find evidence that different like farms (with different pricing schemes) garner likes from overlapping sets of users and, thus, may be managed by the same operator.

Characterizing fake likes. We present the concept of liking a page on Facebook as a binary action where likes received on a page by users who have interest for the content of the page are considered "good" and likes received to manipulate a page's popularity ranking as "fake." We have only considered to mark likes that are meant to manipulate a page's popularity as fake (i.e., by increasing the page's number of fans), as this is the main purpose of like farms. On this note, we start our study with the assumption that likers from farms that like our empty honeypot pages are either fake or compromised real users (i.e., fake likes) as shown in Viswanath et al. (2014). Although, Facebook discourages page owners from buying fake likes, warning that they *"can be harmful to your page,"*[2] they also routinely launch clean-up campaigns to remove fake accounts, including those engaged in like farms. Hence, we also hypothesize that very few or no users from the Facebook ad campaigns will like our honeypot pages, as the pages were empty.

Aiming to counter like farms, researchers as well as Facebook have recently been working on tools to detect fake likes. One currently deployed tool is CopyCatch, which detects lockstep page like patterns by analyzing the social graph among users and pages and the times at which the edges in the graph are created (Beutel et al. 2013). Another one, SynchroTrap, relies on the fact that malicious accounts usually perform loosely synchronized actions in a variety of social network context and can cluster malicious accounts that act similarly at around the same time for a sustained period of time (Cao et al. 2014). The issue with these methods, however, is that stealthier (and more expensive) like farms can successfully circumvent them by spreading likes over longer timespans and liking popular pages to mimic normal users.

We systematically evaluate the effectiveness of these graph-based co-clustering fraud detection algorithms (Beutel et al. 2013; Cao et al. 2014) in identifying like farm accounts. We show that

[1] https://goo.gl/OAxgTh (accessed on January 31, 2017).
[2] See https://www.facebook.com/help/241847306001585 (accessed on July 18, 2016).

these tools incur high false-positives rates for stealthy farms, as their accounts mimic normal users.

Characterizing lexical and non-lexical timeline information. Next, we investigate the use of timeline information, including lexical and non-lexical characteristics of user posts, to improve the detection of like farm accounts. To this end, we crawl and analyze timelines of user accounts associated with like farms as well as a baseline of normal user accounts. Our analysis of timeline information highlights several differences in both lexical and non-lexical features of baseline and like farm users. In particular, we find that timeline posts by like farm accounts have 43% fewer words, a more limited vocabulary, and lower readability than normal users' posts. Moreover, like farm accounts' posts generate significantly more comments and likes, and a much larger fraction of their posts consists of "shared activity" (i.e., sharing posts from other users, news articles, videos, and external URLs).

Detection. Based on our characterization, we extract a set of timeline-based features and use them to train three classifiers using supervised two-class support vector machines (SVM) (Müller et al. 2001). Our first and second classifiers use, respectively, lexical and non-lexical features extracted from timeline posts, while the third one uses both. We evaluate the classifiers using the ground-truth dataset of like farm accounts and show that they achieve 99–100% precision and 93–97% recall in detecting like farm accounts. Finally, we generalize our approach using other classification algorithms, namely, decision tree (Breiman et al. 1984), AdaBoost (Freund and Schapire 1997), kNN (Andoni and Indyk 2008), random forest (Breiman 2001), and naïve Bayes (Zhang 2004), and empirically confirm that the SVM classifier achieves higher accuracy across the board.

1.2 Paper Organization

The rest of the article is organized as follows. Section 2 presents our honeypot-based comparative measurement of likes garnered using farms and legitimate Facebook ad campaigns. Then, Section 3 evaluates the accuracy of state-of-the-art co-clustering techniques to detect like farm accounts in our datasets. Next, we study timeline-based features (both non-lexical and lexical) in Section 4 and evaluate the classifiers built using these features in Section 5. After reviewing related work in Section 6, the article concludes in Section 7.

2 HONEYPOT-BASED MEASUREMENT OF FACEBOOK LIKE FARMS

This section details our honeypot-based comparative measurement study of page likes garnered via Facebook ads and by like farms.

2.1 Datasets

In the following, we present the methodology used to deploy, monitor, and promote our Facebook honeypot pages.

Honeypot pages. In March 2014, we created 13 Facebook pages called "Virtual Electricity" and intentionally kept them empty (i.e., no posts or pictures). Their description included the following text: *"This is not a real page, so please do not like it."* Five pages were promoted using legitimate Facebook (FB) ad campaigns targeting users, respectively, in USA, France, India, Egypt, and worldwide. The remaining 8 pages were promoted using four popular like farms: BoostLikes.com (BL), SocialFormula.com (SF), AuthenticLikes.com (AL), and MammothSocials.com (MS), targeting worldwide or USA users.

In Table 1, we provide details of the honeypot pages, along with the corresponding ad campaigns. All campaigns were launched on March 12, 2014, using a different administrator account (owner)

Table 1. Facebook and Like Farm Campaigns Used to Promote the Facebook Honeypot Pages. Like Farms Promised to Deliver 1000 Likes in 15 Days at Differing Prices Depending on the Geographical Target (i.e., USA and Worldwide), Whereas on Facebook, We Budgeted $6 Per Day for the Promotion of Each Page for a Period of 15 Days

Campaign ID	Provider	Location	Budget	Duration	Monitoring	#Likes	#Terminated
FB-USA	Facebook.com	USA	$6/day	15 days	22 days	32	0
FB-FRA	Facebook.com	France	$6/day	15 days	22 days	44	0
FB-IND	Facebook.com	India	$6/day	15 days	22 days	518	2
FB-EGY	Facebook.com	Egypt	$6/day	15 days	22 days	691	6
FB-ALL	Facebook.com	Worldwide	$6/day	15 days	22 days	484	3
BL-ALL	BoostLikes.com	Worldwide	$70.00	15 days	—	—	—
BL-USA	BoostLikes.com	USA only	$190.00	15 days	22 days	621	1
SF-ALL	SocialFormula.com	Worldwide	$14.99	3 days	10 days	984	11
SF-USA	SocialFormula.com	USA	$69.99	3 days	10 days	738	9
AL-ALL	AuthenticLikes.com	Worldwide	$49.95	3–5 days	12 days	755	8
AL-USA	AuthenticLikes.com	USA	$59.95	3–5 days	22 days	1038	36
MS-ALL	MammothSocials.com	Worldwide	$20.00	—	—	—	—
MS-USA	MammothSocials.com	USA only	$95.00	—	12 days	317	9

for each page. Each Facebook campaign was budgeted at a maximum of $6/day to a total of $90 for 15 days. The price for buying likes varied across like farms: BoostLikes charged the highest price for "100% real likes" ($70 and $190 for 1000 likes in 15 days from, respectively, worldwide and USA). Other like farms also claimed to deliver likes from "genuine," "real," and "active" profiles but promised to deliver them in fewer days. Overall, the price of 1,000 likes varied between $14.99 and $70 for worldwide users and $59.95 and $190 for USA users.

Data collection. We monitored the "liking" activity on the honeypot pages by crawling them every 2 hours using Selenium web driver. At the end of the campaigns, we reduced the frequency of monitoring to once a day and stopped monitoring when a page did not receive a like for more than a week. We used Facebook's reports tool for page administrators, which provides a variety of aggregated statistics about attributes and profiles of page likers. Facebook also provides these statistics for the global Facebook population. Since a majority of Facebook users do not set the visibility of their age and location to public (Chaabane et al. 2012), we used these reports to collect statistics about likers' gender, age, country, home, and current town. Later in this section, we will use these statistics to compare distributions of our honeypot pages' likers to that of the overall Facebook population. We also crawled public information from the likers' profiles, obtaining the lists of liked pages as well as friend lists, which are not provided in the reports. Overall, we identify more than 6.3 million total likes by users who liked our honeypot pages and more than 1 million friendship relations.

Campaign summary. In Table 1, we report the total number of likes garnered by each campaign, along with the number of days we monitored the honeypot pages. Note that the BL-ALL and MS-ALL campaigns remained inactive, that is, they did not result in any likes even though we were charged in advance. We tried to reach the like farm admins several times but received no response. Overall, we collected a total of 6,222 likes (4,453 from like farms and 1,769 from Facebook ads). The largest number of likes were garnered by AL-USA and the lowest (excluding inactive campaigns) by FB-USA.

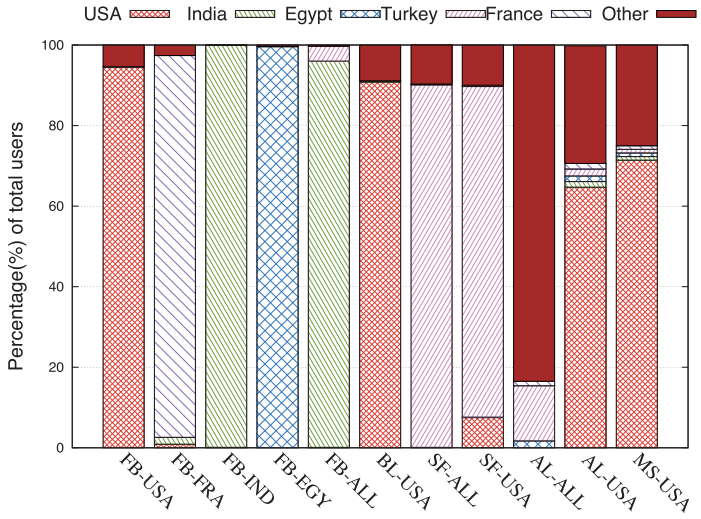


Fig. 1. Geolocation of the likers (per campaign).

Ethics considerations. Although we only collected openly available data, we did collect (public) profile information from our honeypot pages' likers, for example, friend lists and page likes. We could not request consent but enforced a few mechanisms to protect user privacy: All data were encrypted at rest and not re-distributed, and no personal information was extracted, that is, we only analyzed aggregated statistics. We are also aware that paying farms to generate fake likes might raise ethical concerns; however, this was crucial to create the honeypots and observe the like farms' behavior. We believe that the study will help, in turn, to understand and counter these activities. Also note that the amount of money each farm received was small ($190 at most) and that this research was reviewed and approved by Data61's legal team. We also received ethical approval from the ethics committee of UCL where, in conjunction with Data61, data were collected and analyzed.

