SocialDistance: How Far Are You from Verified Users in Online Social Media?

Ang Li
Arizonal State University
anglee@asu.edu

Tao Li
Arizonal State University
tli@asu.edu

Yan Zhang
Arizonal State University
yanzhangyz@asu.edu

Lili Zhang
Arizonal State University
lilizhang@asu.edu

Yanchao Zhang
Arizonal State University
yczhang@asu.edu

ABSTRACT

Verified users on online social media (OSM) largely determine the quality of OSM services and applications, but most OSM users are unverified due to the significant effort involved in becoming a verified user. This paper presents SocialDistance, a novel technique to identify unverified users that can be considered as trustworthy as verified users. SocialDistance is motivated by the observation that online interactions initiated from verified users towards unverified users can translate into some sort of trustworthiness. It treats all verified users equally and assigns a trust score between 0 and 1 to each unverified user. The higher the trust score, the closer an unverified user to verified users. We propose various metrics to model the interactions from verified to unverified users and then derive corresponding trust scores. SocialDistance is thoroughly evaluated with large Twitter datasets containing 276,143 verified users and 19,047,202 unverified users. Our results demonstrate that SocialDistance can produce a non-trivial number of unverified users that can be regarded as verified users for OSM applications. We also show the high efficacy of SocialDistance in sybil detection, a fundamental operation performed by virtually every OSM operator.

CCS CONCEPTS

• **Information systems → Social networks**.

KEYWORDS

Quality of online social media; verified user; unverified user; interaction graph; sybil detection

ACM Reference Format:
Ang Li, Tao Li, Yan Zhang, Lili Zhang, and Yanchao Zhang. 2019. SocialDistance: How Far Are You from Verified Users in Online Social Media?. In *IEEE/ACM International Symposium on Quality of Service (IWQoS '19), June 24–25, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3326285.3329075

IWQoS '19, June 24–25, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6778-3/19/06…$15.00
https://doi.org/10.1145/3326285.3329075

1 INTRODUCTION

Online social media (OSM) such as Twitter, Facebook, and Instagram have penetrated into the fabric of everyday life. According to [1], OSM users accounted for about 71% of Internet users by the end of 2017; and there are over 330 million monthly active users on Twitter [5], 2.38 billion on Facebook [4], and 1 billion on Instagram [3] in Q1 2019. In addition to facilitating online social interactions, OSM have found tremendous applications in public and private sectors, including political campaigns, public relations, marketing, propaganda and counter-propaganda, crisis and emergency response, crowdsourcing, scientific and social studies, etc.

OSM users can be classified as either *verified* or *unverified*. Each verified user is identified by a special icon that varies on different OSM. It involves non-trivial effort to become a verified OSM user. For example, Twitter not only requires a verification applicant to confirm the phone number and email address but also asks the applicant to provide a personal photo, a photocopy of a government-issued ID, an associated personal/official website, and/or other identify-proof information [2]. In addition, the verification applicant needs to write a statement that describes his/her impact in an associated domain and why he/she wants to be verified. Other OSM all adopt similar procedures to handle verification requests. As a result, only high-profile OSM users—such as public figures, celebrities, journalists, politicians, governments, organizations, and businesses—may be eligible or bother to be verified. For instance, there are only about 306,885 verified accounts out of 336 million active users on Twitter in Q1 2018. Other popular OSM also have a very small percentage of verified users.

Verified OSM users largely determine the quality of OSM services and applications. For example, there are numerous fake users on various OSM who pose probably the greatest challenge to OSM operations and applications. Since verified users correspond to known people or organizations, they are commonly considered more credible and trustworthy than average unverified users. But the very small population of verified users in contrast to unverified users highly limits the information and services they can offer.

A natural question arises whether we can explore unverified OSM users to complement verified users for enhancing the quality of OSM services and applications. In this paper, we seek an interaction-based solution to this question. Our study is motivated by the observation that online interactions initiated from verified users towards unverified users can translate into some sort of trustworthiness. The more interactions an unverified user receives from verified users (e.g., retweets, mentions, and replies), the

more similar the unverified user is to a verified user with regard to trustworthiness. Although such directional interactions on OSM have been explored in [15, 35], they have not been used to quantify the similarity between unverified and verified users.

We make the following contributions in this paper.

- We conduct a comprehensive study about the interactions initiated by verified OSM users. A weighted directed *interaction graph* is constructed from verified users, unverified users, and the online interactions from verified users to unverified users. We explore various link metrics for the interaction graph and then analyze the graphical properties based on real Twitter data.
- We propose *SocialDistance*, the first technique to measure how far an unverified user is from verified users on OSM. In SocialDistance, each verified user is treated equally, and each unverified user is assigned a trust score between 0 and 1. The higher the trust score of an unverified user, the closer he/she is from becoming a verified user. We explore multiple metrics to assign trust scores based on the interaction graphs with different link metrics.
- We thoroughly evaluate SocialDistance with large Twitter datasets including 276,143 verified users and 19,047,202 unverified users in total. We show that SocialDistance can produce a non-trivial number of unverified users that can be regarded as verified users. We also demonstrate the efficacy of SocialDistance in sybil detection, a fundamental operation performed by virtually every OSM operator. In particular, with the output of Botometer [13] as the ground truth, we confirm that the unverified users with higher (lower) SocialDistance trust scores are less (more) likely to be sybil users. Finally, we discuss how SocialDistance and Botometer can well complement each other due to their very different technical principles.

SocialDistance can be a third-party service to OSM applications or be implemented by OSM operators themselves. For example, OSM operators can explore SocialDistance to significantly reduce their effort in sybil detection by focusing prohibitive human and computational resources on those unverified users with low SocialDistance trust scores.

The remainder of this paper is organized as follows. Section 2 describes the construction of interaction graphs with various link metrics. Section 3 presents the SocialDistance design. Section 4 evaluates SocialDistance with large Twitter datasets. Section 5 introduces the related work. Section 6 concludes this paper.

