Finding Users We Trust: Scaling Up Verified Twitter Users Using Their Communication Patterns

Martin Hentschel **Omar Alonso** **Scott Counts**[†] **Vasileios Kandylas**
Microsoft, Sunnyvale CA 94089 [†]Microsoft Research, Redmond WA 98052
{hemartin, omar.alonso, vakandyl}@microsoft.com counts@microsoft.com

Abstract

We present a technique to identify Twitter users we trust to be regular Twitter users and not spam or fake accounts. The technique starts with an initial set of trusted users based on Twitter's verified users and recursively includes other users the trusted users communicate with. Conversations must be initiated by trusted users. We show that this technique produces a set of users that is over 200 times larger than Twitter's verified users. Our evaluation shows that the share of non-spam users within the resulting set of trusted users is more than 92% while the share of non-spam users within all tweeting users is 74%.

Introduction

When building applications on top of Twitter, a common problem is avoiding spam accounts and fake accounts. A simple solution is to use blacklists, to exclude spammers, or whitelists, to only include non-spam users. Blacklists are curated manually, via hand-tailored lists, or automatically, via machine-learned models. Whitelists are typically created manually, for example by relying on Twitter's verified users. All of these approaches have disadvantages: Manually curated blacklists do not scale because of time and work constraints. Automatically curated blacklists do not filter fresh accounts that use new spam patterns. Whitelists, based on Twitter's verified users, have a number of issues: (a) the set of verified users is too small (only $\sim 64,000$ users, see our data analysis), (b) they are dependent on Twitter to update the set of users, and (c) there are interesting accounts which are not verified but should be whitelisted (e.g., Dick Costolo, @dickc, the CEO of Twitter). These disadvantages led us to search for new approaches to find trustworthy, non-spam users on Twitter.

Our goal is to generate a set of Twitter users we can trust to be regular (non-spam) Twitter users. Ideally, this set of users should have the following properties: (a) it should be much larger than the current set of verified users on Twitter, (b) it should not contain spam accounts or fake accounts, and (c) it should be created automatically and on a continuous basis to include new, regular Twitter accounts over time. We call this set of users *trusted users* because we trust them to be regular Twitter accounts and not spam or fake accounts.

To compute the set of trusted users, we start with an initial set of trusted users that contains official Twitter-verified users only. We then recursively include the set of non-trusted users the trusted users communicate with. Conversations must be initiated by trusted users. The reason behind this approach is the following. Because verified users are manually verified by Twitter we can trust them to be regular Twitter users. Because trusted users start conversations manually, we can trust the users they communicate with to be regular Twitter users as well—a self-curating whitelist.

This paper presents an analysis of this approach and reports on our findings. We show that we can create a set of trusted users that is over 200 times larger than the current set of verified users and that this set contains less spam accounts and fake accounts than the complete Twitter universe. We describe the characteristics of the generated user set and present the results of a crowdsourcing study that evaluated our approach.

Related Work

Spam detection on social networks is an ongoing field of research. Yardi et al. were among to first to study spam on Twitter specifically (Yardi et al. 2009). In their study, the authors examine user age, tweet frequency, friend-follower ratio, and user clusters as possible features for spam detection. Follow-up work studies features such as text content and timing of posts (Chu et al. 2010), network distance and connectivity (Song, Lee, and Kim 2011), and keywords and URLs (McCord and Chuah 2011) to discriminate spammers from regular users. In addition, all of this work utilizes network-related features (typically friend-follower ratios) for spam detection. This is only possible (at scale) if the complete follower graph is available. Instead, we only need a stream of Twitter messages to extract conversations and do not need knowledge of the follower graph.

Similar to our approach, work on trust computation tries to find trustworthy users in social networks. Wilson et al. compare user interactions to social relationships and find that users of social networks communicate only with a small fraction of their "friends" (Wilson et al. 2009). They propose an interaction graph based on user interactions to capture more significant social relationships. Their interaction graph

is undirected whereas our method uses directed interactions. Lumbreras and Gavaldà propagate trust through the network graph for user recommendations (Lumbreras and Gavaldà 2012). Canini et al. propose a model for trust and influence computation based on friend-follower ratios (Canini, Suh, and Pirolli 2011). Both methods need knowledge of the follower graph.