2.2 Location and Demographics Analysis

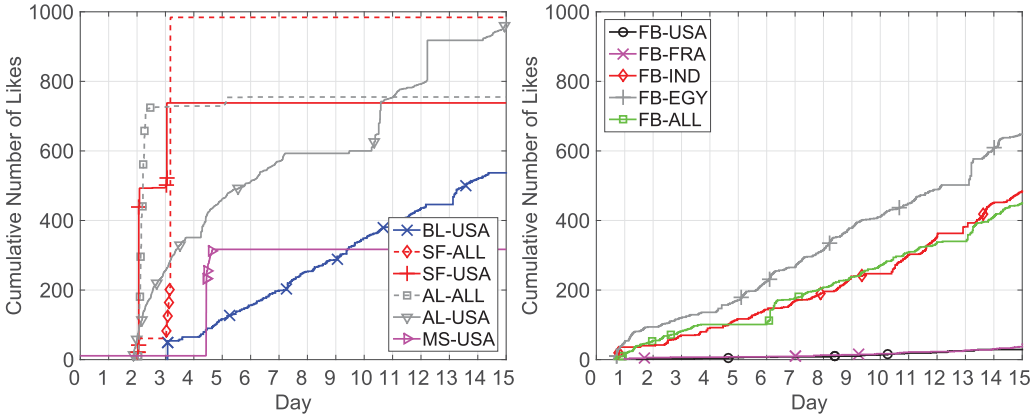
We now set to compare the characteristics of the likes garnered by the honeypot pages promoted via legitimate Facebook campaigns and those obtained via like farms.

Location. For each campaign, we looked at the distribution of likers' countries: As shown in Figure 1, for the first four Facebook campaigns (FB-USA, FB-FRA, FB-IND, and FB-EGY), we mainly received likes from the targeted country (87–99.8%), even though FB-USA and FB-FRA generated far fewer likes than any other campaign. When we targeted Facebook users worldwide (FB-ALL), we almost exclusively received likes from India (96%). Looking at the like farms, most likers from SocialFormula were based in Turkey, regardless of whether we requested a US-only campaign. The other three farms delivered likes complying to our requests, for example, for US-only campaigns, the pages received a majority of likes from US profiles. The location result supports Facebook's claim that "the percentage of accounts that are duplicate or false is meaningfully lower in developed markets such as the United States or United Kingdom and higher in developing markets such as India and Turkey."[3] It also potentially supports the claim that like farm accounts diversify their

[3]https://goo.gl/OAxgTh (accessed on January 31,, 2017).

Table 2. Gender and Age Statistics of Likers

Campaign ID	Gender % F/M	Age Distribution (%) 13–17	18–24	25–34	35–44	45–54	55+	KL
FB-USA	54/46	54.0	27.0	6.8	6.8	1.4	4.1	0.45
FB-FR	46/54	60.8	20.8	8.7	2.6	5.2	1.7	0.54
FB-IND	**7/93**	52.7	43.5	2.3	0.7	0.5	0.3	**1.12**
FB-EGY	**18/82**	54.6	34.4	6.4	2.9	0.8	0.8	**0.64**
FB-ALL	**6/94**	51.3	44.4	2.1	1.1	0.5	0.6	**1.04**
BL-USA	53/47	34.2	54.5	8.8	1.5	0.7	0.5	0.60
SF-ALL	37/63	19.8	33.3	21.0	15.2	7.2	2.8	0.04
SF-USA	37/63	22.3	34.6	22.9	11.6	5.4	2.9	0.04
AL-ALL	42/58	15.8	52.8	13.4	9.7	5.2	3.0	0.12
AL-USA	31/68	7.2	41.0	35.0	10.0	3.5	2.8	0.09
MS-USA	26/74	8.6	46.9	34.5	6.4	1.9	1.4	0.17
Facebook	46/54	14.9	32.3	26.6	13.2	7.2	5.9	—



(a) Like Farms (b) Facebook Campaigns

Fig. 2. Time series of cumulative number of likes for Facebook and like farms campaigns.

liking activities by liking pages promoted via Facebook ads to avoid Facebook's fraud detection algorithms (we further explore this in Section 2.5).

Other demographics. In Table 2, we show the distribution of likers' gender and age and also compare them to the global Facebook network (last row). The last column reports the KL -divergence between the age distribution of the campaign users and that of the entire Facebook population, highlighting large divergence for FB-IND, FB-EGY, and FB-ALL, which are biased toward younger users. These three campaigns also appear to be skewed toward male profiles. In contrast, the demographics of likers from SocialFormula and, to a lesser extent, AuhtenticLikes and MammothSocials, are much more similar to those of the entire network, even though male users are still over-represented.

2.3 Temporal Analysis

We also analyzed temporal patterns observed for each of the campaigns. In Figure 2, we plot the cumulative number of likes observed on each honeypot page over our observation period (15 days).

Table 3. Likers and Friendships Between Likers

Provider	#Likers	#Likers with Public Friend Lists	Avg (± Std) #Friends	Median #Friends	#Friendships Between Likers	#Two-Hop Friendship Relations Between Likers
FB	1448	261 (18.0%)	315 ± 454	198	6	169
BL	621	161 (25.9%)	1171 ± 1096	850	540	2987
SF	1644	954 (58.0%)	246 ± 330	155	50	1132
AL	1597	680 (42.6%)	719 ± 973	343	64	1174
MS	121	62 (51.2%)	250 ± 585	68	4	129
ALMS	213	101 (47.4%)	426 ± 961	46	27	229

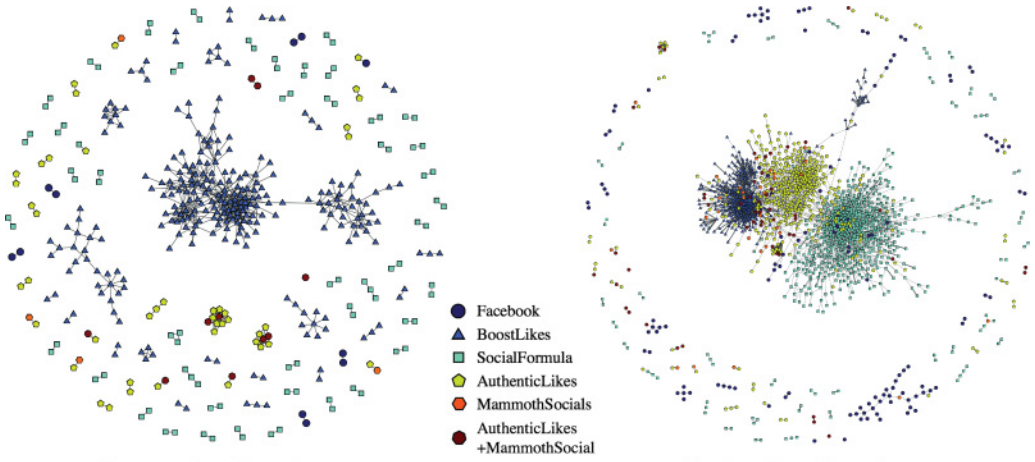
We observe from Figure 2(a) that all the like farm campaigns, except BoostLikes, exhibit a very similar trend with a few bursts of a large number of likes. Specifically, for the SocialFormula, AuthenticLikes, and MammothSocials campaigns, likes were garnered within a short period of time of 2 hours. With AuthenticLikes, we observed likes from more than 700 profiles within the first 4 hours of the second day of data collection. Interestingly, no more likes were observed later. On the contrary, the BoostLikes campaign targeting US users shows a different temporal behavior: The trend is actually comparable to that observed in the Facebook ads campaigns (see Figure 2(b)). The number of likes steadily increases during the observation period, and no abrupt changes are observed.

This suggests that two different strategies may be adopted by like farms. On the one hand, the abrupt increase in the cumulative number of likes happening during a short period of time might likely be due to automated scripts operating a set of fake profiles. These profiles are instrumented to satisfy the number of likes as per the customer's request. On the other hand, BoostLikes's strategy, which resembles the temporal evolution in Facebook campaigns, seems to rely on the underlying social graph, possibly constituted by fake profiles operated by humans. Results presented in the next section corroborate the existence of these two strategies.

2.4 Social Graph Analysis

Next, we evaluated the social graph induced by the likers' profiles. To this end, we associated each user with one of the like farm services based on the page they liked. Note that a few users liked pages in multiple campaigns, as we will discuss in Section 2.5. A significant fraction of users actually liked pages corresponding to both the AuthenticLikes and the MammothSocials campaigns (see Figure 5): We put these users into a separate group, labelled as ALMS. Table 3 summarizes the number of likers associated with each service, as well as additional details about their friendship networks. Note that the number of likers reported for each campaign in Table 3 is different from the number of campaign likes (Table 1), since some users liked more than one page.

Many likers kept their friend lists private: This occurred for almost 80% of likers in the Facebook campaigns, about 75% in the BoostLikes campaign, and much less frequently for the other like farm campaigns (~40–60%). The number and percentage of users with public friend lists are reported in Table 3. The fourth column reports the average number of friends (± the standard deviation) for profiles with visible friend lists, and the fifth column reports the median. Some friendship relations may be hidden, for example, if a friend chose to be invisible in friend lists, thus, these numbers only represent a *lower bound*. The average number of friends of users associated with the BoostLikes campaign (and to a smaller extent, the AuthenticLikes campaign) was much higher than the average number of friends observed elsewhere.



(a) Direct friendship relations (b) 2-hop friendship relations

Fig. 3. Friendship relations between likers of different campaigns.

To evaluate the social ties between likers, we looked at friendship relations between likers (either originating from the same campaign provider or not), ignoring friendship relations with Facebook users who did not like any of our pages. Table 3 (sixth column) reports, for each provider, the overall number of friendship relationships between likers that involved users associated with the provider.

In Figure 3(a), we plot the social graph induced by such friendship relations (likers who did not have friendship relations with any other likers were excluded from the graph). Based on the resulting social structure, we suggest that:

(1) Dense relations between likers from BoostLikes point to an interconnected network of real users, or fake users who mimic complex ties to pose as real users;
(2) The pairs (and occasionally triplets) that characterize SocialFormula likers might indicate a different strategy of constructing fake networks, mitigating the risk that identification of a user as fake would consequently bring down the whole connected network of fake users; and
(3) The friendship relations between AuthenticLikes and MammothSocials likers might indicate that the same operator manages both services.

We also considered indirect links between likers, through mutual friends. Table 3 reports the overall number of two-hop relationships between likers from the associated provider. Figure 3(b) plots the relations between likers who either have a direct relation or a mutual friend, clearly pointing to the presence of relations between likers from the same provider. These tight connections, along with the number of their friends, suggest that we only see a small part of these networks. For SocialFormula, AuthenticLikes, and MammothSocials, we also observe many isolated pairs and triplets of likers who are not connected. One possible explanation is that farm users create fake Facebook accounts and keep them separate from their personal accounts and friends. In contrast, the BoostLikes network is well connected.