2 CONSTRUCTING INTERACTION GRAPHS

In this section, we describe the construction of interaction graphs for SocialDistance. Any OSM user can initiate interactions towards others, but not all interactions can be associated with social trust. For example, an unverified user on Twitter can be followed by arbitrary users (even spammers), and there is even an active underground market for purchasing fake followers who are almost all unverified users. In contrast, the followings from verified users are almost impossible to fake and thus can relate to social trust. So we only explores interaction graphs built upon online interactions initiated from verified users which are more trustworthy. We define two kinds of interaction graphs: non-dynamic and dynamic.

2.1 Non-Dynamic Interaction Graphs

Non-dynamic interaction graphs characterize one-time interactions associated with befriending requests. For an arbitrary set of verified and unverified users, we build a directed non-dynamic graph $\mathcal{G}(U, E')$, where U denotes the set of verified and unverified users, and E' refers to the set of directed links from verified users to unverified users. A link $e'_{i,j}$ from a verified user i to an unverified user j is formed when user i sent a befriending request to and was accepted by user j. The formation of any link is just a one-time event. In addition, a link may be broken due to unfriending or reestablished, which can also be considered rare or non-dynamic events. So we refer to $\mathcal{G}(U, E')$ as a non-dynamic interaction graph.

2.2 Dynamic Interaction Graphs

Dynamic interaction graphs correspond to temporal interactions initiated by verified users towards unverified users. For example, a verified user on Twitter may reply to, mention, or retweet an unverified user on a dynamic basis. For the same set U of verified and unverified users, we denote the directed dynamic interaction graph by $\mathscr{G}(U, E)$, where E is the set of directed interaction links. There is a link $e_{i,j}$ from a verified user i to an unverified user j if i has ever initiated interactions with j other than sending a befriending request. Assume that there are L types of dynamic interactions on OSM, such as retweets, replies, and mentions on Twitter. The dynamic interaction graph $\mathscr{G}(U, E)$ is weighted in contrast to $\mathcal{G}(U, E')$. There are many ways to define the weight $w_{i,j}$ for link $e_{i,j}$. In this paper, we explore the following metrics which are by no means exhaustive.

- *Unit-based*: This metric assigns $w_{i,j} = 1$ for every link $e_{i,j}$, indicating that the verified user i has ever interacted with the unverified user j.
- *Sum-based*: Intuitively speaking, the more interactions an unverified user received from a verified user, the closer their relationship. The sum-based metric uses the total number of interactions through a link as the link weight. Let $I^k_{i,j}$ denote the number of type-k interactions ($\forall k \in [1, L]$) over link $e_{i,j}$. We define $w_{i,j} = \sum_{k=1}^{L} I^k_{i,j}$.
- *Weighted average-based*: Different types of interactions on OSM may have diverse trust implications. For example, a verified Twitter user may often casually retweet someone else's tweets, but he/she is usually more careful and selective in mentioning other users in his/her own tweets. It is thus meaningful to associate higher trust with mentions than with retweets. Let λ_k denote the trust weight of type-k interactions. We define the weighted average-based link metric $w_{i,j} = \frac{\sum_{k=1}^{L} \lambda_k I^k_{i,j}}{\sum_{k=1}^{L} \lambda_k}$.
- *Consistency-based*: In general, consistent interactions can be viewed as a good representative of a stable relationship between two persons. To capture this observation, we partition a time span into epochs of equal length and then determine the percentage $\rho_{j,i}$ of epochs in which a verified

user i has interacted with an unverified user j. We define the consistency-based link metric $w_{i,j} = \rho_{i,j} \sum_{k=1}^{L} I_{i,j}^{k}$.

- *Epoch-based*: This metric reflects the observation that the temporal importance of an interaction and its implied social trust may age with time. We split an observation time period into equal-length epochs as well. Then we assign a time-decay factor to each epoch i, which is defined as

$$\Delta(i, t_{\text{end}}, t_{\text{start}}) = \frac{g(i - t_{\text{start}})}{g(t_{\text{end}} - t_{\text{start}})}, \quad (1)$$

where t_{end} and t_{start} denote the indexes of the first and last epochs, respectively, and $g(\cdot)$ represents a monotone non-decreasing function. As in [7], this paper considers the following three functions: (1) linear time decay: $g(t) = \alpha \times t$; (2) polynomial time decay: $g(t) = t^{\beta}$; and (3) exponential time decay: $g(t) = \exp(\gamma \times t)$. Let $\overrightarrow{\Delta}$ denote a row vector comprising time-decay factors, one for each epoch in the considered time span. Also let $\overrightarrow{I_{i,j}^{k}}$ denote a column vector, where each element indicates the number of type-k interactions from verified user i to unverified user j in one epoch of the considered time span. Finally, we define

$$w_{i,j} = \frac{\sum_{k=1}^{L} \lambda_k \times (\overrightarrow{\Delta} \cdot \overrightarrow{I_{i,j}^{k}})}{\sum_{k=1}^{L} \lambda_k}. \quad (2)$$

3 THE SOCIALDISTANCE SCHEME

In this section, we present the design of SocialDistance, a novel technique that explores the weighted directed dynamic interaction graph $\mathcal{G}(U, E)$ to quantify how far an unverified user is from verified users on OSM. The link weights such as $w_{i,j}$ can follow any definition in Section 2.2. Let U_{V} and $U_{\bar{\text{V}}}$ denote the set of verified and unverified users, respectively, for which we have $U = U_{\text{V}} \cup U_{\bar{\text{V}}}$.

SocialDistance assigns each unverified user $j \in U_{\bar{\text{V}}}$ a trust score $c_j \in [0, 1]$ according to the volume and features of interactions he/she received from all verified users in U_{V}. The higher c_j, the closer user j is to verified users, the higher trust others can have in him/her, and vice versa. In particular, $c_j = 0$ means that no verified user has ever initiated any interaction to user $j \in U_{\bar{\text{V}}}$, and $c_j = 1$ indicates that user $j \in U_{\bar{\text{V}}}$ is the closest to verified users who have initiated many interactions to user j. In addition, c_j can be dynamically adjusted with changing interactions. There are many ways to derive c_j. SocialDistance adopts the following trust metrics which can certainly be further expanded.

