Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach

Supraja Gurajala gurajas@clarkson.edu
Joshua S. White whitejs@clarkson.edu
Brian Hudson hudsonb@clarkson.edu
Jeanna N. Matthews jnm@clarkson.edu
Department of Computer Science, Clarkson University, Potsdam, NY 13699

ABSTRACT

In Online Social Networks (OSNs), the audience size commanded by an organization or an individual is a critical measure of that entity's popularity. This measure has important economic and/or political implications. Organizations can use information about their audience, such as age, location etc., to tailor their products or their message appropriately. But such tailoring can be biased by the presence of fake profiles on these networks. In this study, analysis of 62 million publicly available Twitter user profiles was conducted and a strategy to retroactively identify automatically generated fake profiles was established. Using a pattern-matching algorithm on screen-names with an analysis of tweet update times, a highly reliable sub-set of fake user accounts were identified. Analysis of profile creation times and URLs of these fake accounts revealed distinct behavior of the fake users relative to a ground truth data set. The combination of this scheme with established social graph analysis will allow for time-efficient detection of fake profiles in OSNs.

Categories and Subject Descriptors

Collaborative and Social Computing – Social Media, Collaborative and Social Computing design and evaluations – Social network analysis, Knowledge representation and reasoning, Law, social and behavioral sciences—Sociology.

1. INTRODUCTION

Online social networks (OSNs), such as Facebook, Twitter, Google+, Instagram, LinkedIn, Weibo, and RenRen, have become the preferred means of communication amongst a diverse set of users including: individuals, companies, and families divided over several continents. Of the different OSNs, Twitter has become popular with users such as young adults, governments, commercial enterprises, and politicians as a way of instantaneously connecting with their audience and directly conveying their message [1].

The success of Twitter (and other OSNs) as a platform for large-scale communication and the expansion of efforts to mine their data for new and novel applications related to public health, economic development, scientific dissemination etc., critically hinges on the authenticity of their user database. A sub-section of OSN users on most platforms are not authentic. These "fake" users (or Sybils/Socialbots) are generated intentionally, and often automatically/semi-automatically, by cyber-opportunists (or cyber-criminals) [2]. The fake users (or their operators) may send requests to 'follow' or 'friend' other OSN users and gain popularity when these requests are accepted [3]. The presence of fake followers can: bias an individual or organization's popularity based on follower number count [4]; alter the characteristics of the audience [5]; or create a legitimacy problem for individuals/organizations [1]. Sometimes a fake profile is created to essentially duplicate a user's online presence, and these "identity clone attacks" [6] are devised to direct online fraud.

To tackle the problem of spam in social networks, several graph-theory related detection techniques have been developed to identify Sybil accounts by their social graph properties [3]. In response, "spammers" have worked to integrate Sybils into authentic user communities by creating accounts with full profiles and background information similar to authentic users [3]. Such techniques have complicated detection efforts, requiring continued development of new spam-recognition approaches.

Machine learning techniques and honeypot harvesting approaches have been used to classify Twitter accounts as legitimate or not. In social honeypot techniques, user profiles are specifically created to attract spammers so as to harvest their profile information [7]. These spammer profiles are then analyzed using machine learning techniques to understand spammer behavior and thus aid the development of detection techniques [8]. Other Twitter specific approaches to identify spammers and fake profiles include: detection based on tweet-content (e.g., "number of hashtags per word of each tweet" [9]); use of tweet/tweeter characteristics such as "reputation score", "number of duplicate Tweets" and "number of URLs" [10]; and comparison of tweet links (URLs) to publicly blacklisted URLs/domains [11].

For fast detection of spam accounts, a simple profile-pattern detection without detailed "Tweet" analysis has been proposed [12]. For example, Benevenuto et al. [9] used profile information such as "number of followers and number of followings" to identify fake profiles. Thomas et al. [12] used a multi-variable pattern-recognition approach based on user-profile-name, screen-name, and email parameters. They determined that there was strong and consistent correlation for the three parameters for all fake accounts. Here, we extend these user profile-pattern detection based approaches with the inclusion of user activity time-stamp information, to develop a new process for detection of fake profiles with high reliability. The details of our schema and the analysis of the fake profile set enabled with this approach are presented below. This paper is organized under the following sections: Acquisition of Twitter

 SMSociety '15, July 27 - 29, 2015, Toronto, ON, Canada © 2015 ACM. ISBN 978-1-4503-3923-0/15/07…$15.00
DOI: http://dx.doi.org/10.1145/2789187.2789206

data; Design and Methodology; Analysis of Results; and Conclusions.