To further compare connectivity of BoostLikes versus SocialFormula, AuthenticLikes, and MammothSocials, we analyze the structural properties of the social graph visualized in Figure 3(b). Figure 4 plots distributions of degree, number of triangles, clustering coefficient, and cliques for these like farms. The distributions demonstrate that BoostLikes accounts have dense connectivity

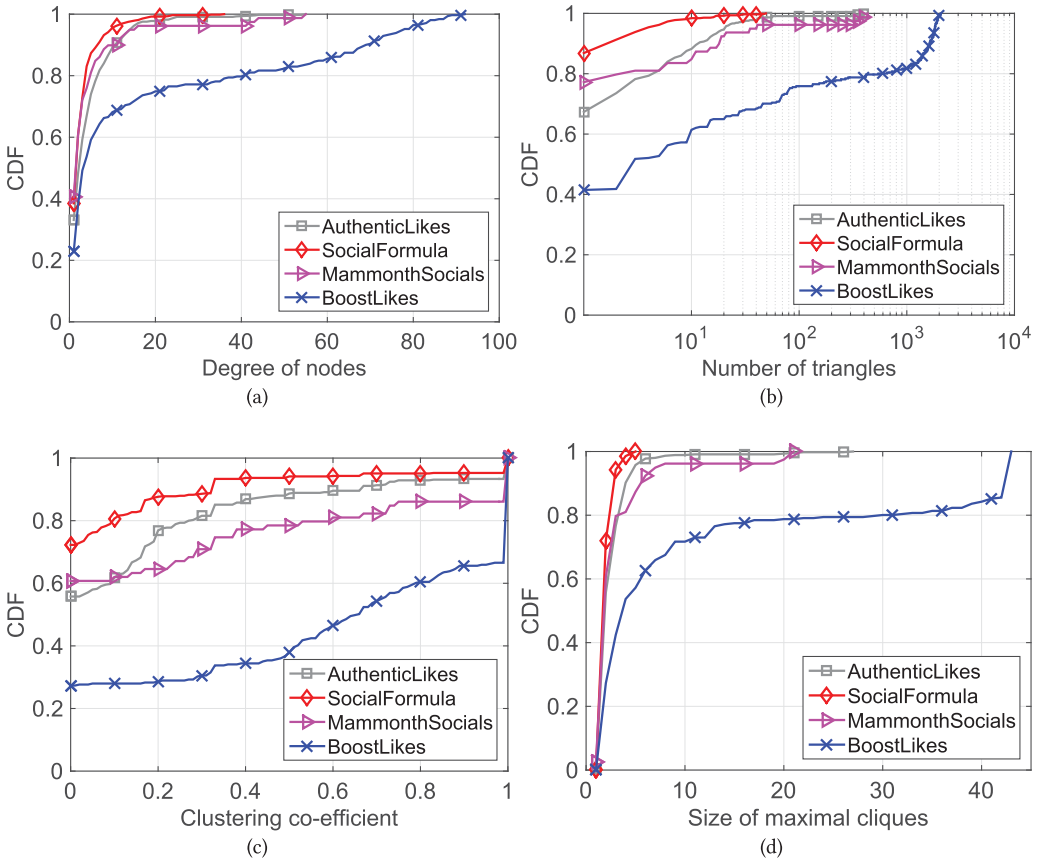


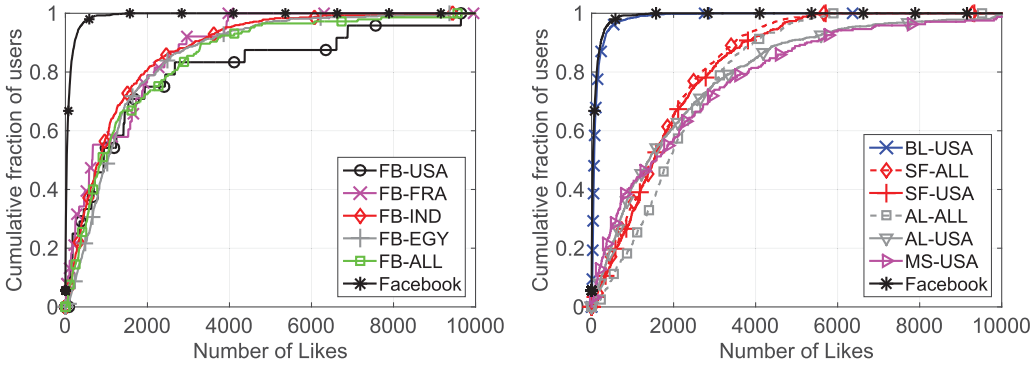
Fig. 4. Structural properties of the graph of two-hop relationships among likers of like farm campaigns.

as compared to accounts belonging to SocialFormula, AuthenticLikes, and MammothSocials. More specifically, BoostLikes accounts have higher degree, are part of more triangles, have higher clustering coefficient, and have larger maximal cliques than other like farms. For example, the average degree of BoostLikes accounts is 18 while other like farms have average degrees of less than 5. Moreover, more than 25% of BoostLikes accounts make maximal cliques of size greater than 10 while less than 1% accounts of the other like farms make maximal cliques of size greater than 10.

2.5 Page Like Analysis

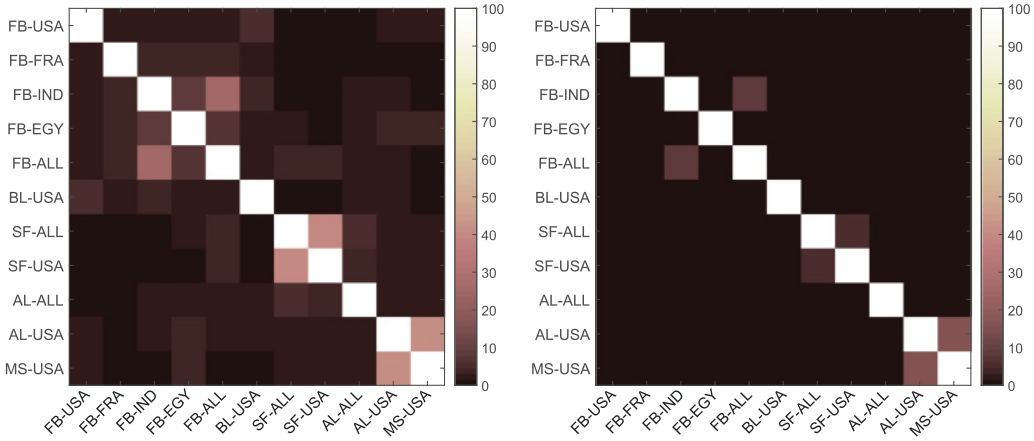
We then looked at the *other* pages liked by profiles attracted to our honeypot pages. In Figure 5(a) and 5(b), respectively, we plot the distribution of the number of page likes for Facebook ads' and like farm campaigns' users. To draw a baseline comparison, we also collected page like counts from a random set of 2,000 Facebook users, extracted from an unbiased sample of Facebook user population. The original sample was crawled for another project (Chen et al. 2013), obtained by randomly sampling Facebook public directory that lists all the IDs of searchable profiles.

We observed a large variance in the number of pages liked, ranging from 1 to 10,000. The median page like count ranged between 600 and 1,000 for users from the Facebook campaigns and between 1,200 and 1,800 for those from like farm campaigns, with the exception of the BL-USA campaign (median was 63). In contrast, the median page like count for our baseline Facebook user



(a) Facebook Campaigns (b) Like Farms

Fig. 5. Distribution of the number of likes by users in Facebook and like farm campaigns.



(a) Page Like (b) User

Fig. 6. Jaccard index similarity (×100) matrices of page likes and likers across different campaigns.

sample was 34. The page like counts of our baseline sample mirrored numbers reported in prior work, for example, according to Lafferty (2013), the average number of pages liked by Facebook users amounts to roughly 40. In other words, our honeypot pages attracted users that tend to like significantly more pages than regular Facebook users. Since our honeypot pages both for Facebook and like farm campaigns explicitly indicated they were not "real," we argue that a vast majority of the garnered likes are fake. We argue that these users like a large number of pages because they are probably reused for multiple "jobs" and also like "normal" pages to mimic real users.[4]

To confirm our hypothesis, for each pair of campaigns, we plot their Jaccard similarity. Specifically, let S_k denote the set of pages liked by a user k: The Jaccard similarity between the set of likes by likers of two campaigns A and B, which we plot in Figure 6(a), is defined as $|A \cap B|/|A \cup B|$, where $A = \bigcup_{\forall i \in A} S_i$ and $B = \bigcup_{\forall j \in B} S_j$. We also plot, in Figure 6(b), the similarity between $A' = \bigcup_{\forall i \in A} i$ and $B' = \bigcup_{\forall j \in B} j$, that is, the similarity between the set of likers of the different campaigns.

[4]Facebook does not impose any limit on the maximum number of page likes per user.

Note from Figure 6 that FB-IND, FB-EGY, and FB-ALL have relatively large (Jaccard) similarity with each other. In addition, the SF-USA and SF-ALL pair and the AL-USA and MS-USA pair also have relatively large Jaccard similarity. These findings suggest that the same fake profiles are used in multiple campaigns by a like farm (e.g., SF-ALL and SF-USA). Moreover, some fake profiles seem to be shared by different like farms (e.g., AL-USA and MS-USA), suggesting that they are run by the same operator.

2.6 Discussion

Overall, we identified two main *modi operandi*: (1) some farms, like SocialFormula and AuthenticLikes, seem to be operated by bots and do not really try to hide the nature of their operations, as demonstrated by large bursts of likes and the limited number of friends per profile; (2) other farms, like BoostLikes, follow a much stealthier approach, aiming to mimic regular users' behavior, and rely on their large and well-connected network structure to disseminate the target likes while keeping a small count of likes per user. For the latter, we also observed a high number of friends per profile and a "reasonable" number of likes.

A month after the campaigns, we checked whether or not likers' accounts were still active: As shown in Table 1, only one account associated with BoostLikes was terminated, as opposed to 9, 20, and 44 for the other like farms. Eleven accounts from the regular Facebook campaigns were also terminated. Although occurring not so frequently, the accounts' termination might be indicative of the disposable nature of fake accounts on most like farms, where "bot-like" patterns are actually easy to detect. It also mirrors the challenge Facebook is confronted by, with like farms such as BoostLikes that exhibit patterns closely resembling real users' behavior, thus making fake like detection quite difficult.

We stress that our findings do not necessarily imply that advertising on Facebook is ineffective, since our campaigns were specifically designed to avert real users. However, we do provide strong evidence that likers attracted to our honeypot pages, even when using legitimate Facebook campaigns, are significantly different from typical Facebook users, which confirms the concerns about the genuineness of these likes. We also show that most fake likes exhibit some peculiar characteristics—including demographics, likes, and temporal and social graph patterns—that can and should be exploited by like fraud detection algorithms.

3 LIMITATIONS OF GRAPH CO-CLUSTERING TECHNIQUES

Aiming to counter fraudulent activities, including like farms, Facebook has recently deployed detection tools such as CopyCatch (Beutel et al. 2013) and SynchroTrap (Cao et al. 2014). These tools use graph co-clustering algorithms to detect large groups of malicious accounts that like similar pages around the same time frame. However, as shown in Section 2, some stealthy like farms seem to deliberately modify their behavior to avoid synchronized patterns, which might reduce the effectiveness of these detection tools. Specifically, while several farms use a large number of accounts (possibly fake or compromised) liking target pages within a short timespan, some spread likes over longer timespans and onto popular pages aiming to circumvent fraud detection algorithms. In this section, we analyze the efficacy of state-of-the-art co-clustering algorithms on our dataset of like farm users.