- *Strength of Interactions*: In reality, we tend to spend much more time communicating with our family, friends, or even colleagues, rather than with acquaintances or strangers. So the strength of interactions between an unverified user and a verified user is a good trustworthiness indicator of the unverified user. Based on this observation, we abuse the notation and redefine $\eta_j = \sum_{e_{i,j} \in E} w_{i,j}$ for each unverified user $j \in U_{\bar{\text{V}}}$.
- *Hybrid*: The hybrid metric considers both the strength of interactions and the number of verified users having interacted with each unverified user. Since the later factor is usually much smaller than the former, we use the logarithm of the interaction volume instead to balance these two factors. In particular, we redefine $\eta_j = \#_j \log(\sum_{e_{i,j} \in E} w_{i,j})$, where $\#_j$ denotes the number of incoming edges to each unverified user $j \in U_{\bar{\text{V}}}$.
- *Difference in Verified Users*: This metric considers the difference among verified users for interacting with the same unverified user. To incorporate this observation, we redefine η_j for each unverified user $j \in U_{\bar{\text{V}}}$ as $\eta_j = \sum_{e_{i,j} \in E} \theta_i \cdot w_{i,j}$, where θ_i denotes the weight assigned to verified user i that can also be defined in many ways. For example, our experiments use θ_i as the ratio of user i's outgoing interactions towards all unverified users to all his/her tweets.

To alleviate skewness, we adopt the log-max root to transform the raw trust values for each trust metric above. Let n denote the common logarithm of the maximum trust value. We transform η_j into $\eta'_j = (\eta_j)^{1/n}$ and then define the trust value c_j ($\forall j \in U_{\bar{\text{V}}}$) as

$$c_j = \frac{\eta'_j}{\max\{\eta'_i | i \in U_{\bar{\text{V}}}\}}, \quad (3)$$

which is in the range of [0,1].

4 EXPERIMENTAL STUDIES

In this section, we present experimental studies of interaction graphs and SocialDistance based on real Twitter datasets.

4.1 Datasets

We implemented a data crawler using Java based on official Twitter APIs. Since all verified users on Twitter are followed by the account named "Twitter Verified", we explored the API "GET friends/ids" and managed to get 276,143 of 301,000 verified users (about 91%). Then we randomly partitioned them into four subsets, each including at least 55,000 verified users. Finally, we crawled the followings, mentions, replies, retweets, and most recent 3,200 public tweets of each verified user. In addition, we obtained all the users who received interactions initiated by a verified user. The data-crawling process lasted from October 13, 2017 to December 1, 2017, and the crawled data cover from October 2012 to October 2017. Table 1 shows the basic statistics of the four resulting datasets. As we can see, the four datasets have very similar statistics, which confirm that our crawled datasets are quite representative. So our experimental results can be very trustworthy.

4.2 Construction of Interaction Graphs

Our experiments ran on a Dell OptiPlex 7010 desktop with Intel i7-3770 3.4 GHz CPU, 16 GB memory, and Windows 10 Pro. It took less than 3 minutes from constructing an interaction graph till generating the ranked list of trust scores for all unverified users in all cases. So SocialDistance incurs very low computational overhead.

We built non-dynamic and dynamic interaction graphs for each dataset. It is straightforward to build the non-dynamic interaction graph and dynamic interaction graphs with the unit-based and sum-based link metrics in Section 2. In contrast, it is a little more complex to construct dynamic interaction graphs with the other three link metrics.

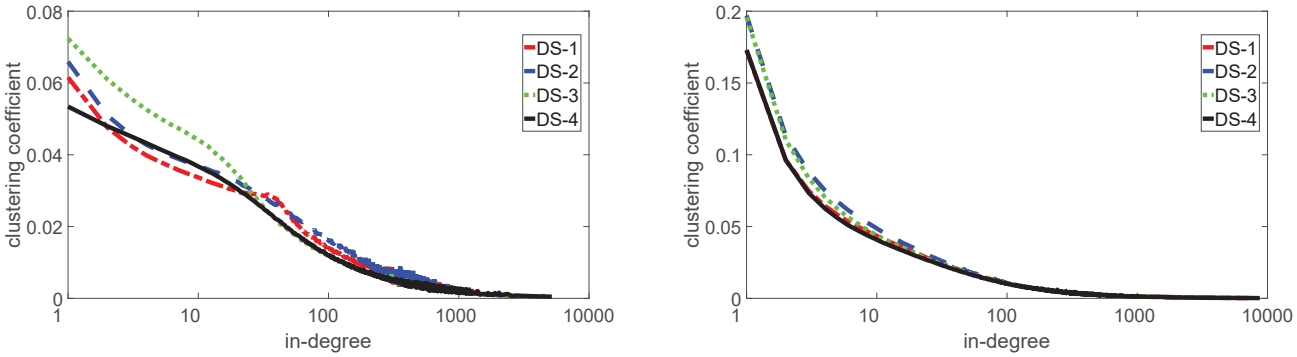
To construct the weighted average-based dynamic interaction graph, we should decide the weights for all interaction types. In

Table 1: Dataset descriptions.