2. DESIGN AND IMPLEMENTATION

For analysis of user profiles, the first step was to obtain publicly available Twitter user profile information. Utilizing social web-crawling technique, we gathered profiles of ~62 million users and used map-reducing techniques and pattern recognition approaches to detect fake profiles. The details of the Twitter user profile acquisition approach and analysis techniques are discussed below.

2.1 Acquisition of Twitter User Profiles

Twitter provides access to user and social graph data through the Twitter REST API v1.1.2 [13]. For non-protected users, the majority of the user's profile information is publicly available and accessible through this API. To maintain reliability of its services and control costs, Twitter employs rate-limiting steps, restricting the number of calls that can be made to the API per rate limit window, currently defined as a 15-minute interval.

As recently as 2010, when the maximum number of Twitter user IDs was estimated to be less than 800 million [14], it was feasible to crawl over the entire Twitter user ID space. Since then, the number of users has grown significantly. In addition to the size of the database, the sparsity of the Twitter user ID space also complicates the search. Spam accounts constitute as much as 93% of all new accounts, 68% of which are detected and automatically suspended or deleted by Twitter [15]. Thus, in an exhaustive search, we are requesting information for a significant (~1.4B) number of accounts that no longer exist or are unavailable due to suspension. With user IDs now exceeding 2 billion, the sparsity of the user ID space, and the rate limits imposed by Twitter with the rollout of the Twitter REST API v1.1, an exhaustive search over the entire user ID space is no longer feasible.

To overcome these issues, our approach performs a Breadth First Search (BFS) over a given set of seed users, which we specify. As the social graph is crawled, previously unknown Twitter user IDs obtained from the list of the user's followers are pursued and eventually the user profiles for these IDs are acquired. This ensures that all user profile requests we make to Twitter include only valid Twitter user IDs. Effectively, this adds each previously unknown follower of a user as a seed user for the next iteration of the search. We used a multi-account approach with application access to crawl the Twitter social graph and gathered ~ 62 million Twitter user profiles, within a three-month period in late 2013.

3. METHODOLOGY

Our crawler obtained 33 different attributes for each Twitter profile, the descriptions of which are available from Twitter [16]. Our schema then analyzed patterns amongst combinations of these attributes to identify a highly reliable core set of fake profiles, which provided the basis for identifying key distinguishing characteristics of fake accounts based on their publicly available profile information. To limit the parameter space of our analysis, we initially investigated the database semi-manually to determine the primary attributes that differed amongst most users. From this analysis, it was established that several of the 33 attributes were largely unused (or left as default) by most users. The key attributes that were either user selected or varied with account usage were *id, followers_count, friends_count, verified, created_at, description, location, updated, profile_image_url and screen_name*. Using this reduced attribute set, we used the analysis approach described below for identification of a reliable fake profile set.

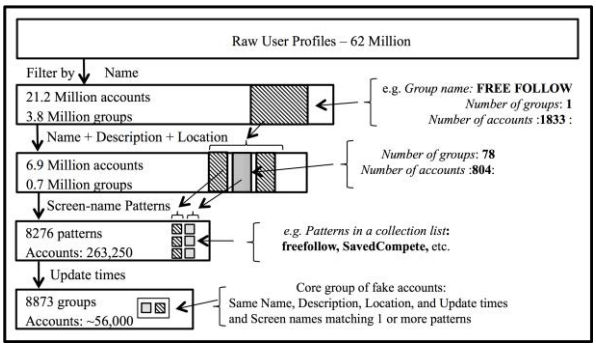


Figure 1: Schematic diagram of the algorithm for identification of fake profile groups.