3.1 Re-Crawling

Our experiments use, as ground truth, the Facebook accounts gathered as part of the honeypot-based measurement of like farms. Recall (from Section 2) that we garnered 5,918 likes from 5,616 unique users—specifically, 1,437 unique accounts from Facebook ad campaigns and 4,179 unique accounts from the like farm campaigns (note that some users liked more than one honeypot pages).

Table 4. Overview of the Datasets Used in Our Study

Campaign	#Users	#Pages Liked	#Pages Liked (Unique)	#Posts
BL-USA	583	79,025	37,283	44,566
SF-ALL	870	879,369	108,020	46,394
SF-USA	653	340,964	75,404	38,999
AL-ALL	707	162,686	46,230	61,575
AL-USA	827	441,187	141,214	30,715
MS-USA	259	412,258	141,262	12,280
Tot. Farms	*3,899*	*2,315,489*	*549,413*	*234,529*
Baseline	1,408	79,247	57,384	34,903

In summer 2015, we checked how many accounts had been closed or terminated and found that 624 of 5,616 accounts (11%) were no longer active. We then began to crawl the pages liked by each of the 4,179 like farm users (again, using Selenium web driver). We collected basic information associated with each page, such as the total number of likes, category, and location, using the page identifier. Unlike our previous crawl, we now also collected the timelines of the like farm accounts—specifically, timeline posts (up to a maximum of 500 recent posts), the comments on each post, as well as the associated number of likes and comments on each post.

Besides some accounts having become inactive (376), we also could not crawl the timeline of 24 users who had restricted the visibility of their timeline. Moreover, in fall 2015, Facebook blocked all the accounts we were using for crawling, and so we stopped our data collection before we could completely finish our data collection, and, hence, we missed an additional 109 users. In summary, our new dataset consists of 3,670 users (of the initial 4,179), with more than 234K posts (messages, shared content, check-ins, etc.) for these accounts. In our experiments, we will also rely on a baseline of 1,408 random accounts from Chen et al. (2013) that we use to form a baseline of "normal" accounts. For each of these accounts, we again collected posts from their timeline, their page likes, and information from these pages. 53% of the accounts had at least 10 visible posts on the timeline, and in total we collected about 35K posts.

Table 4 summarizes the data used in the experiments presented in the rest of the article. Note that users who like more than one honeypot pages are included in all rows, hence the disparity between the number of unique users (3,670) and the total reported in the table (3,899). Overall, we gathered information from 600K unique pages, liked by 3,670 like farm accounts and 1,408 baseline accounts, and around 270K posts.

Again, note that we collected openly available data such as (public) profile and timeline information, as well as page likes. Also, all data were encrypted at rest and have not been re-distributed. No personal information was extracted, as we only analyzed aggregated statistics. We also consulted Data61's legal team, which classified our research as exempt and, likewise, received approval from the ethics committee of UCL.

3.2 Experimental Evaluation of Co-Clustering

We use the labeled dataset of 3,670 users from six different like farms and the 1,408 baseline users and employ a graph co-clustering algorithm to divide the user-page bipartite graph into distinct clusters (Kluger et al. 2003). Similarly to CopyCatch (Beutel et al. 2013) and SynchroTrap (Cao et al. 2014), the clusters identified in the user-page bipartite graph represent near-bipartite cores, and the set of users in a near-bipartite core like the same set of pages. Since we are interested in distinguishing between two classes of users (like farm users and normal users), we set the target

Table 5. Effectiveness of the Graph Co-clustering Algorithm

Campaign	TP	FP	TN	FN	Precision	Recall	F1-Score
AL-USA	681	9	569	4	98%	99%	99%
AL-ALL	448	53	527	1	89%	99%	94%
BL-USA	523	588	18	0	**47%**	100%	**64%**
SF-USA	428	67	512	1	86%	100%	94%
SF-ALL	431	48	530	2	90%	99%	95%
MS-USA	201	22	549	2	90%	99%	93%

number of clusters at 2. Given that our crawlers were restricted to crawl the behavior of all like farms and baseline users on daily basis, we do not have fine-grained features to further analyze CopyCatch and SynchroTrap. Aiming to reveal the liking behavior of like farms users, we evaluate the employed graph co-clustering schemes of CopyCatch and SynchroTrap on our collected datasets. Moreover, our analysis has a limitation that it relies on the data of a small number of like farm accounts.

Results. In Table 5, we report the receiver operating characteristic (ROC) statistics of the graph co-clustering algorithm—specifically, true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), Precision: $(TP)/(TP + FP)$, Recall: $(TP)/(TP + FN)$, and F1-Score, that is, the harmonic average of precision and recall. Figure 7 visualizes the clustering results as user-page scatter plots. The x-axis represents the user index and the y-axis the page index.[5] The vertical black line marks the separation between two clusters. The points in the scatter plot are colored to indicate true positives (green), true negatives (blue), false positives (red), and false negatives (black).

Analysis. We observe two distinct behaviors in the scatter plots: (1) *"liking everything"* (vertical streaks) and (2) *"everyone liking a particular page"* (horizontal streaks). Both like farms and normal users exhibit vertical and horizontal streaks in the scatter plots.

While the graph co-clustering algorithm neatly separates users for AL-USA, it incurs false positives for other like farms. In particular, the co-clustering algorithm fails to achieve a good separation for BL-USA, where it incurs a large number of false positives, resulting in 47% precision. Further analysis reveals that the horizontal false positive streaks in BL-USA include popular pages, such as "Fast & Furious" and "SpongeBob SquarePants," each with millions of likes. We deduce that stealthy like farms, such as BL-USA, use the tactic of liking popular pages aiming to mimic normal users, which reduces the accuracy of the graph co-clustering algorithm.

Our results highlight the limitations of prior graph co-clustering algorithms in detecting fake likes by like farm accounts. We argue that fake liking activity is challenging to detect when only relying on monitoring the liking activity due to the increased sophistication of stealthier like farms. Therefore, as we discuss next, we plan to leverage the characteristics of timeline features to improve accuracy.

4 CHARACTERIZING TIMELINE FEATURES

Motivated by the poor accuracy of graph co-clustering based detection tools on stealthy farms, we set to evaluate the feasibility of timeline-based detection of like farm accounts. To this end, we characterize timeline activities for users in our datasets (cf. Section 3.1) with respect to two categories of features, *non-lexical* and *lexical*, aiming to identify the most distinguishing features

[5]To ease presentation, we exclude users and pages with fewer than 10 likes.

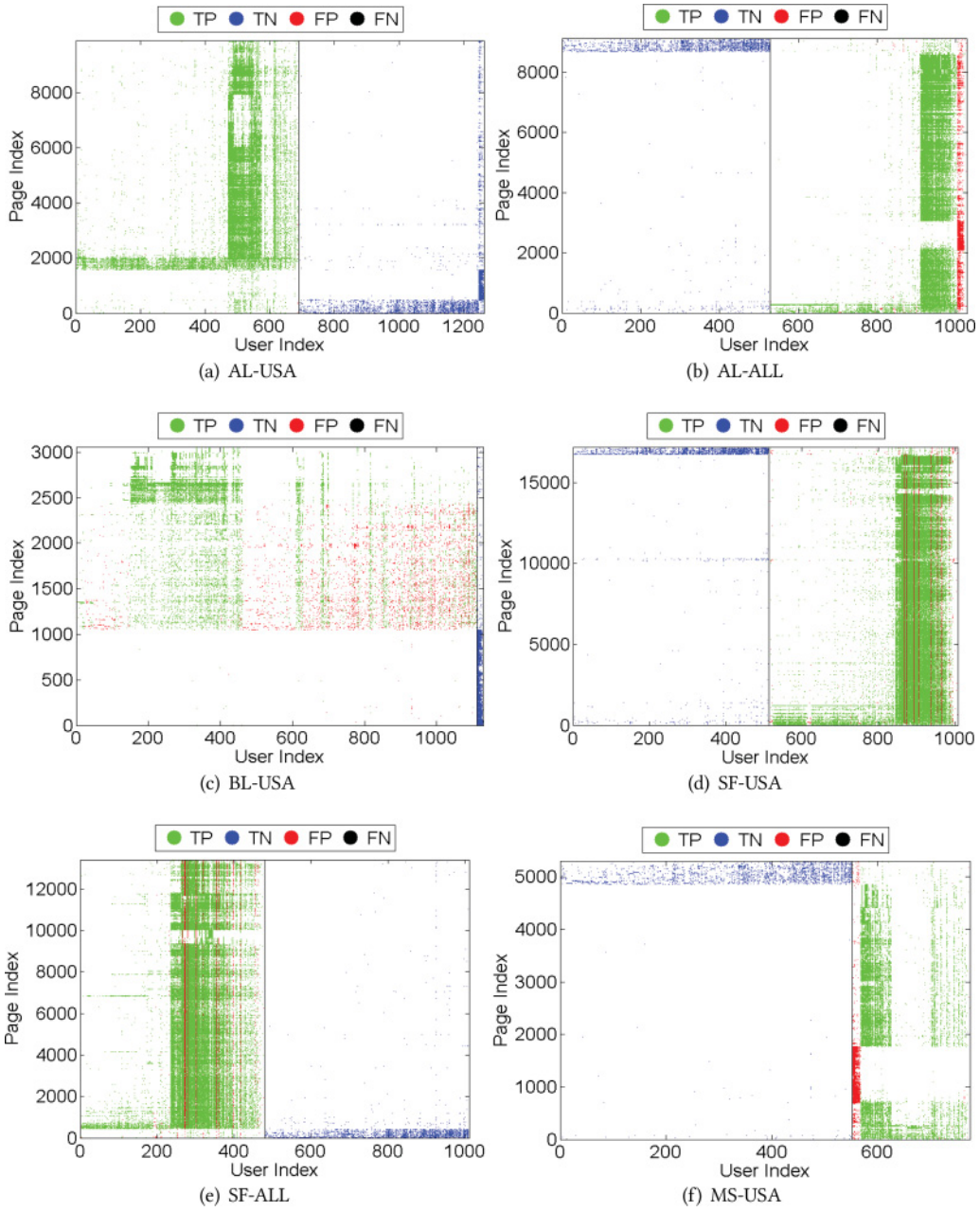


Fig. 7. Visualization of graph co-clustering results. The vertical black line indicates the separation between two clusters. We note that the clustering algorithm fails to achieve good separation, leading to a large number of false positives (red dots).