	DS-1	DS-2	DS-3	DS-4
# of verified users	70,286	55,289	68,443	82,125
# of followings	48,551,836	44,060,094	47,793,809	80,989,248
# of mentioned users	10,701,077	9,258,628	10,393,613	14,015,006
# of mentioned unverified users	5,911,417	5,094,862	5,488,068	7,685,652
% of mentioned unverified users	55.2%	55.0%	52.8%	54.8%
# of replied users	7,502,955	6,693,586	7,681,198	10,348,813
# of replied unverified users	4,167,513	3,707,679	4,091,733	5,735,839
% of replied unverified users	55.5%	55.4%	53.3%	55.4%
# of retweeted users	6,145,964	4,870,857	4,937,427	6,104,188
# of retweeted unverified users	3,369,291	2,646,794	2,591,535	3,310,432
% of retweeted unverified users	54.8%	54.3%	52.3%	54.2%
# of mentioned+replied+retweeted users	13,576,502	11,618,797	12,811,752	16,956,141
# of mentioned+replied+retweeted unverified users	7,540,051	6,435,208	6,795,182	9,340,799
% of mentioned+replied+retweeted unverified users	55.5%	55.4%	53.3%	55.1%

Table 2: Graph density.

	DS-1	DS-2	DS-3	DS-4
non-dynamic interaction graph	0.000051	0.000062	0.000047	0.000051
unit-based dynamic interaction graph	0.000031	0.000037	0.000031	0.000027



(a) Non-dynamic interaction graph

(b) Unit-based dynamic interaction graph

Figure 1: Average clustering coefficients of unverified users with different in-degrees.

contrast to retweeting and replying, mentioning is more like a proactive action. People tend to be more prudent in selecting users to mention than to retweet or reply to. In addition, retweeting is more casual than replying. According to this observation, we assign the highest weight to mentioning, the second highest to replying, and the lowest to retweeting. In our experiment, we empirically assigned 1, 0.8, and 0.5 to mentioning, replying, and retweeting, respectively. The social-trust implications and relative weights of these interaction types may be worthy of an independent study.

To construct the consistency-based and epoch-based dynamic interaction graphs, we must first determine time epochs. Since our crawled data cover October 2012 to October 2017, we chose month as the epoch length for simplicity and obtained 60 epochs in total. For the epoch-based dynamic interaction graph, we further set the coefficients for linear, polynomial, and exponential time-decay functions to $\alpha = 1$, and $\beta = 2$, and $\gamma = 1$, respectively.

4.3 Analysis of Interaction Graphs

Since the topology of the five dynamic interaction graphs is the same, we focus on comparing the topological properties of the non-dynamic interaction graph and the unit-based dynamic interaction graph only. Due to space constraints, we only report the graph density

First, we check the graph density which is defined as the number of edges divided by $|U_V| \times |U_{\bar{V}}|$, where $|U_V|$ and $|U_{\bar{V}}|$ denote the number of verified and unverified users, respectively. The results

are shown in Table 2 for each dataset. We can see that the density of both graphs is very low, so both graphs can be viewed as sparse graphs. There are two main reasons for this result. First, unverified users are a few orders more than verified users in each dataset according to Table 1. Second, both graphs only consist of links (or interactions) initiated from verified users to unverified users. Since verified users are quite selective in interacting with unverified users, the number of links in both graphs are quite limited.

Second, we compare the out-degree and in-degree cumulative distribution functions (CDFs) of the two graphs. We found that about 80% to 90% of unverified users received interactions from less than 10 verified users, and most verified users initiated 100 or more interactions with unverified users. Furthermore, we utilize the method in [11] to fit the power-law distribution to the in-degree and out-degree CDFs. We found that the in-degree power-law fitting curves of the non-dynamic interaction graphs for all four datasets have the alpha values from 3.3 to 5 and the p-values between 0.127 and 0.986. In addition, the in-degree alpha values for the unit-based dynamic interaction graphs lie in $[3.1, 3.4]$, with the p-values between 0.346 and 0.895. These two facts indicate that both graphs conform to power-law distributions with regard to the in-degree, which coincides with our observation that most unverified users on Twitter received interactions from a relatively small number of verified users. In contrast, the alpha values of out-degree CDFs for the non-dynamic interaction graphs lie in $[2.1, 2.4]$, with the p-values ranging from 0.0129 to 0.03; and the alpha values of out-degree CDFs for the unit-based dynamic interaction graphs are all about 5.2, with the p-values all about 0.02. This later result manifests that the out-degree distributions of both graphs are less consistent with power-law scaling.

Third, we study the clustering coefficients of both graphs, which measure the extent to which nodes in a graph tend to cluster together. Our experiment adopted the definition of clustering coefficients in [19] which take values in $[0, 1]$. According to this definition, the more common neighbors two users have, the higher the clustering coefficient. In other words, a high clustering coefficient indicates that nodes tend to tightly form a small group. Table 3 shows the average clustering coefficients of verified users, unverified users, and all users for both graphs with respect to each dataset. We can see that the unit-based dynamic interaction graph always has a higher clustering coefficient than that of the non-dynamic interaction graph. This result further supports our conjecture that verified users tend to be more selective in interacting with unverified users than in following them. In addition, the correlations of the node in-degree and out-degree with the clustering coefficients are displayed in Fig. 1 and Fig. 2, respectively. Since the clustering coefficients decrease with both the in-degree and out-degree, the unverified users receiving lots of interactions from verified users and also the verified users initiating extensive interactions to unverified users are quite spread out on Twitter, which follows our intuition.

To summarize our results above, both non-dynamic and dynamic interaction graphs in our definitions are sparse graphs, but there are many verified users with outgoing links to a non-trivial number of unverified users. So we can explore this observation to identify those unverified users who are more trustworthy than other unverified users.

4.4 Evaluation of SocialDistance

We evaluate SocialDistance on each dataset in Table 1. Given six link metrics for dynamic interaction graphs and three trust metrics, we obtained 18 lists of unverified users in the descending order of their trust scores for each dataset. We analyze the results as follows.

4.4.1 Trust-score distribution. The absolute trust scores for different pairs of trust and link metrics cannot be directly compared because they have different implications. It is, however, still interesting to compare their trust-score distributions. Since the results across different datasets are quite similar, we only focus on DS-1 in Table 4, Table 5, and Table 6 for the three trust metrics, respectively. We have the following remarks on the results.