A schematic diagram outlining our algorithm for detection of the core fake group is shown in Figure 1. Starting with our 62 million user profile database, we obtained groups (containing at least two accounts) with the same user profile information of name, description, and location. This filtering process resulted in generating 724,494 groups containing a total of 6,958,523 accounts. To further refine the 724,494 groups, their screen names were analyzed and near-identical ones were identified using a pattern-recognition algorithm procedure described below.

To identify screen name patterns in the 724,494 groups, a Shannon entropy-based analysis of the screen names in each group was conducted. For this, first, one of the screen names in a group was selected as a base screen name and its Shannon entropy was determined. Secondly, the next screen name in the group was concatenated with the selected base name and the Shannon entropy of the concatenated string was calculated. If the entropy of the concatenated string was greater than that of the base name by a threshold value (0.1), then the concatenated screen name was added to a collection list. This entropy comparison with the selected base screen name was repeated with all screen names in the group. All screen names that were not accumulated in the collection list associated with the first screen name were then re-grouped and analyzed with the above described pattern recognition procedure to generate other collection lists. This procedure was repeated until all screen names were either placed in a collection list or identified as not being a part of any collection. This procedure resulted in the division of the 724,494 groups into several collection lists with cohesive screen name entropy distributions.

A regular expression pattern (more than 4 characters long) search was then conducted within each collection list to obtain any pattern(s) that might exist in their screen names. The screen names associated with a pattern formed a "pattern list". From a manual inspection of the pattern lists, it was determined that this procedure was able to identify and group mass fake profiles with screen names that were seemingly generated automatically (e.g., freefollow1, freefollow3, etc.). The procedure was even able to group profiles with a common string well hidden within the screen name, with substantial lengths of prefixes and suffixes to

them. We did notice, however, that the entropies of some pattern lists were not tightly bound, suggesting a need to revise the entropy filtration procedure. This was accomplished by analyzing the broadness of the Shannon entropy distributions of the screen names in each pattern list, which was quantified by the normalized standard deviation of their entropies $\bar{\sigma}$ as shown in Equation 1.

$$(1)\ \bar{\sigma} = \frac{\left[\sum_i^N (x_i - \bar{x})^2\right]^{\frac{1}{2}}}{\bar{x}}$$

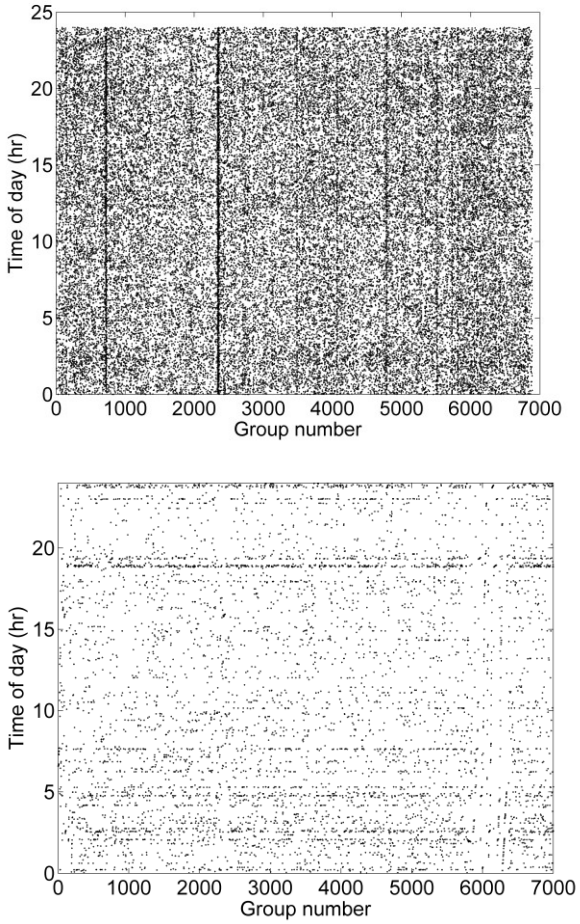
Where N represents the total number of screen names in a pattern list, x_i is the Shannon entropy of a i^{th} screen name, and $\bar{x}$ is the mean Shannon entropy of a pattern list.