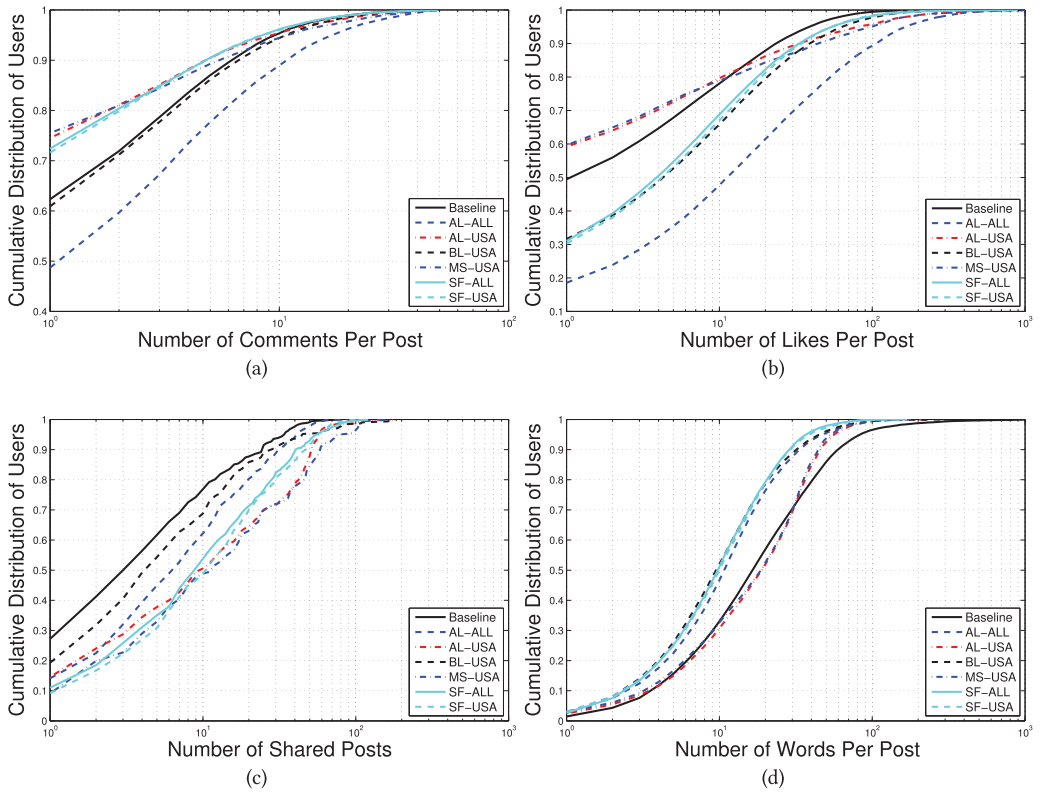


Fig. 8. Distribution of non-lexical features for like farm and baseline accounts.

to be used by machine-learning algorithms (in Section 5) for accurately classifying like farms vs. regular accounts.

4.1 Analysis of Non-Lexical Features

Comments and likes. In Figure 8(a), we plot the distributions of the number of comments a post attracts, revealing that users of AL-ALL like farm generate many more comments than the baseline users. We note that BL-USA is almost identical to the baseline users. Next, Figure 8(b) shows the number of likes associated with users' posts, highlighting that posts of like farm users attract much more likes than those of baseline users. Therefore, posts produced by the former gather more likes (and also have lower lexical richness as shown later on in Table 6), which might actually indicate their attempt to mask suspicious activities.

Shared content. We next study the distributions of posts that are classified as "shared activity," that is, originally made by another user, or articles, images, or videos linked from an external URL (e.g., a blog or YouTube). Figure 8(c) shows that baseline users generate more original posts, and share fewer posts or links, compared to farm users.

Words per post. Figure 8(d) plots the distributions of number of words that make up a text-based post, highlighting that posts of like farm users tend to have fewer words. Roughly half of the users in four of the like farms (AL-ALL, BL-USA, SF-ALL, and SF-USA) use 10 or fewer words in their posts vs. 17 words by baseline users.

Table 6. Lexical Analysis of Timeline Posts

Campaign	Avg Chars	Avg Words	Avg Sents	Avg Sent Length	Avg Word Length	Richness	ARI	Flesch Score
Baseline	4,477	780	67	6.9	17.6	0.70	20.2	55.1
BL-USA	7,356	1,330	63	5.7	22.8	0.58	16.9	51.5
AL-ALL	2,835	464	32	6.2	13.9	0.59	14.8	43.6
AL-USA	2,475	394	33	6.2	12.7	0.49	14.1	54.0
SF-ALL	1,438	227	19	6.3	11.7	0.58	14.1	45.2
SF-USA	1,637	259	22	6.3	12.0	0.55	14.4	45.6
MS-USA	6,227	1,047	66	6.1	17.8	0.53	16.2	50.1

4.2 Analysis of Lexical Features

We now look at features that relate to the content of timeline posts, and similar lexical features could be extracted for other non-English languages. We acknowledge that the extraction of lexical features of a non-English language is a challenging task and the extraction models might be prone to errors. We constrain our analysis to only English language and argue that lexical features extractions and analysis could be extended for other non-English Language such as Chinese (Zhang et al. 2003a, 2003b), French (Silberztein 1989), Arabic (Farghaly and Shaalan 2009), and Hindi/Urdu (Tiwary and Siddiqui 2008). We refer the reader to Silberztein (1997) for more details about lexical features used in this article.

We have also considered user timelines as the collection of posts and the corresponding comments on each post (i.e., all textual content) and build a corpus of words extracted from the timelines by applying the term frequency-inverse document frequency (TF-IDF) statistical tool (Salton and McGill 1986). However, the overall performance of this "bag-of-words" approach was poor, which can be explained with the short nature of the posts. Indeed, Hogenboom et al. (2015) has shown that the word frequency approach to analyze short text on social media and blogs does not perform well. Thus, in our work, we disregard simple TF-IDF based analysis of user timelines and identify other lexical features.

Language. Next, we analyze the ratio of posts in English, that is, for every post we filter out all non-English ones using a standard language detection library.[6] For each user, we count the number of English-language posts and calculate its ratio with respect to the total number of posts. Figure 9 shows that the baseline users and like farm users in the USA (i.e., MS-USA, BL-USA, and AL-USA) mostly post in English, while users of worldwide campaigns (MS-ALL, BL-ALL, AL-ALL) have significantly fewer posts in English. For example, the median ratio of English posts for AL-ALL campaign is around 10% and that for SF-ALL around 15%. We acknowledge that our analysis is limited to English-only content and may be statistically biased toward non-native English speakers that is, non-USA campaign users. While our analysis could be extended to other languages, we argue that English-based lexical analysis provides sufficient differences across different categories of users. Thus, developing algorithms for language detection and processing on non-English posts is out of the scope of this article.

Readability. We further analyze posts for grammatical and semantic correctness. We parse each post to extract the number of words, sentences, punctuation, and non-letters (e.g., emoticons) and measure the lexical richness, as well as the Automated Readability Index (ARI) (Senter and Smith 1967) and Flesch score (Flesch 1948). Lexical richness, defined as the ratio of number of unique

[6] https://python.org/pypi/langdetect (accessed on July 18, 2016).

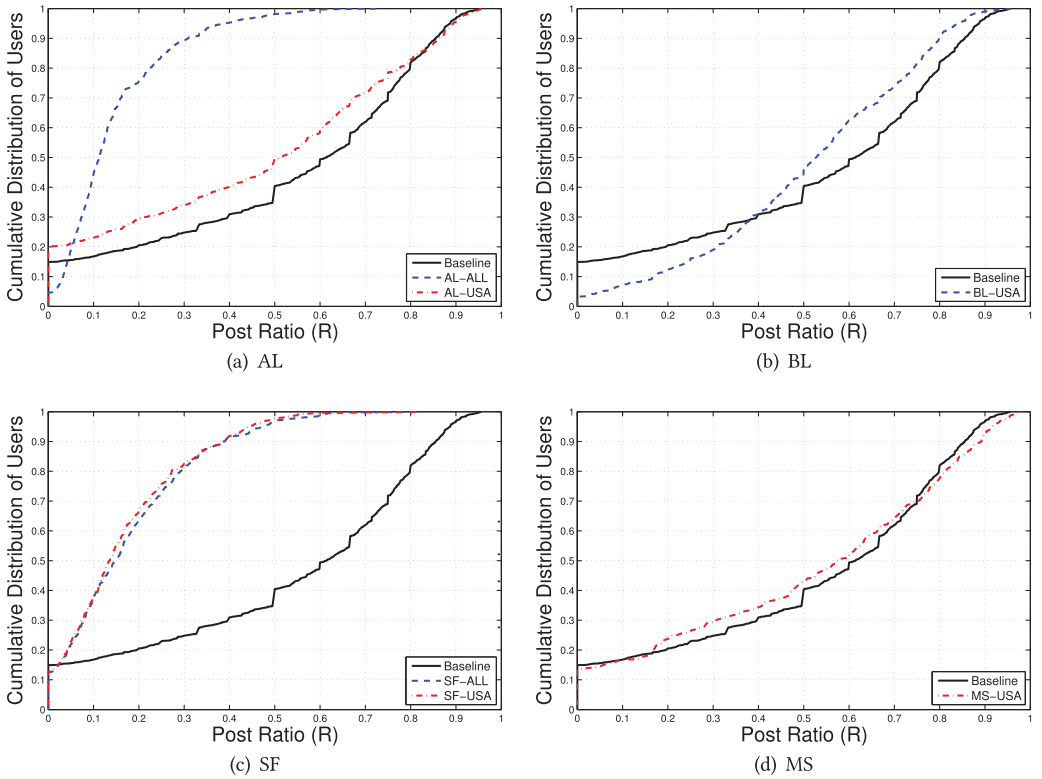


Fig. 9. Distributions of the ratio of English posts to non-English posts.

words to total number of words, reveals noticeable repetitions of distinct words, while the ARI, computed as 4.71 × average word length) + (0.5 × average sentence length) −21.43, estimates the comprehensibility of a text corpus. Table 6 shows a summary of the results. In comparison to like farm users, baseline users post text with higher lexical richness (70% vs. 55%), ARI (20 vs. 15), and Flesch score (55 vs. 48), thus suggesting that normal users use a richer vocabulary and that their posts have higher readability.

4.3 Remarks

Our analysis of user timelines highlights several differences in both lexical and non-lexical features of normal and like farm users. In particular, we find that posts made by like farm accounts have 43% fewer words, a more limited vocabulary, and lower readability than normal users' posts. Moreover, like farm users generate significantly more comments and likes and a large fraction of their posts consists of non-original and often redundant "shared activity."

In the next section, we will use these timelines features to automatically detect like farm users using a machine-learning classifier.