- For the Strength-of-Interaction trust metric, over 90% of unverified users have their trust scores in $[0, 0.2)$ for both sum-based and weighted average-based link metrics. The main reason is that most unverified users on Twitter seldom receive interactions from verified users.
- For the Hybrid trust metric, over 90% of unverified users also have their trust scores in $[0, 0.2)$ for both sum-based and weighted average-based link metrics. Since the number of verified users is the dominating factor for the Hybrid trust metric, this result is anticipated because most unverified users received interactions from a limited number of verified users.
- The majority of trust scores with the consistency-based link metric and all three trust metrics are below 0.4, corresponding to the observation that most unverified users lack consistent interactions from verified users.
- The epoch-based link metrics lead to much more evenly distributed trust scores for all three trust metrics. Therefore, they can provide much better trustworthiness distinction among unverified users than other link metrics.
- Although the fractions of unverified users whose trust scores are greater than 0.4 under all metrics are relatively small, a non-trivial number of unverified users still received a lot of interactions from verified users and thus can be trustworthy.

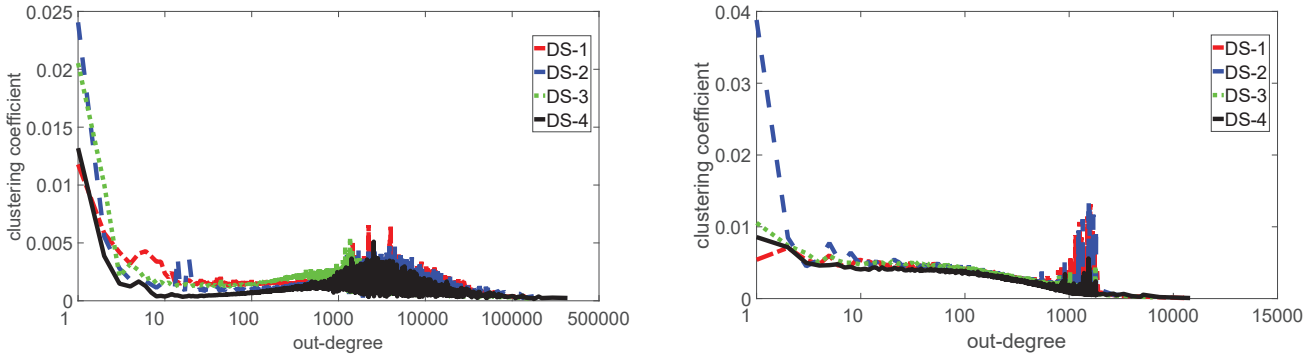
4.4.2 Impact of link metrics on rankings. Now we evaluate the impact of different link metrics on the rankings of unverified users. Given each trust metric, we use Kendall's tau-b [18] to conduct pairwise comparisons among the six ranked lists. Kendall's tau-b measures the ordinal association between two ranked lists, and its value ranges from -1 to $+1$. The larger the Kendall's tau-b value, the stronger agreement between two ranked lists, and vice versa. Due to space constraints and similar results, we only present the results for DS-1 with the three trust metrics in Table 7, Table 8, and Table 9, respectively.

We mainly have two observations. First, since the epoch-based link metrics take into consideration the temporal factor, they lead to much more dissimilar rankings to those of the other link metrics. Second, the exponential-decay metric produces the most dissimilar list to all the others, which coincides with our observation in the datasets that most interactions occurred in the early epochs of the observation period.

4.4.3 Impact of trust metrics on rankings. We also evaluate how different trust metrics affect the rankings of unverified users and

Table 3: Cluster coefficients of non-dynamic and unit-based dynamic interaction graphs.

	non-dynamic interaction graph				unit-based dynamic interaction graph			
average clustering coefficient	DS-1	DS-2	DS-3	DS-4	DS-1	DS-2	DS-3	DS-4
verified users	0.001833	0.001607	0.001967	0.001061	0.003506	0.003801	0.003777	0.003090
unverified users	0.04601	0.05019	0.05829	0.04444	0.1420	0.1650	0.1621	0.1422
verified+Unverified users	0.04670	0.04908	0.05734	0.04386	0.1407	0.1637	0.1605	0.1409



(a) Non-dynamic interaction graph

(b) Unit-based dynamic interaction graph

Figure 2: Average clustering coefficients of verified users with different out-degrees.

Table 4: Distribution of trust scores under Strength of Interaction (DS-1).

Link Metrics	[0,0.2) (#)	[0.2,0.4) (#)	[0.4,0.6) (#)	[0.6,0.8) (#)	[0.8,1] (#)
sum	94.66% (7,137,256)	5.22% (393,591)	0.12% (9,116)	0.00% (84)	0.00% (3)
weighted average	92.76% (6,994,362)	7.05% (531,385)	0.19% (14,190)	0.00% (100)	0.00% (13)
consistency	4.69% (353,945)	94.48% (7,123,966)	0.82% (62,000)	0.00% (133)	0.00% (3)
linear decay ($\alpha = 1$)	4.69% (353,944)	79.34% (5,982,334)	15.87% (1,196,733)	0.09% (7,014)	0.00% (25)
polynomial decay ($\beta = 2$)	5.92% (446,690)	80.36% (6,058,968)	13.62% (1,026,724)	0.10% (7,640)	0.00% (28)
exponential decay ($\gamma = 1$)	68.91% (5,196,070)	29.56% (2,228,469)	1.51% (113,677)	0.02% (1,824)	0.00% (10)

Table 5: Distribution of trust scores under Hybrid (DS-1).