A refined pattern list was generated by eliminating all pattern lists with relative standard deviations greater than a critical value (0.03; established empirically). The procedure, thus far, identified closely associated accounts that are very likely to be automatically generated. This procedure, however, results in false positives associated with highly popular names. As a final filter, the accounts were examined to determine their distribution of update times. The update times of a pattern list with genuine accounts is likely to be uniform or broadly distributed, while the fake accounts that are automatically updated from a single operator will likely have closely related update times. Pattern lists with broad update time distributions were then eliminated.

The above procedure resulted in the generation of a fake profile set that contains 8,873 groups with ~56,000 accounts that have identical name, description, location, and a narrow range of update times. In addition, the accounts in each group had screen names with matching patterns. Further investigation of the fake profile set revealed that all the identified accounts were updated not just within a narrow time distribution but as groups (i.e. with identical update times). While this list is not very large in size, the highly refined constraints applied to produce this fake profile set ensure a high likelihood of reducing false positives. A manual inspection of some of the accounts in the fake profile set did not reveal any false positives. This highly reliable set of fake profiles was then analyzed to determine their profile-based distinguishing features.

4. ANALYSIS OF RESULTS

To determine the characteristics of this set of fake accounts, the generation of a ground truth dataset was required. The ground truth dataset was obtained from a random sample of our Twitter user profile database. For consistency with the identified fake group datasets, the ground-truth set was selected to be of similar size and from a similar timeline. Analysis of the updated times, creation times, and profile URLs of the two datasets was then conducted to understand the relative characteristics of the two sets.



(a) Ground Truth (b) Fake Profiles

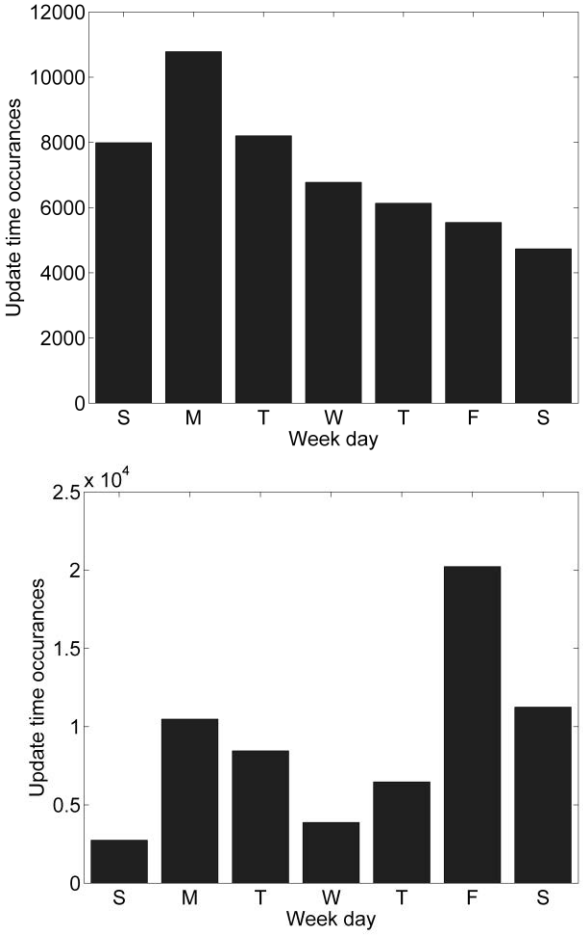
Figure 2: Comparison of update times for (a) ground truth and (b) fake profile datasets.

4.1 Update Times

A comparison of the update times of all 8,873 groups of fake profiles and of the generated truth dataset is shown in Figure 2. The truth dataset was randomly divided into a similar number of groups and group sizes to closely match the fake profile set. The update times of the truth dataset (Figure 2a) was seen to be almost uniformly distributed over the entire day, with no obvious time bias. The update times of the fake profile set (Figure 2b), as observed earlier, were non-uniformly distributed with significant time periods in a day when there was no update activity. Analysis of our dataset showed that the maximum number of profiles updated at a given time was 100 and this limit was achieved by one group in the fake profile set.