There is a large body of work on Sybil defense in social networks. Systems such as SybilLimit (Yu et al. 2008) and SybilRank (Cao et al. 2012) leverage trust networks to defend against fake accounts by utilizing random walks over the undirected social network graph. Xie et al. detect legitimate users in online email services by analyzing email conversations between trusted users and non-trusted users (Xie et al. 2012). The authors adapt this method to Twitter using undirected mentions. Our work differs from these systems in that we use directed relationships.

Scaling up Verified Users

This section explains our approach of finding trusted users. We first explain verified users on Twitter. We then describe interactions through mentions and define conversations. Finally we formulate our algorithm to compute trusted users.

Verified Users. Twitter has the concept of verified users. A verified user is a user for which Twitter manually authenticates their identity. Typically these users are celebrities from many different areas including music, acting, politics, and journalism (Twitter 2013). Because Twitter manually and pro-actively verifies their authenticity, we can trust these users not to be spam or fake accounts. Visually, verified users are marked with a special symbol that represents a check mark.

Conversations. In this work we focus on directed tweets, a common type of interaction on Twitter. There are other types of interaction such as sending private messages, retweeting others' tweets, or favoriting tweets, which we will not further discuss. A directed tweet is a tweet that starts with the @-mention of another user. For example the Twitter user Alice may direct a tweet to user Bob by tweeting: "@Bob How are you?" Bob may answer this tweet by replying with: "@Alice Very well, thanks." To each tweet, Twitter assigns a unique tweet identifier (TweetID). When replying to a tweet (by clicking on the reply button on Twitter's webpage or apps), Twitter internally assigns a field InReplyToTweetID that specifies for which TweetID this tweet is a reply to.

We can now define conversations:

Definition 1. *A conversation between a user A and a user B happens when there are two tweets. The first tweet is an initial tweet from user A directed at user B. This tweet must start with user B's @-mention and the tweet must not have the field InReplyToTweetID set. The second tweet is the reply tweet of user B to user A. This tweet must have the field InReplyToTweetID set to the TweetID of the first tweet.*

By this definition, conversations are directional. Because user A initiates the conversation, the conversation is directed from user A to user B. This allows us to scale up verified users from within.

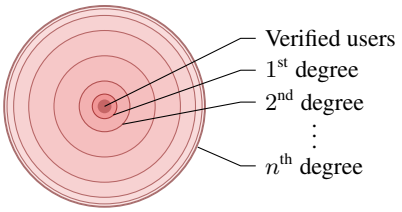


Figure 1: Scaling up verified Twitter users.

Algorithm. The idea of finding users we trust is to include users that verified users communicate with. It is important that the conversation is initiated by a verified user, not vice versa. Because verified users are hand-picked by Twitter and because they initiate conversations, we can trust the users they communicate with not to be spammers or fake accounts.

The algorithm to compute trusted users is the following. We start with a set trusted users that contains verified users only. We add to this set all (non-trusted) users the trusted users have conversations with. Conversations must be initiated by trusted users. The result set contains verified users and users that are 1 degree away from verified users. Repeating this process, we add all (non-trusted) users the trusted users have conversations with. We get a set of trusted users that includes users that are 2 degrees away from verified users. We repeat this process until the n^{th} degree, where n is a free parameter. The process is illustrated in Figure 1.

Data Analysis

We analyze our approach using Twitter data. We report on sizes of user sets per degree, follower distributions, and time span of conversations.

We extracted all public tweets in English language between August 1, 2013 and October 31, 2013. During that time period, the number of verified users that tweeted at least once (in English) was 63,639. The number of conversations (as defined in the last section) we extracted was 100,027,355.