Link Metrics	[0,0.2) (#)	[0.2,0.4) (#)	[0.4,0.6) (#)	[0.6,0.8) (#)	[0.8,1] (#)
sum	95.98% (7,237,231)	3.92% (295,703)	0.09% (6,934)	0.00% (180)	0.00% (2)
weighted averaged	94.23% (7,104,705)	5.64% (424,974)	0.13% (10,141)	0.00% (228)	0.00% (2)
consistency	96.18% (6,911,838)	3.69% (265,481)	0.12% (8,543)	0.00% (238)	0.00% (2)
linear decay ($\alpha = 1$)	75.94% (5,456,867)	23.38% (1,680,046)	0.68% (48,592)	0.01% (595)	0.00% (6)
polynomial decay ($\beta = 2$)	38.01% (2,731,368)	60.28% (4,331,845)	1.69% (121,772)	0.02% (1,113)	0.00% (8)
exponential decay ($\gamma = 1$)	61.55% (4,423,383)	36.06% (2,591,083)	2.35% (168,807)	0.04% (2,794)	0.00% (39)

report the Kendall's tau-b values in Table 10. The results show that all three trust metrics lead to similar trust rankings and have relatively less impact than link metrics. In other words, there are strong positive correlations among the three trust metrics.

4.4.4 Sybil detection. One important application of SocialDistance is to aid the detection of fake accounts (i.e., sybil or bot detection). Intuitively speaking, we would like that the unverified users with higher (lower) trust scores are less (more) likely to be sybil users and vice versa. To evaluate SocialDistance in sybil detection, the ground truth is needed but impractical to manually obtain because each of our ranked lists contains hundreds of millions of users. Therefore, we use the public APIs of Botometer [13] to automatically examine unverified users. Botometer is a public web service which can be used to evaluate if a Twitter account is a sybil user. Given a Twitter account, Botometer returns a score between 0 to 5 along with the Complete Automation Probability (CAP) which indicates the confidence level. The higher the score and corresponding CAP, the more likely the account is a sybil. To quantify the efficacy

Table 6: Distribution of trust scores under Difference in Verified Users (DS-1).

Link Metrics	[0,0.2) (#)	[0.2,0.4) (#)	[0.4,0.6) (#)	[0.6,0.8) (#)	[0.8,1] (#)
sum-based	3.23% (243,768)	93.69% (7,064,345)	3.06% (230,835)	0.01% (1,088)	0.00% (14)
weighted average-based	13.09% (987,311)	84.17% (6,346,792)	2.72% (204,772)	0.00% (1,153)	0.00% (22)
consistency-based	34.32% (2,588,010)	65.18% (4,914,789)	0.49% (37,133)	0.00% (116)	0.00% (2)
linear decay ($\alpha = 1$)	4.79% (361,507)	83.67% (6,309,084)	11.46% (864,275)	0.07% (5,159)	0.00% (25)
polynomial decay ($\beta = 2$)	6.68% (503,567)	83.39% (6,287,957)	9.85% (742,833)	0.08% (5,669)	0.00% (24)
exponential decay ($\gamma = 1$)	72.37% (5,457,047)	26.60% (2,005,431)	1.01% (76,468)	0.01% (1,097)	0.00% (7)

Table 7: Comparison of rankings under Strength of Interaction (DS-1).

Link Metrics	sum	weighted average	consistency	linear decay	polynomial decay	exponential decay
sum	—	0.87992	0.85469	0.63266	0.56051	0.26496
weighted average	0.87992	—	0.77105	0.67446	0.58925	0.26061
consistency	0.85469	0.77105	—	0.73408	0.65693	0.34668
linear decay ($\alpha = 1$)	0.63266	0.67446	0.73408	—	0.90595	0.54717
polynomial decay ($\beta = 2$)	0.56051	0.58925	0.65693	0.90595	—	0.63463
exponential decay ($\gamma = 1$)	0.26496	0.26061	0.34668	0.54717	0.63463	—

Table 8: Comparison of rankings under Hybrid (DS-1).

Link Metrics	sum	weighted average	consistency	linear decay	polynomial decay	exponential decay
sum	—	0.89782	0.89724	0.72650	0.65980	0.33111
weighted average	0.89782	—	0.84290	0.76507	0.68653	0.33117
consistency	0.89724	0.84290	—	0.75724	0.68718	0.33051
linear decay ($\alpha = 1$)	0.72650	0.76507	0.75724	—	0.91426	0.51266
polynomial decay ($\beta = 2$)	0.65980	0.68653	0.68718	0.91426	—	0.59323
exponential decay ($\gamma = 1$)	0.33111	0.33117	0.33051	0.51266	0.59323	—

Table 9: Comparison of rankings under Difference in Verified Users (DS-1).

Link Metrics	sum	weighted average	consistency	linear decay	polynomial decay	exponential decay
sum	—	0.88898	0.86520	0.68350	0.60733	0.29757
weighted average	0.88898	—	0.78823	0.71670	0.62793	0.28899
consistency	0.86520	0.78823	—	0.78256	0.69992	0.37326
linear decay ($\alpha = 1$)	0.68350	0.71670	0.78256	—	0.90139	0.53222
polynomial decay ($\beta = 2$)	0.60733	0.62793	0.69992	0.90139	—	0.62408
exponential decay ($\gamma = 1$)	0.29757	0.28899	0.37326	0.53222	0.62408	—

of SocialDistance, we adopt quite conservative parameters: if the Botometer score of a user is above 2 with CAP no less than 70%, he/she is considered a sybil user in the ground truth. Since Table 10 shows that the ranked list based on the exponential-decay link metric and the Difference-in-Verified-Users trust metric is overall most similar to those with other combinations of link and trust metrics, we focus on this list for sybil detection due to space constraints.

Figs. 3 and 4 show the number and fraction of sybil users with the top-K trust scores for different Ks and CAPs. According to our evaluations, the number of top-K sybil users remains zero until $K = 543$ and then undergoes a sudden jump when K is between 965 and 1,000 for all three CAPs. It is anticipated to see that the number of top-K sybil users increases with K. In addition, a larger CAP leads to a smaller number top-K users labeled as sybil users by Botometer with higher confidence. Furthermore, the fraction of top-K sybil users is sufficiently small for all cases, and it does not monotonically increase with K because the number of top-K sybil users grows slower than K.