The distribution of the days of the week when the groups in the fake profile set and the ground truth dataset were updated is shown in Figure 3. For the ground truth data, the frequency distribution reveals that these users preferentially updated on Sundays and Mondays (UTC time). A decreasing number of users updated as the week progressed. Considering that the data is in UTC time (and could not be converted to local time, as location information was not always available), there is some

uncertainty in the actual update days. The distribution of update days for the fake profiles was seen to have a highly non-uniform distribution, with a bias for update during the later part of the week.



(a) Ground Truth (b) Fake Profiles

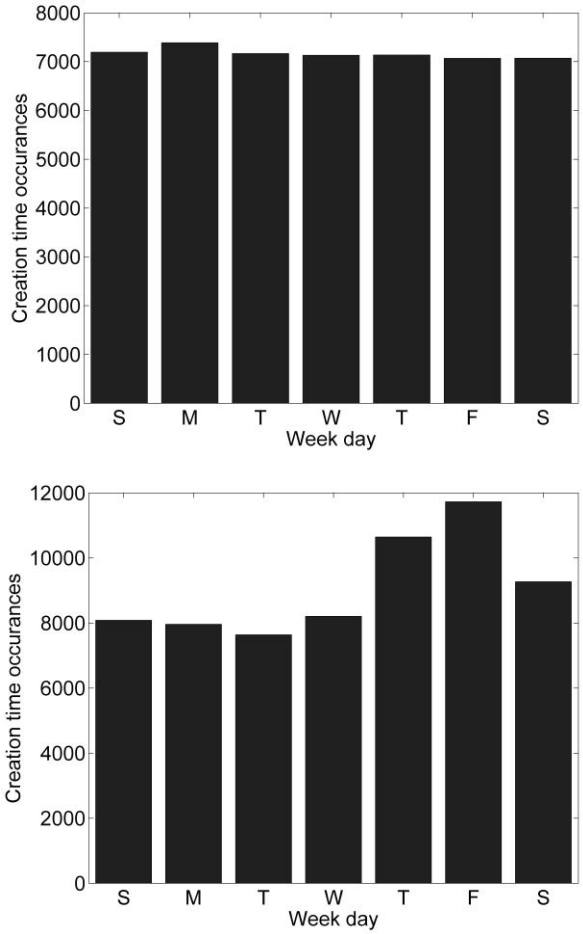
Figure 3: Occurrence frequency of update days for (a) ground truth and (b) fake profile datasets.

Our observation that fake accounts are updated more during the later part of the week may seem to match earlier observations of increasing retweets by the general population over the course of the week [16]. But it must be noted that for the general population, only a 4% difference in retweets over the course of the week was observed, while the current study suggests that fake accounts are ~ 100% more active on Fridays than on any other day of the week. Thus, the fake account behavior observed here is quite distinct from that of the general population.

4.2 Creation Times

The difference in the update time distributions of the two datasets provides confirmation of the distinct nature of our generated fake profile set. It is, however, not entirely surprising that the two datasets have different update characteristics, given that the update time was a factor in filtering the dataset. A better measure of the difference between the two datasets is the distribution of creation times.