Size of User Sets. First we analyze the size of the user sets per degree and see if our method converges. Figure 2 shows the cumulative number of users per degree. The number of verified users, our starting point, is 63,639 users. The 1st degree adds 136,145 users. The 4th degree is the largest set, adding 4,235,449 users. The 10th degree only adds 28,840 users. The curve flattens after 10 degrees at 13.6 million trusted users, which is 9.5% of all active users (that tweeted in English language during the 3 months time period we analyzed). Interestingly, the number of trusted users does not increase strongly after the 6th degree, which relates to the small-world phenomenon that states that every person is related to every other person by six or less degrees of separation (Kleinberg 2000).

Our method converges after 20 degrees. That is, after the 20th degree there are no more non-trusted users any trusted user communicates with. The final number of all trusted users up to the 20th degree is 13,608,816.

Follower distribution. Next we analyze the follower distribution of the different user sets. Figure 3 displays the fol-

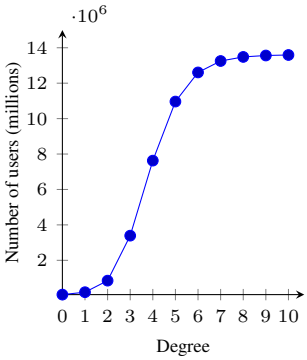


Figure 2: Cumulative number of users per degree.

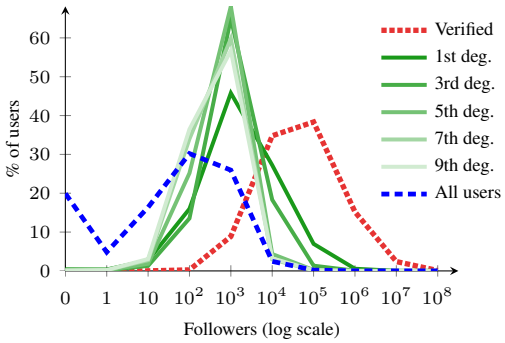


Figure 3: Follower distribution per degree.

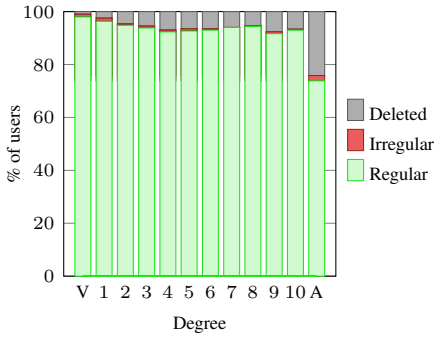


Figure 4: % of regular users within verified users (V), degrees 1–10, and all users (A).

lower distribution of verified users, users of the 1st, 3rd, 5th, 7th, and 9th degree, and all tweeting users. Verified users typically have many followers (e.g., 38% of verified users have between 10,001 and 100,000 followers). The main observation is that the follower distribution of trusted users within degrees 1–10 (not all shown in Figure 3) differs from the follower distribution of all users. This is a hint that the set of trusted users differs from the complete Twitter universe.

Conversation time spans. Last we analyze the time span of conversations. The table below shows the median as well as 90th, 95th, and 99th percentile of conversation time spans. Surprisingly, the median time span is short. Half of all directed tweets are replied to in under 4 minutes and 43 seconds. The 99th percentile is roughly 2 days and 16 hours. This allows us to capture at least 99% of all conversations by setting up a recurring process that analyzes 6 days of tweets (twice the interval of the 99th percentile) every 3 days.

Metric	Conversation time span
median	4 min 43 sec
90th percentile	5 hours 15 min 16 sec
95th percentile	12 hours 24 min 8 sec
99th percentile	2 days 16 hours 2 min 16 sec

Evaluation

We carried out a crowdsourcing study to gain more insights into the characteristics of the user sets. Most importantly, we wanted to know if the set of trusted users indeed contains more regular users than the average of all tweeting users. For the study, we sampled 2,400 random users (200 verified users, 200 users per degree 1–10, and 200 users among all users that tweeted at least once during the time we collected messages). The crowdsourcing task presented a link to a single user's Twitter profile along with the following four questions:

1. Is this user a regular Twitter user? (Regular, irregular, deleted.)
2. What is the type of the Twitter account? (Person, company, organization, product, location, other.)
3. What type of profile picture does this Twitter account have? (Face, body, cartoon version of a face, icon, landmark, other.)
4. What is the style of the biography? (Professional, informal, inappropriate, none.)