Figs. 5 and 6 show the number and fraction of sybil users with the bottom-K trust scores for different Ks and CAPs. As expected, there are a significant number and fraction of sybil users among the unverified users with very low SocialDistance trust scores. In particular, about 47.84%, 35.36%, and 24.11% of the bottom-8000 unverified users are labeled as sybils by Botometer with CAPs equal to 70%, 80%, and 90%, respectively.

To sum up, the SocialDistance trust score is a highly credible indicator of the trustworthiness of an unverified user. OSM operators can explore SocialDistance to significantly

Table 10: Comparison of rankings under different trust metrics (DS-1).

Link Metrics	Trust Metrics	Strength of Interactions	Hybrid	Difference in Verified Users
sum	Strength of Interactions	—	0.88268	0.79652
	Hybrid	0.88268	—	0.88315
	Difference in Verified Users	0.79652	0.88315	—
weighted average	Strength of Interactions	—	0.87885	0.82420
	Hybrid	0.87885	—	0.87935
	Difference in Verified Users	0.82420	0.87935	—
consistency	Strength of Interactions	—	0.91016	0.77067
	Hybrid	0.91016	—	0.91016
	Difference in Verified Users	0.77067	0.91016	—
linear decay	Strength of Interactions	—	0.88944	0.86065
	Hybrid	0.88944	—	0.88944
	Difference in Verified Users	0.86065	0.88944	—
polynomial decay	Strength of Interactions	—	0.88905	0.87911
	Hybrid	0.88905	—	0.88905
	Difference in Verified Users	0.87911	0.88905	—
exponential decay	Strength of Interactions	—	0.94776	0.97424
	Hybrid	0.94776	—	0.94776
	Difference in Verified Users	0.97424	0.94776	—

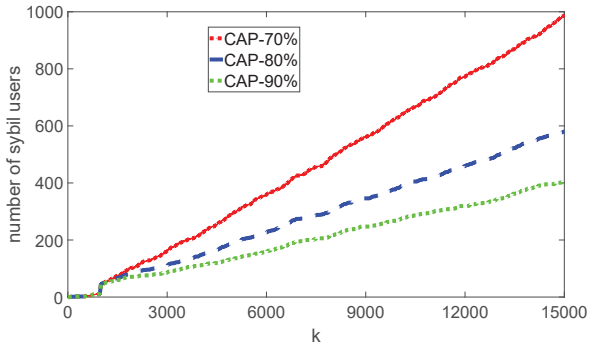


Figure 3: Number of top-K sybil users.

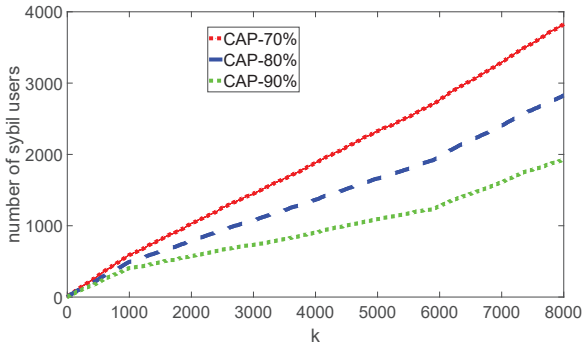


Figure 5: Number of bottom-K sybil users.

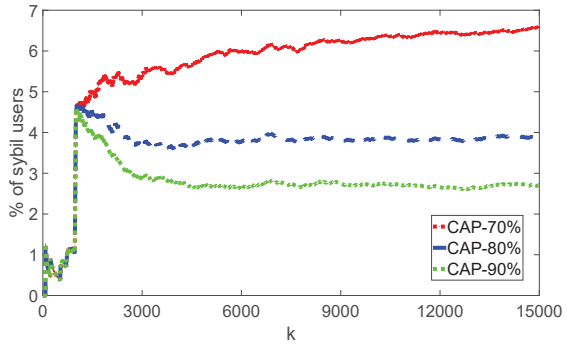


Figure 4: Fraction of top-K sybil users.

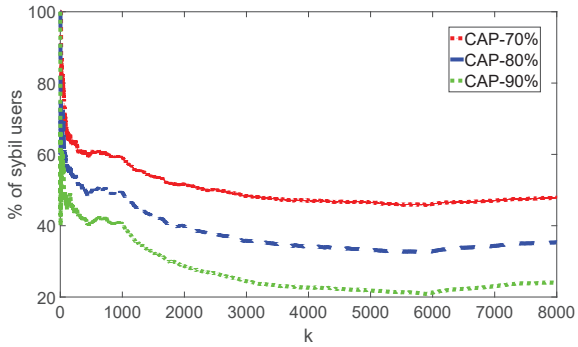


Figure 6: Fraction of bottom-K sybil users.

reduce their effort in sybil detection by focusing on those unverified with lower social scores. The parameter K that marks the region of most trustworthy users may be trained

through machine learning methods. We postpone this study to an extended version of this work.

4.4.5 Comparison with Botometer. Botometer provides a probabilistic assessment about whether a given Twitter user is a sybil. Now we briefly discuss how our SocialDistance scheme can well complement Botometer in sybil detection.

First, Botometer needs to access the most recent 100 tweets of a Twitter user to be evaluated. In practice, many users choose to keep their tweets private or seldom post tweets, but their interaction data are still much easier to obtain. So Botometer cannot evaluate such users, but SocialDistance can. For example, we found that Botometer cannot evaluate 11 users in the top-1000 unverified users output by SocialDistance. We manually confirmed that seven of them are non-sybil users and can be trustworthy.

Second, Botometer is a machine-learning technique that incorporates 1,200 features, while SocialDistance is a measurement-based technique purely based on interaction data. Neither technique is perfect and can have different false positives and/or negatives. Naturally speaking, we could consolidate their results for better sybil detection. For instance, we manually checked the first 100 unverified users (ranked high to low with regard to their trust scores) that are considered sybils by Botometer with CAP equal to 90%. We found that 49 of them are actually not sybils and can be trustworthy.