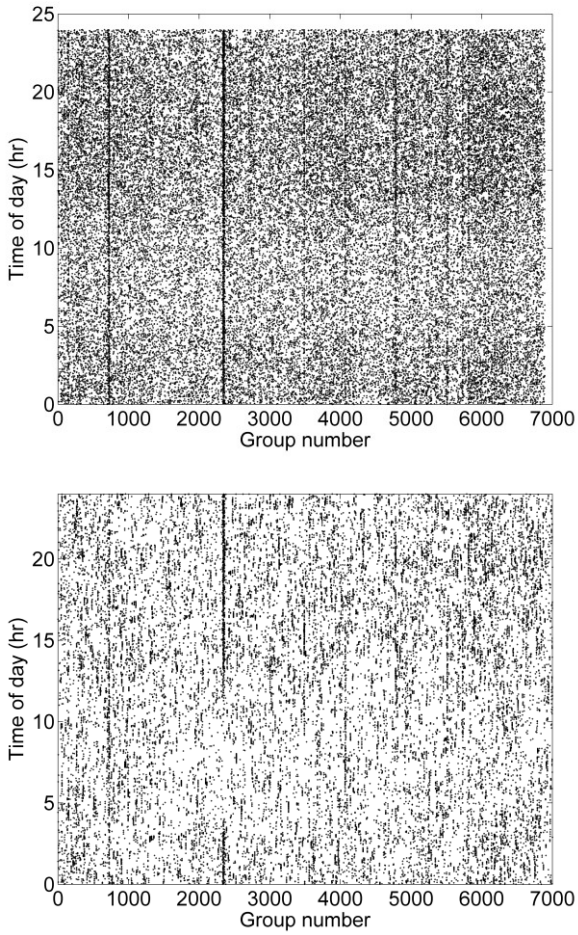
The distribution of days of the week when profiles in the two datasets were created, is shown in Figure 4. For the ground truth data, the distribution is nearly uniform (Figure 4a), with no preference for any particular day of the week. This suggests that a typical legitimate user would create a Twitter profile any day of the week. For the fake profile set, the creation days are biased towards the later part of the week (Figure 4b). While it is not obvious why this bias exists, it could feasibly relate to a possible manual element in the creation of these fake profiles.



(a) Ground Truth (b) Fake Profiles

Figure 4: Occurrence frequency of creation days for (a) ground truth and (b) fake profile datasets.

The creation times for the different groups in the two datasets are shown in Figure 5. The ground truth profile creation times were largely distributed uniformly during the day, with some reduction in the number of created accounts during the 5 to 10 hr time period (Figure 5a). The distributions of fake profile creation times (Figure 5b), however, were seen to be very different from the ground truth dataset. The creation times were found to be significantly more non-uniformly distributed during the day than the ground truth dataset, suggesting that the accounts in the fake profile set were largely created in batches.

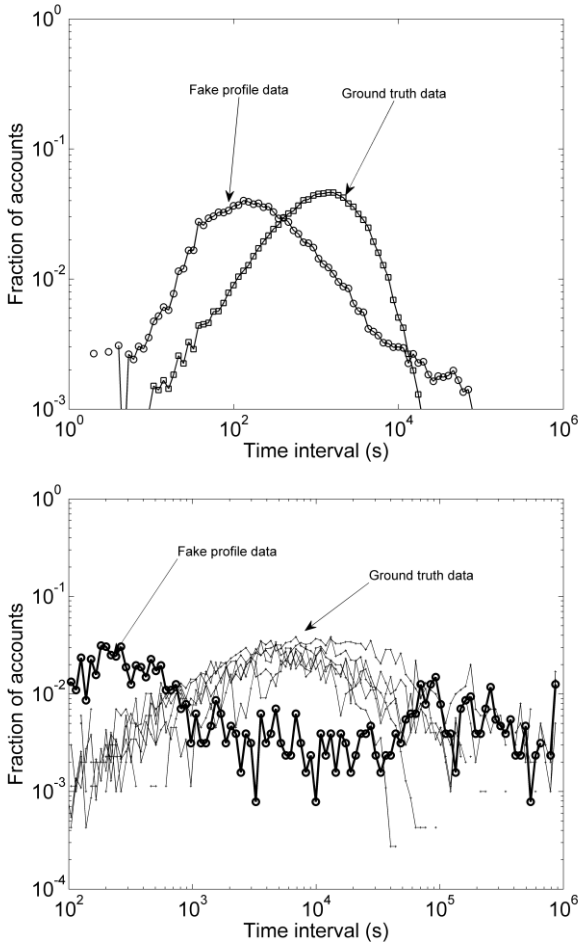


(a) Ground Truth (b) Fake Profiles

Figure 5: Comparison of creation times for (a) ground truth and (b) fake profile datasets.