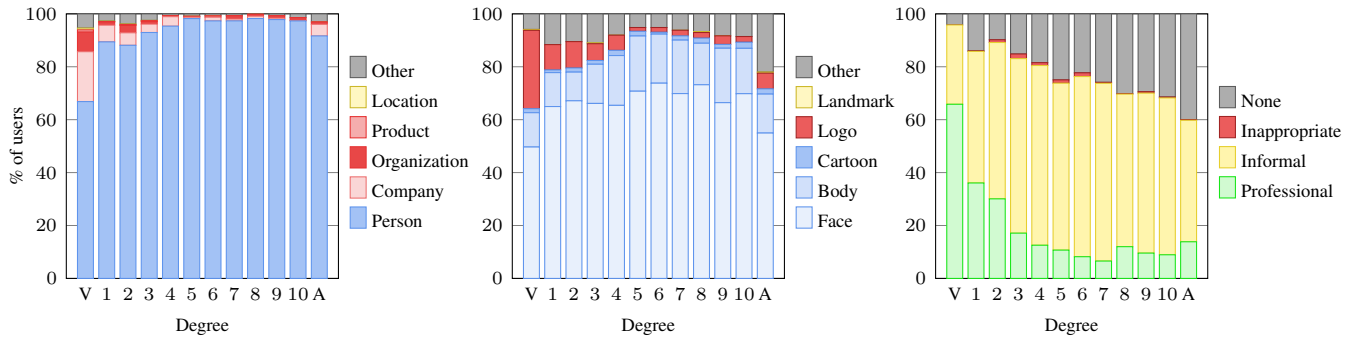
The study was executed through UHRS, a proprietary crowdsourcing platform. To ensure high-quality judgments, we used a manually created gold-standard set of 100 Twitter users to remove crowdsourcing spammers.

The crowdsourcing study was carried out between January 7–9, 2014. We collected 7,200 judgments, 3 judgments per Twitter user. Interrater agreement was high, with Fleiss' kappa values ranging from .70 to .90 across the four questions. We will now summarize the results of our study.

Regular Twitter users. The results of Question 1, "Is this user a regular Twitter user?", are presented in Figure 4. We make the following observations: (1) Verified accounts have the highest percentage of regular users (98%). Only 1% of accounts were judged irregular (e.g., Twitter user @modernwar) and 1% of accounts were deleted (e.g., @frenchie917). (2) Degrees 1–4 have a slightly decreasing share of regular accounts, declining from 96% to 93%. The share of regular users remains roughly constant at 93% for the remaining degrees. (3) Only 74% of all tweeting users are regular users. Most of the other users were deleted (24%).

Why do we need to consider deleted accounts in this question? In our observation, accounts are removed from Twitter mostly because they are spam, much less so because they chose to quit Twitter. Twitter deletes spammers that are reported to their service (Stringhini, Kruegel, and Vigna 2010). Because of the long time span between data gathering and the crowdsourcing study (4 months), we assume most spammers have been deleted by Twitter already. However, as soon as spam accounts are deleted, other spam accounts will be created. Therefore, if 26% of all tweeting users are either irregular or deleted, there is a constant 26% share of spam and other irregular accounts present on Twitter.

Account types. The results of Question 2, "What is the type of the Twitter account?" are shown in Figure 5a. Here, we make the following observations. (1) Verified accounts consist of 67% persons, 19% companies, 7% organizations, 1% products (e.g., Surface), 1% locations (e.g., Paris) and 5% others (e.g., music bands). (2) Degrees 1–10 as well as



(a) User types (b) Profile images (c) Quality of biographies

Figure 5: Characteristics of user sets of verified users (V), degrees 1–10, and all users (A). (Y-axes always in %.)

all users consist mostly of persons, with the first 4 degrees and all users consisting of around 4–5% company accounts.