5 TIMELINE-BASED DETECTION OF LIKE FARMS

Aiming to automatically distinguish like farm users from normal (baseline) users, we use a supervised two-class SVM classifier (Müller et al. 2001), implemented using *scikit-learn* (Buitinck et al. 2013) (an open source machine-learning library for Python). We later compare this classifier with

Table 7. Effectiveness of Non-Lexical Features (+SVM) in Detecting Like Farm Users

Campaign	Total Users	Training Set	Testing Set	TP	FP	TN	FN	Precision	Recall	Accuracy	F1-Score
BL-USA	583	466	117	37	80	270	12	76%	32%	77%	45%
AL-ALL	707	566	141	132	9	278	4	96%	94%	97%	95%
AL-USA	827	662	164	113	51	278	4	97%	69%	88%	81%
SF-ALL	870	696	174	139	35	273	9	94%	80%	90%	86%
SF-USA	653	522	131	110	21	277	5	96%	84%	94%	90%
MS-USA	259	207	52	39	13	280	2	95%	75%	96%	84%

Table 8. Effectiveness of Lexical Features (+SVM) in Detecting Like Farm Users

Campaign	Total Users	Training Set	Testing Set	TP	FP	TN	FN	Precision	Recall	Accuracy	F1-Score
BL-USA	564	451	113	113	0	240	0	100%	100%	100%	100%
AL-ALL	675	540	135	133	2	238	2	99%	99%	99%	99%
AL-USA	570	456	114	113	1	239	1	99%	99%	99%	99%
SF-ALL	761	609	152	151	1	238	2	99%	99%	99%	99%
SF-USA	570	456	114	113	1	225	15	99%	87%	95%	92%
MS-USA	224	179	45	45	0	240	0	100%	100%	100%	100%

other well-known supervised classifiers such as Decision Tree (Breiman et al. 1984), AdaBoost (Freund and Schapire 1997), kNN (Andoni and Indyk 2008), Random Forest (Breiman 2001), and Naïve Bayes (Zhang 2004) and confirm that the two-class SVM is the most effective in detecting like farms users.

We extract 4 non-lexical features and 12 distinct lexical features from the timelines of baseline and like farm users, as explained in Section 4, using the datasets presented in Section 3.1. The non-lexical features are the average number of words, comments, likes per post, and re-shares. The lexical features include the following: the number of characters, words, and sentences; the average word length, sentence length, and number of upper case letters; the average percentage of punctuation, numbers, and non-letter characters; richness, ARI, and Flesch Score.

We form two classes by labeling like farm and baseline users' lexical and non-lexical features as positives and negatives, respectively. We use 80% and 20% of the features to build the training and testing sets, respectively. Appropriate values for parameters γ (*radial basis function kernel* parameter (Schölkopf et al. 2001)) and υ (SVM parameter) are set empirically by performing a greedy grid search on ranges $2^{-10} \leq \gamma \leq 2^0$ and $2^{-10} \leq \upsilon \leq 2^0$, respectively, on each training group.

Non-lexical features. Table 7 reports on the accuracy of our classifier with non-lexical features, that is, users interactions with posts as described in Section 4.1. Note that for each campaign, we train the classifier with 80% of the non-lexical features from baseline and campaign training sets derived from the campaign users timelines. The poor classification performance for the stealthiest like farm (BL-USA) suggests that non-lexical features alone are not sufficient to accurately detect like farm users.

Lexical features. Next, we evaluate the accuracy of our classifier with lexical features, reported in Table 8. We filter out all users with no English-language posts (i.e., with the ratio of English posts to non-English posts, $R = 0$, see Figure 9). Again, we train the classifier with 80% lexical features from baseline and like farm training sets. We observe that our classifier achieves very high precision and recall for MS-USA, BL-USA, and AL-USA. Although the accuracy decreases by

Table 9. Effectiveness of Both Lexical and Non-Lexical Features (+SVM) in Detecting Like Farm Users

Campaign	Total Users	Training Set	Testing Set	TP	FP	TN	FN	Precision	Recall	Accuracy	F1-Score
BL-USA	583	466	117	116	1	278	4	99%	97%	99%	98%
AL-ALL	707	566	141	140	1	278	4	99%	97%	99%	98%
AL-USA	827	662	164	164	0	275	7	100%	96%	98%	97%
SF-ALL	870	696	174	172	2	271	11	99%	94%	97%	96%
SF-USA	653	522	131	130	1	273	9	99%	93%	98%	96%
MS-USA	259	207	52	52	0	280	2	100%	96%	99%	98%

Table 10. F1-Score Obtained with Different Classification Methods, Using Both Lexical and Non-lexical Features, in Detecting Like Farm Users

Campaign	SVM	Decision Tree	AdaBoost	kNN	Random Forest	Naïve Bayes
BL-USA	98%	96%	96%	91%	88%	53%
AL-ALL	98%	84%	95%	86%	84%	75%
AL-USA	97%	88%	90%	91%	86%	81%
SF-ALL	96%	90%	94%	89%	87%	67%
SF-USA	96%	83%	92%	79%	78%	61%
MS-USA	98%	90%	89%	89%	87%	74%

approximately 8% for SF-USA, the overall performance suggests that lexical features are useful in automatically detecting like farm users.

Combining lexical and non-lexical features. While building a classifier based on lexical features performs very well in detecting fake accounts, we acknowledge that lexical features may be affected by geographical location, especially if one set of users who write in English are native speakers while the other set is not. Therefore, we further combine both lexical and non-lexical features to build a more robust classifier. We also note that approximately 3% to 22% of like farm users and 14% of baseline users do not have English language posts and are not considered in the lexical features based classification. To include these users in our classification, for each like farm and baseline, we set their lexical features to zeros and aggregate the lexical features with non-lexical features and evaluate our classifier with the same classification methodology as detailed above. Results are summarized in Table 9, which shows high accuracy for all like farms (F1-Score $\geq$ 96%), thus confirming the effectiveness of our timeline-based features in detecting like farm users.

Comparison with other machinelearning classifiers. To generalize our approach, we have also used other machine-learning classification algorithms, that is, Decision Tree, AdaBoost, kNN, Random Forest, and Naïve Bayes. The training and testing of all these classifiers follow the same setup as the SVM approach. We again use 80% and 20% of the combined lexical and non-lexical features to build the training and testing sets, respectively. We summarize the performance of the classifiers in Table 10. Our results show that the SVM classifier achieves the highest F1-Scores across the board. Due to overfitting on our dataset, Random Forest and Naïve Bayes show poor results and require mechanism such as pruning, detailed analysis of parameters, as well as selection of the optimal set of prominent features to improve classification performance (Breiman 2001; Kohavi and Sommerfield 1995).

Analysis. We now analyze in more detail the classification performance (in terms of F1-Score) to identify the most distinctive features. Specifically, we incrementally add lexical and non-lexical

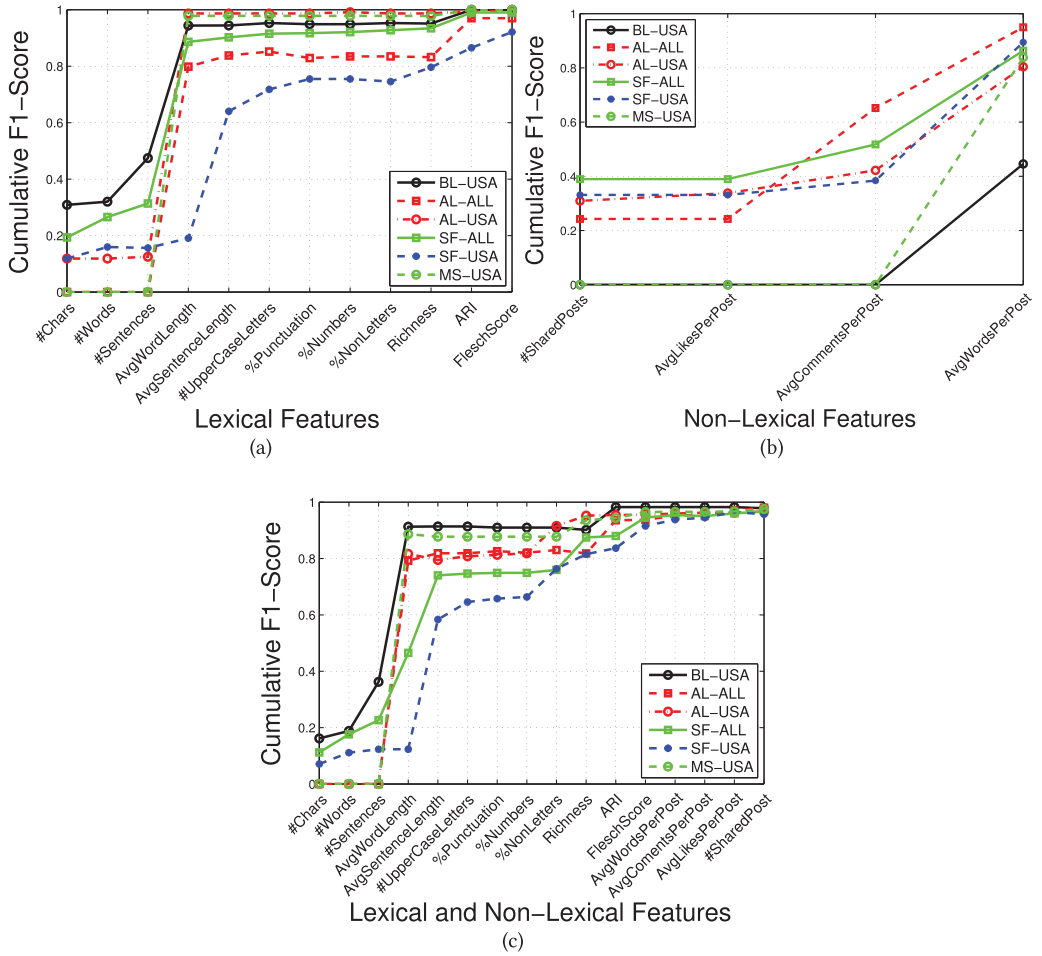


Fig. 10. Cumulative F1-Score for all lexical and non-lexical features measured. The x-axis shows the incremental inclusion of features in both training and testing of SVM. Details of the classification performance for all features are listed in Table 9.

features to train and test our classifier for all campaigns. We observe that the average word length (cf. Figure 10(a)) and average number of words per post (cf. Figure 10(b)) provide the most improvement in the F1-Score for all campaigns. This finding suggests that like farm users use shorter words and fewer number of words in their timeline posts as compared to baseline users. While these features provide the largest improvement in detecting a like farm account, an attempt to circumvent detection by increasing the word length or number of words per post will also effect the ARI, Flesch score, and richness. That is, increasing word length and number of words on posts in a way that is not readable nor understandable will not improve the overall outlook of the account to appear real. Therefore, combining several features increases the workload required to appear real on like farm accounts. The overall classification accuracy with both lexical and non-lexical features is reported in Figure 10(c).

Robustness of our approach. The like farms users may evade our detection system by mimicking the behavior of real users. To test the effectiveness of our features and classifiers, we assume two

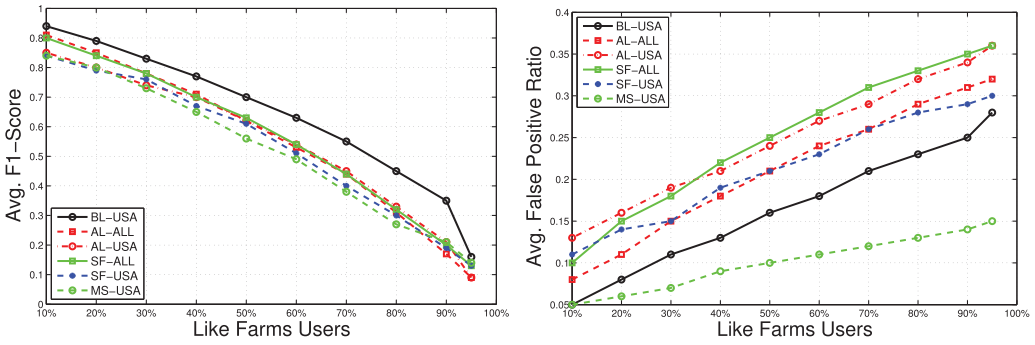


Fig. 11. Average F1-Score and false-positive ratio measured when fractions of like farms users mimic all lexical and non-lexical features (+SVM). The x-axis shows the percentage of like farms users who are mimicking baseline users.