The creation rate of fake profiles was investigated by first sorting the creation times of the two sets and then calculating the interval times between consecutive profile creations in each dataset (Figure 6). The median time intervals for the identified fake profiles were seen to be at least one order of magnitude less than that of the ground truth profiles (Figure 6a). The faster creation times of the fake profiles are consistent with our earlier observations of batch-creation of these accounts. The shortest creation time difference between two account creations in the fake profiles group was generally ~20-40 seconds, with some groups exhibiting even faster generation rates (3-5 seconds).

In Figure 6b, the creation time interval distribution of nine large groups in the fake profile list were compared against ground truth groups of similar sizes. The ground truth data for this comparison was obtained using a variety of aggregation approaches, including random collection and using profiles with matching popular first and/or last names. A two-sample t-test analysis confirmed that, independent of the aggregation approach used, the creation time distributions of the ground truth data were distinct from that of the fake profiles (p-value less than 5e-4). The median interval times of the ground truth data were seen to be significantly larger than that of the fake profiles. Another interesting observation that can be made is the similarity of the distributions at large time intervals ($> 10^5$). This could suggest that the profiles contributing to the tail of the fake profile distribution represent the false positive fraction of the fake profile set, and these were ~1% of the total number of profiles.



(a) Entire dataset (b) Limited dataset

Figure 6: The fraction of accounts created in a selected time interval for the fake profile and ground truth datasets for two cases: (a) Entire dataset of the fake profiles and a ground truth dataset of a similar size; (b) Limited dataset of manually-confirmed fake profiles compared against different groups of ground truth datasets.

4.3 URL Analysis

The profile_image_url attribute allows users to upload an image to personalize their account. To determine the diversity of the URLs in the fake and the ground truth datasets, the Shannon entropy values of the URLs were obtained for the different groups in the two datasets. The normalized standard deviation (Equation 1) of the entropies in each group was then determined.

The complementary cumulative distribution function (CCDF) at a desired normalized standard deviation of Shannon entropies (x) was calculated as:

$$(2)\ CCDF(x) = 1 - \frac{\sum_{x_i<x} E(x_i)}{\sum_{x_i<\infty} E(x_i)}$$

Where E is the number of groups with a selected normalized Shannon entropy (x_i). For the fake profiles, the CCDF (Figure 7) shows that the URLs are not very dissimilar compared to the URLs of the ground truth data. A large fraction of profiles in the fake group were actually seen to have similar or the same URL, resulting is very small values of normalized standard deviation of entropies.

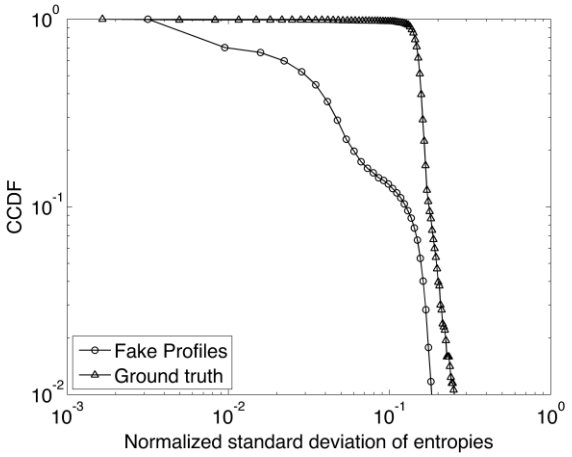


Figure 7: Complementary cumulative distribution function of the normalized standard deviation of the URL entropies.

For fake accounts, even when their URLs were very different, the images were seen to be the same. For one of the groups in the fake profile set containing 659 accounts, a collage of images from distinct URLs is shown in Figure 8. The number of distinct images for the 659 accounts were just 14. Thus, using image analysis of profile URLs, we could further refine our groups within the fake profile set.