It is interesting that trusted users are mostly persons instead of entities such as companies or organizations. This is useful if we want to aggregate user-generated content instead of entity-curated content (e.g., by news agencies).

Profile images. Figure 5b presents the results of Question 3, "What type of profile picture does this Twitter account have?" We make the following observations: (1) The most commonly used profile images on Twitter are faces. The second most commonly used images are bodies. Cartoon version of faces are only a small fraction of around 2–3%. (2) Because verified accounts have a high share of companies and organizations, we see a high share of logos as profile images of verified accounts. (3) Trusted users have the higher share of images that are faces or bodies of people.

Biographies. Figure 5c presents the results of Question 4, "What is the style of the biography?" Our observations are: (1) 66% of the verified users have professional biographies. Interestingly, this number drops sharply to 36% for the 1st degree users. Professionalism of biographies decreases until the 7th degree and slightly improves for the remaining degrees. (2) Only 4% of verified users have no biography. With increasing degrees, more users have no biographies. 40% of all tweeting Twitter users do not have a biography.

This shows that we cannot blindly rely on biographies created by trusted users. If we want to use biographies in our applications, we need to either curate these biographies (e.g., through crowdsourcing) or create our own summaries for these users (e.g., through expert profiling techniques).

Conclusions

In this paper, we showed a technique to create a set of users we trust to be regular Twitter users. Our technique uses directed conversations between users on Twitter. We start with an initial set of trusted users that contains verified users and we recursively include non-trusted users which trusted users communicate with. Conversations must be initiated by trusted users.

For our study, we analyzed 3 months of tweets and 100 million conversations. The computed set of trusted users is over 200 times larger than the set of verified users on Twitter. We showed that the set of trusted users (verified users and degrees 1–10) contains more than 92% regular Twitter accounts which are not spam or fake accounts. In contrast, only 74% of all tweeting accounts are regular accounts. An analysis of attributes of the resulting trusted user set shows they are people with faces or bodies for profile photos and professional or informal biographies.

References

Canini, K. R.; Suh, B.; and Pirolli, P. 2011. Finding credible information sources in social networks based on content and social structure. In *PASSAT/SocialCom*, 1–8.

Cao, Q.; Sirivianos, M.; Yang, X.; and Pregueiro, T. 2012. Aiding the detection of fake accounts in large scale social online services. In *NSDI*, 197–210.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2010. Who is tweeting on Twitter: human, bot, or cyborg? In *ACSAC*, 21–30.

Kleinberg, J. M. 2000. The small-world phenomenon: an algorithm perspective. In *STOC*, 163–170.

Lumbreras, A., and Gavaldà, R. 2012. Applying trust metrics based on user interactions to recommendation in social networks. In *ASONAM*, 1159–1164.

McCord, M., and Chuah, M. 2011. Spam detection on Twitter using traditional classifiers. In *ATC*, 175–186.

Song, J.; Lee, S.; and Kim, J. 2011. Spam filtering in Twitter using sender-receiver relationship. In *RAID*, 301–317.

Stringhini, G.; Kruegel, C.; and Vigna, G. 2010. Detecting spammers on social networks. In *ACSAC*, 1–9.

Twitter. 2013. FAQs about verified accounts. http://twitter.com/help/verified.

Wilson, C.; Boe, B.; Sala, A.; Puttaswamy, K. P. N.; and Zhao, B. Y. 2009. User interactions in social networks and their implications. In *EuroSys*, 205–218.

Xie, Y.; Yu, F.; Ke, Q.; Abadi, M.; Gillum, E.; Vitaldevaria, K.; Walter, J.; Huang, J.; and Mao, Z. M. 2012. Innocent by association: early recognition of legitimate users. In *CCS*, 353–364.

Yardi, S.; Romero, D.; Schoenebeck, G.; and boyd d. 2009. Detecting spam in a Twitter network. *First Monday* 15(1).

Yu, H.; Gibbons, P. B.; Kaminsky, M.; and Xiao, F. 2008. SybilLimit: A near-optimal social network defense against sybil attacks. In *SP*, 3–17.