Table 11. The Difference in F1-Score Obtained When All Like Farm Users Coordinate and Mimic Sets of Lexical and Non-Lexical Features of Baseline Users. F1-Score in Table 9 Is Used as a Reference to Compute the Δ in F1-Score

	Δ F1-Score				
Campaign	**1-Feature**	**2-Features**	**3-Features**	**4-Features**	**8-Features**
BL-USA	1%	2%	3%	5%	42%
AL-ALL	2%	3%	4%	5%	47%
AL-USA	2%	4%	5%	10%	20%
SF-ALL	3%	4%	6%	6%	56%
SF-USA	8%	9%	11%	13%	55%
MS-USA	5%	6%	6%	7%	26%

worst-case attacking scenarios: (i) fractions of like farms users mimic all features of baseline users; and (ii) all like farm users mimic sets of baseline users' features.

We simulate the first scenario by assuming that sets of like farm users randomly select baseline users and aggressively replace the values of all their features with that of the selected baseline users. We use the aforementioned settings of the best of our classifiers, SVM, and run the experiments for each like farm 10 times. Figure 11 shows the effect on F1-Score of our classifier when fractions of like farm users aggressively mimic all the lexical and non-lexical features of baseline users. When 30% of like farms users coordinate and mimic all features of baseline users, we observe that our classifier achieves at least 73% F1-Score and at most 17% false positive ratio, decreasing 26% F1-Score compared to our approach (cf. Table 9).

For the latter case, we assume that all like farms users coordinate and select sets of features from randomly selected baseline users that they copy or mimic. We use identical configuration of our SVM classifier and conduct experiments for each like farm 10 times. Table 11 summarizes the results of our experiments. With this attack strategy, we observe that when only one feature is mimicked, the F1-Score of our approach (cf. Table 9) decreases by between 1% and 8%. The F1-Score of our classifier decreases by between 26% and 56% when the like farm users target sets of eight features including prominent ones (cf. Figure 10).

Note that any feature used to identify fake like farms' behavior can be either circumvented or manipulated by the like farms users by behaving more like real users. We believe that this is a

typical arms race that eventually raises the bar for the like farms—the more effort they need to invest in appearing as real users, the lower their incentive is to do this.

Remarks. Our results demonstrate that it is possible to accurately detect like farm users from both sophisticated and naïve farms by incorporating additional account information—specifically, timeline activities. The low false-positive ratio ($\ll$1%, cf. Table 9) highlights the effectiveness of our approach as well as the limitations of prior graph co-clustering algorithms in detecting like farms users (cf. Section 3). Unfortunately, we do not have access to a larger dataset to measure and discuss the effects on the false-positive ratio of our approach. We believe then that without an evaluation of our approach at a larger scale, further discussion would be speculative, so we refrained from further interpretation of those results. We also argue that the use of a variety of lexical and non-lexical features will make it difficult for like farm operators to circumvent detection. Like farms typically rely on pre-defined lists of comments, resulting in word repetition and lower lexical richness. As a result, we argue that, should our proposed techniques be deployed by Facebook, it will be challenging, as well as costly, for fraudsters to modify their behavior and evade detection, since this would require instructing automated scripts and/or cheap human labor to match the diversity and richness of real users' timeline posts.

6 RELATED WORK

Prior work has focused quite extensively on the analysis and the detection of sybil and/or fake accounts in online social networks by relying on tightly knit community structures (Boshmaf et al. 2015; Cao et al. 2012; Danezis and Mittal 2009; Yang et al. 2011, 2012; Yu et al. 2006). By contrast, we work to detect accounts that are employed by like farms to boost the number of Facebook page likes, whether they are operated by a bot or a human. We highlight several characteristics about the social structure and activity of fake profiles attracted by the honeypot pages, for example, their interconnected nature or the activity bursts. In fact, our analysis not only confirms a few insights used by sybil detection algorithms but also reveals new patterns that could complement them. Fraud and fake activities are not restricted to social network but are widespread also on other platforms, such as online gaming. In this context, Lee et al. (2016) rely on self-similarity to effectively measure the frequency of repeated activities per player over time and use it to identify bots. Also, Kwon et al. (2017) analyze the characteristics of the ecosystem of multiplayer online role-playing games and devise a method for detecting gold farming groups, based on graph techniques.

Prior work on reputation manipulation on social networks include a few *passive* measurement studies have also focused on characterizing fake user accounts and their activity. Nazir et al. (2010) studied phantom profiles in Facebook gaming applications, while Thomas et al. (2011) analyzed over 1.1 million accounts suspended by Twitter. Gao et al. (2010) studied spam campaigns on Facebook originating from approximately 57,000 user accounts. Yang et al. (2012) performed an empirical analysis of social relationships between spam accounts on Twitter, and Dave et al. (2012) proposed a methodology to measure and fingerprint click-spam in ad networks. Our work differs from these studies, as they all conducted passive measurements, whereas we rely on the deployment of several honeypot pages and (paid) campaigns to actively engage with fake profiles. Lee et al. (2010) and Stringhini et al. (2010) created honeypot profiles in Facebook, MySpace, and Twitter to detect spammers while we use accounts attracted by our honeypot Facebook pages that actively engage like farms. Unlike Lee et al. (2010) and Stringhini et al. (2010), we leverage timeline-based features for the detection of fake accounts. Our work also differs from theirs in that (1) their honeypot profiles were designed to look legitimate, while our honeypot pages explicitly indicated

they were not "real" (to deflect real profiles), and (2) our honeypot pages *actively* attracted fake profiles by means of paid campaigns as opposed to passive honeypot profiles.

Thomas et al. (2013) analyzed trafficking of fake accounts in Twitter. They bought fake profiles from 27 merchants and developed a classifier to detect these fake accounts. In a similar study, Stringhini et al. (2012, 2013) analyzed the market of *Twitter followers*, which, akin to Facebook like farms, provide Twitter followers for sale. Note that Twitter follower markets differ from Facebook like farms, as Twitter entails a *follower-followee* relationship among users, while Facebook friendships imply a bidirectional relationships. Also, there is no equivalent of liking a Facebook page in the Twitter ecosystem.

Wang et al. (2014) studied human involvement in Weibo's reputation manipulation services, showing that simple evasion attacks (e.g., workers modifying their behavior) as well as poisoning attacks (e.g., administrators tampering with the training set) can severely affect the effectiveness of machine-learning algorithms to detect malicious crowd-sourcing workers. Song et al. (2015) also looked at *crowdturfing* services that manipulate account popularity on Twitter through artificial retweet and developed "CrowdTarget" to detect such tweets. Partially informed by these studies, we not only cluster like activity performed by users but also build on lexical and non-lexical features.

Specific to Facebook fraud is CopyCatch (Beutel et al. 2013), a technique deployed by Facebook to detect fraudulent accounts by identifying groups of connected users liking a set of pages within a short time frame. SynchroTrap (Cao et al. 2014) extended CopyCatch by clustering accounts that perform similar, possibly malicious, synchronized actions, using tunable parameters such as time-window and similarity thresholds to improve detection accuracy. However, as discussed earlier, while some farms seem to be operated by bots (producing large bursts of likes and having limited numbers of friends) that do not really try to hide their activities, other *stealthier* farms exhibit behavior that may be challenging to detect with tools like CopyCatch and SynchroTrap. In fact, our evaluation of graph co-clustering techniques shows that these farms successfully evade detection by avoiding lockstep behavior and liking sets of seemingly random pages. As a result, we use timeline features, relying on both lexical and non-lexical features, to build a classifier that detects stealthy like farm users with high accuracy. Finally, we highlight that our work can complement other methods used in prior work to detect fake and compromised accounts, such as using unsupervised anomaly detection techniques (Viswanath et al. 2014), temporal features (Jiang et al. 2014a, 2014b), IP addresses (Stringhini et al. 2015), as well as generic supervised learning (Badri Satya et al. 2016).

Remarks on "new material." Compared to our preliminary results (published in De Cristofaro et al. (2014) and reported in Section 2), this article clearly introduces significant additional new material. Specifically, (i) we introduce an empirical evaluation demonstrating that temporal and social graph analysis can only be used to detect naive farms (Section 3), and (ii) we present a novel timeline-based classifier geared to detect accounts from stealthy like farms with a remarkably high degree of accuracy (Sections 4 and 5).

7 CONCLUSION

Minimizing fraud in online social networks is crucial for maintaining the confidence and trust of the user base and investors. In this article, we presented the results of a measurement study of Facebook like farms, that is, paid services artificially boosting the number of likes on a Facebook page, aiming to identify characteristics and accurately detect the accounts used by them. We crawled profile information, liking patterns, and timeline activities from like farms accounts. Our demographic, temporal, and social graph analysis highlighted similar patterns between accounts

across different like farms and revealed two main modi operandi: Some farms seem to be operated by bots and do not really try to hide the nature of their operations, while others follow a stealthier approach, mimicking regular users' behavior.

We then evaluated the effectiveness of existing graph-based fraud detection algorithms, such as CopyCatch (Beutel et al. 2013) and SynchroTrap (Cao et al. 2014), and demonstrated that sophisticated like farms can successfully evade detection.

Next, aiming to address their shortcomings, we focused on incorporating additional profile information from accounts' timelines to train machine-learning classifiers geared to distinguish between like farm users from normal ones. We extracted lexical and non-lexical features from user timelines, finding that posts by like farm accounts have 43% fewer words, a more limited vocabulary, and lower readability than normal users' posts. Moreover, like farm posts generated significantly more comments and likes, and a large fraction of their posts consists of non original and often redundant "shared activity" (i.e., repeatedly sharing posts made by other users, articles, videos, and external URLs). By leveraging both lexical and non-lexical features, we experimented with several machine-learning classifiers, with the best of our classifiers (SVM) achieving as high as 100% precision and 97% of recall, and at least 99% and 93%, respectively, across all campaigns—significantly higher than graph co-clustering techniques.

In theory, fraudsters could try to modify their behavior to evade our proposed timeline-based detection. However, like farms either heavily automate mechanisms or rely on manual input of cheap human labor. Since non-lexical features are extracted from users' interactions with timeline posts, imitating normal users' behaviors will likely incur an remarkably higher cost. Even higher would be the cost to interfere with lexical features, since this would entail modifying or imitating normal users' writing style.