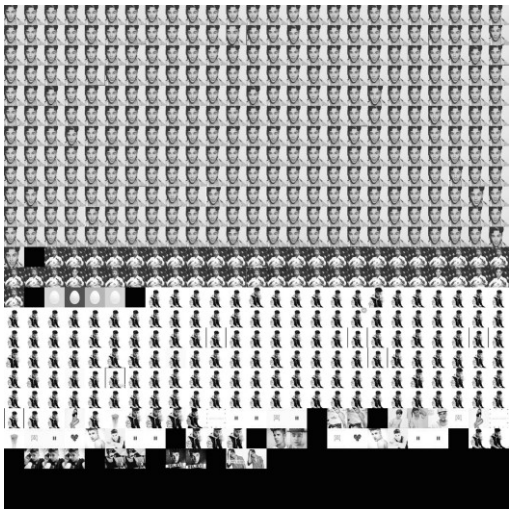


Figure 8: Collage of images from distinct URLs for a group of 659 accounts within the fake profile set

5. CONCLUSIONS

Using a crawler, a large Twitter user profile database of 62 million user accounts was obtained and analyzed to understand the characteristics of fake account creation. A highly reliable fake profile set was generated by grouping user accounts based on: matched multiple-profile-attributes; patterns in their screen names; and an update-time distribution filter. A subset of the accounts identified as fake by our algorithm were manually inspected and verified as all being fake (based on their Tweet activity). Analysis of the characteristics of the fake profile set revealed that these fake profiles were almost always created in batches and over intervals of less than 40 seconds. These accounts were created preferentially on some weekdays and during select times of the day, suggesting some manual element in the generation and maintenance of the profiles. The creation time characteristics of our identified fake profile set were very different from that of a ground truth dataset of similar size. The URLs of the fake profile set were seen to have lower diversity than the ground truth images. Our activity based profile-pattern detection scheme provides a means to identify potential spammers without detailed analysis of their tweets. One limitation of our approach is that it only identifies a relatively small percentage of fake accounts. But the low number of false positives that are likely in the obtained fake profiles make it an ideal seed database for use with social graph techniques for efficient spam detection. The effective employment of spam-detection approaches, such as ours, will enable Twitter to maintain a platform that is populated with real users and, thus, be a valuable tool for accurate data gathering and dissemination.

6. ACKNOWLEDGMENT

The authors would like to kindly acknowledge assistance from Benjamin Petroski for the URL analysis.

7. REFERENCES

[1] Parmelee, J. H., and Bichard S. L. 2011. *Politics and the Twitter revolution: How tweets influence the*

relationship between political leaders and the public. Lexington Books.

[2] Douceur, J. R. 2002. The sybil attack. *Peer-to-peer Systems*, Springer, 251–260.

[3] Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., and Dai, Y. 2011. Uncovering Social Network Sybils in the Wild. *ACM Trans. Knowl. Discov. from Data* 8, 1, 7.

[4] Kwak, H., Lee, C., Park, H., and Moon, S. (2010).What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*, 591–600.

[5] Stringhini, G., Wang, G., Egele, M., Kruegel, C., Vigna, G., Zheng, H., and Zhao, B. Y. 2013. Follow the green: growth and dynamics in twitter follower markets. In *Proceedings of the 2013 conference on Internet measurement*, 163–176.

[6] Jin, L., Takabi, H., and Joshi, J. B. D. 2011. Towards active detection of identity clone attacks on online social networks. In *Proceedings of the first ACM conference on Data and application security and privacy*, 27–38.

[7] Lee, K., Caverlee, J., and Webb, S. 2010. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 435–442.

[8] Stringhini, G., Kruegel, C., and Vigna, G. 2010. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, 1–9.

[9] Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference*, 6,12.

[10] Wang, A. H. 2010. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proc. of the International Conf.*, 2010, 1–10.

[11] Grier, C., Thomas, K., Paxson, V., and Zhang, M. 2010. @ spam: the underground on 140 characters or less. In *Proc. of the 17th ACM conference on Computer and communications security*, 27–37.

[12] Thomas, K., McCoy, D., Grier, C., Kolcz, A., and Paxson, V. 2013. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *USENIX Security*, 2013,195–210.

[13] Twitter\ Inc., "The Twitter Rules." .

[14] Gabielkov, M., and Legout, A. 2012. The complete picture of the Twitter social graph. In *CoNEXT Student '12 Proc. of ACM conference on CoNEXT student workshop*,19–20.

[15] S. D, "Is Twitter telling the truth about their 'active user' stats?" Nov-2013.

[16] Twitter\ Inc., "The Twitter Users." .