REFERENCES

Alexandr Andoni and Piotr Indyk. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM* 51, 1 (Jan. 2008), 117–122. DOI: http://dx.doi.org/10.1145/1327452.1327494

Charles Arthur. 2013. How Low-Paid Workers at 'Click Farms' Create Appearance of Online Popularity. Retrieved from http://gu.com/p/3hmn3/stw.

Prudhvi Ratna Badri Satya, Kyumin Lee, Dongwon Lee, Thanh Tran, and Jason (Jiasheng) Zhang. 2016. Uncovering fake likers in online social networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM'16)*. ACM, New York, NY, 2365–2370. DOI: http://dx.doi.org/10.1145/2983323.2983695

Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. 2013. CopyCatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. ACM, New York, NY, 119–130. DOI: http://dx.doi.org/10.1145/2488388.2488400

Yazan Boshmaf, Dionysios Logothetis, Georgos Siganos, Jorge Lería, José Lorenzo, Matei Ripeanu, and Konstantin Beznosov. 2015. Integro: Leveraging victim prediction for robust fake account detection in OSNs. In *Proceedings of the 22nd Annual Network and Distributed System Security Symposium (NDSS'15)*.

Leo Breiman. 2001. Random forests. *Machine Learning*.

Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees*. CRC Press.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: Experiences from the scikit-learn project. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases Workshop (ECML/PKDD LML'13)*.

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 (NSDI'12)*. USENIX Association, Berkeley, CA, 15–15. http://dl.acm.org/citation.cfm?id=2228298.2228319

Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. 2014. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS'14)*. ACM, New York, NY, 477–488. DOI: http://dx.doi.org/10.1145/2660267.2660269

Brian Carter. 2013. *The Like Economy: How Businesses Make Money with Facebook*. QUE Publishing.

Rory Cellan-Jones. 2012. Who 'likes' my Virtual Bagels? (July 2012). Retrieved from http://www.bbc.co.uk/news/technology-18819338.

A. Chaabane, G. Acs, and Mohamed Ali Kaafar. 2012. You are what you like! Information leakage through users' interests. In *Proceedings of the 19th Annual Network and Distributed System Security Symposium (NDSS'12)*. https://hal.inria.fr/hal-00748162.

Terence Chen, Abdelberi Chaabane, Pierre Ugo Tournoux, Mohamed-Ali Kaafar, and Roksana Boreli. 2013. How much is too much? Leveraging ads audience estimation to evaluate public profile uniqueness. In *Proceedings of the 13th International Symposium on Privacy Enhancing Technologies (PETS'13)*. Springer, Berlin, 225–244. DOI: http://dx.doi.org/10.1007/978-3-642-39077-7_12

George Danezis and Prateek Mittal. 2009. SybilInfer: Detecting sybil nodes using social networks. In *Proceedings of the 16th Annual Network and Distributed System Security Symposium (NDSS'09)*. The Internet Society.

Vacha Dave, Saikat Guha, and Yin Zhang. 2012. Measuring and fingerprinting click-spam in ad networks. *SIGCOMM Comput. Commun. Rev.* 42, 4 (Aug. 2012), 175–186. DOI: http://dx.doi.org/10.1145/2377677.2377715

Emiliano De Cristofaro, Arik Friedman, Guillaume Jourjon, Mohamed Ali Kaafar, and M. Zubair Shafiq. 2014. Paying for likes?: Understanding facebook like fraud using honeypots. In *Proceedings of the 2014 Conference on Internet Measurement Conference (IMC'14)*. ACM, New York, NY, 129–136. DOI: http://dx.doi.org/10.1145/2663716.2663729

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Trans. Asian Lang. Inf. Process.* 8, 4 (2009), 14.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32, 3 (1948), 221–233.

Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 1 (Aug. 1997), 119–139.

Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. 2010. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC'10)*. ACM, New York, NY, 35–47. DOI: http://dx.doi.org/10.1145/1879141.1879147

Alexander Hogenboom, Flavius Frasincar, Franciska de Jong, and Uzay Kaymak. 2015. Using rhetorical structure in sentiment analysis. *Commun. ACM* 58, 7 (June 2015), 9.

Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. 2014a. CatchSync: Catching synchronized behavior in large directed graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. ACM, New York, NY, 941–950. DOI: http://dx.doi.org/10.1145/2623330.2623632

Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. 2014b. Inferring strange behavior from connectivity pattern in social networks. In *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference (PAKDD'14)*. Springer International, 126–138. DOI: http://dx.doi.org/10.1007/978-3-319-06608-0_11

Yuval Kluger, Ronen Basri, Joseph T. Chang, and Mark Gerstein. 2003. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Res.* 13, 4 (2003), 703–716. DOI: http://dx.doi.org/10.1101/gr.648603

Ron Kohavi and Dan Sommerfield. 1995. Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining (KDD'95)*. 192–197.

Hyukmin Kwon, Aziz Mohaisen, Jiyoung Woo, Yongdae Kim, Eunjo Lee, and Huy Kang Kim. 2017. Crime scene reconstruction: Online gold farming network analysis. *IEEE Trans. Inf. Forens. Secur.* 12, 3 (2017), 544–556.

Justin Lafferty. 2013. How Many Pages Does The Average Facebook User Like? Retrieved from http://allfacebook.com/how-many-pages-does-the-average-facebook-user-like_b115098.

Eunjo Lee, Jiyoung Woo, Hyoungshick Kim, Aziz Mohaisen, and Huy Kang Kim. 2016. You are a game bot!: Uncovering game bots in MMORPGs via self-similarity in the wild. In *NDSS*.

Kyumin Lee, James Caverlee, and Steve Webb. 2010. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, USA, 435–442. DOI: http://dx.doi.org/10.1145/1835449.1835522

Richard Metzger. 2012. Facebook: I Want My Friends Back. Retrieved from http://dangerousminds.net/comments/facebook_i_want_my_friends_back.

Derek Muller. 2014. Facebook Fraud. Retrieved from https://www.youtube.com/watch?v=oVfHeWTKjag.

Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf. 2001. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks* 12, 2 (2001).

Atif Nazir, Saqib Raza, Chen-Nee Chuah, and Burkhard Schipper. 2010. Ghostbusting facebook: Detecting and characterizing phantom profiles in online social gaming applications. In *Proceedings of the 3rd Wonference on Online Social Networks (WOSN'10)*. USENIX Association, Berkeley, CA, 1–1. http://dl.acm.org/citation.cfm?id=1863190.1863191

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval.* McGraw–Hill, Inc., New York, NY.

Jaron Schneider. 2014. Likes or lies? How perfectly honest businesses can be overrun by Facebook spammers. Retrieved from http://thenextweb.com/facebook/2014/01/23/likes-lies-perfectly-honest-businesses-can-overrun-facebook-spammers/.

Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neur. Comput.* 13, 7 (July 2001), 29.

R. J. Senter and Edgar A. Smith. 1967. Automated Readability Index. AMRL-TR-66-22. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/667273.pdf.

Max Silberztein. 1989. The lexical analysis of French. In *Proceedings of the LITP Spring School on Theoretical Computer Science: Electronic Dictionaries and Automata in Computational Linguistics*, Maurice Gross and Dominique Perrin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 93–110.

Max Silberztein. 1997. The lexical analysis of natural languages. In *Finite-State Language Processing*, Emmanuel Roche and Yves Schabes (Eds.). MIT Press, Chapter 6, 175–203.

Benjamin Snyder. 2015. Facebook added 10 million small business pages in a year. (April 2015). Retrieved from http://fortune.com/2015/04/30/facebook-small-business.

Jonghyuk Song, Sangho Lee, and Jong Kim. 2015. CrowdTarget: Target-based detection of crowdturfing in online social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS'15)*. ACM, New York, NY, 793–804. DOI: http://dx.doi.org/10.1145/2810103.2813661

Gianluca Stringhini, Manuel Egele, Christopher Kruegel, and Giovanni Vigna. 2012. Poultry markets: On the underground economy of twitter followers. In *Proceedings of the 2012 ACM Workshop on Online Social Networks (WOSN'12)*. ACM, New York, NY, 1–6. DOI: http://dx.doi.org/10.1145/2342549.2342551

Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC'10)*. ACM, New York, NY, 1–9. DOI: http://dx.doi.org/10.1145/1920261.1920263

Gianluca Stringhini, Pierre Mourlanne, Grgoire Jacob, Manuel Egele, Christopher Kruegel, and Giovanni Vigna. 2015. EVILCOHORT: Detecting communities of malicious accounts on online services. In *USENIX Security*. USENIX Association, 563–578. http://dblp.uni-trier.de/db/conf/uss/uss2015.html#StringhiniMJEKV15

Gianluca Stringhini, Gang Wang, Manuel Egele, Christopher Kruegel, Giovanni Vigna, Haitao Zheng, and Ben Y. Zhao. 2013. Follow the green: Growth and dynamics in twitter follower markets. In *Proceedings of the 2013 Conference on Internet Measurement Conference (IMC'13)*. ACM, New York, NY, 163–176. DOI: http://dx.doi.org/10.1145/2504730.2504731

Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. 2011. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC'11)*. ACM, New York, NY, 243–258. DOI: http://dx.doi.org/10.1145/2068816.2068840

Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. 2013. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *Proceedings of the 22nd USENIX Conference on Security (SEC'13)*. USENIX Association, Berkeley, CA, 195–210. http://dl.acm.org/citation.cfm?id=2534766.2534784

U. S. Tiwary and Tanveer Siddiqui. 2008. *Natural Language Processing and Information Retrieval.* Oxford University Press.

Bimal Viswanath, M. Ahmad Bashir, Mark Crovella, Saikat Guha, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2014. Towards detecting anomalous user behavior in online social networks. In *Proceedings of the 23rd USENIX Security Symposium (USENIX Security'14)*. USENIX Association, San Diego, CA, 223–238. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/viswanath.

Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y. Zhao. 2014. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *Proceedings of the 23rd USENIX Security Symposium (USENIX Security'14)*. USENIX Association, San Diego, CA, 239–254. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/wang.

Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. ACM, New York, NY, 71–80. DOI: http://dx.doi.org/10.1145/2187836.2187847

Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y. Zhao, and Yafei Dai. 2011. Uncovering social network sybils in the wild. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC'11)*. ACM, New York, NY, 259–268. DOI: http://dx.doi.org/10.1145/2068816.2068841

Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. 2006. SybilGuard: Defending against sybil attacks via social networks. In *Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'06)*. ACM, New York, NY, 267–278. DOI: http://dx.doi.org/10.1145/1159913.1159945

Harry Zhang. 2004. The optimality of naive bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS'04)*, Valerie Barr and Zdravko Markov (Eds.). AAAI Press, Miami Beach, Florida, USA.

Hua-Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang, and Hong-Kui Yu. 2003a. Chinese lexical analysis using hierarchical hidden markov model. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, Volume 17 (SIGHAN'03)*.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003b. HHMM-based chinese lexical analyzer ICTCLAS. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, Volume 17 (SIGHAN'03)*. Association for Computational Linguistics, Stroudsburg, PA, 184–187. DOI: http://dx.doi.org/10.3115/1119250.1119280

Received July 2016; revised March 2017; accepted June 2017