4.4.6 Practical interpretation of trust scores. Ideally speaking, we would like SocialDistance trust scores to coincide with real identities to some extent. For this evaluation, we manually checked the occupations of the top-1000 unverified users based on the exponential-decay link metric and the Difference-in-Verified-Users trust metric. There are about 100 of them whose occupations are difficult to tell. Most of the rest are in the categories of organizations, celebrities, politicians, and journalists. According to their tweets and number of followers, we can safely say that these users tend to be influential in their respective domains. So they could easily become verified users per Twitter's criteria as long as they apply. This result further confirms the value of SocialDistance.

5 RELATED WORK

This section outlines the prior work most related to SocialDistance.

There has been significant research on modeling, measuring, and analyzing the interactions in OSM. For instance, the topological structure of the following graph on Twitter is analyzed in [21] and shown to exhibit the structural characteristics of both an information network and a social network. In addition, the interaction graph on Facebook is studied in [27] and demonstrated to carry a much more accurate representation of meaningful peer connectivity on social networks. In [16], the authors investigated latent user interactions such as profile browsing on Renren, which used to be the largest online social network in China. These studies [16, 21, 27] focus on static or seldom-changing interactions typically associated with befriending requests. In contrast, Viswanath *et al.* [25] and Yang *et al.* [30] both presented an in-depth look at the changing dynamics of user interactions on Facebook. To the best of our knowledge, we are the first to build and study a weighted directed interaction graph built upon dynamic interactions (retweets, replies, and mentions) initiated by verified users towards unverified users on Twitter. Our graph analysis results may provide important insights to other relevant OSM research.

People have also explored interaction data on OSM for various applications. For example, the work in [28] infers hidden tie strength from online interactions, and the predictive model in [17] explores online interactions to identify strong ties. Other work [6, 10, 22] tries to measure online influence based on interactions. In addition, Zhang *et al.* [35] built an interaction graph on mutual interactions on Twitter and explored it to achieve sybil-resilient influence measurement. They also applied the similar interaction graph to social botnet detection [37], hidden location inference [34], and hidden age inference [33] on Twitter. Our work differs significantly from these elegant studies in both how the interaction graph is built and the research objective.

Also relevant is the extensive effort on sybil detection or defenses in various distributed systems, e.g., [9, 12, 24, 26, 31, 32]. There are two common assumptions. First, each node can be mapped into a vertex in an undirected social network graph where every edge corresponds to a human-established trust relation which is quite easy to fake. Second, the honest region is fast mixing and separate from the sybil region. These two assumptions have been challenged by many recent studies such as [8, 14, 20, 23, 29, 36]. In addition, these schemes [9, 12, 24, 26, 31, 32] cannot be directly applied to directed graphs. By comparison, SocialDistance explores a weighted directed graph built upon online interaction data only without any strong assumption, and its high efficacy in sybil detection is corroborated by Botometer [13], a practical online tool.

6 CONCLUSION

The qualify of OSM services and applications rely on the availability of many trustworthy OSM users. In this paper, we proposed SocialDistance, a novel scheme to identify unverified users on OSM that can be trusted as verified users. Thorough evaluations on large Twitter datasets confirmed that SocialDistance can produce a large list of unverified yet trustworthy users to well complement very limited verified users for rendering high-quality OSM services and applications. We also showed the high efficacy of SocialDistance in sybil detection.

7 ACKNOWLEDGE

We would like to thank the anonymous reviewers for their insightful comments that help improve the quality of this paper. This work was supported in part by Army Research Office under grant W911NF-15-1-0328 and US National Science Foundation under grants CNS-1514381, CNS-1619251, and CNS-1824355.

REFERENCES

[1] 2018. Number of social network users worldwide from 2010 to 2021 (in billions). https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] 2018. Verified account FAQs. https://help.twitter.com/en/managing-your-account/twitter-verified-accounts

[3] 2019. Most popular social networks worldwide as of April 2019, ranked by number of active users (in millions). https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[4] 2019. The Top 20 Valuable Facebook Statistics Updated April 2019. https://zephoria.com/top-15-valuable-facebook-statistics/

[5] 2019. Twitter: number of active users 2010-2018 | Statista. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[6] Isabel Anger and Christian Kittl. 2011. Measuring influence on Twitter. In *i-KNOW*. ACM.

[7] Eytan Bakshy, Jake Hofman, Winter Mason, and Duncan Watts. 2011. Everyone's an influencer: quantifying influence on Twitter. In *WSDM*. Hong Kong, China.

[8] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The socialbot network: when bots socialize for fame and money. In *ACSAC*. Orlando, FL.

[9] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. 2012. Aiding the detection of fake accounts in large scale social online services. In *NSDI*. San Jose, CA.

[10] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*. Washington, DC.

[11] Aaron Clauset, Cosma Shalizi, and Mark Newman. 2009. Power-Law Distributions in Empirical Data. *Society for Industrial and Applied Mathematics review* 51, 4 (2009), 661–703.

[12] George Danezis and Prateek Mittal. 2009. SybilInfer: Detecting Sybil Nodes using Social Networks. In *NDSS*. San Diego, CA.

[13] Clayton Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *WWW'16 Companion*. International World Wide Web Conferences Steering Committee, 273–274.

[14] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Korlam Gautam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. 2012. Understanding and Combating Link Farming in the Twitter Social Network. In *WWW*. Lyon, France.

[15] Martin Hentschel, Omar Alonso, Scott Counts, and Vasileios Kandylas. 2014. Finding Users we Trust: Scaling up Verified Twitter Users Using their Communication Patterns. In *ICWSM*.

[16] Jing Jiang, Christo Wilson, Xiao Wang, Wenpeng Sha, Peng Huang, Yafei Dai, and Ben Zhao. 2013. Understanding latent interactions in online social networks. *ACM Transactions on the Web* 7, 4 (2013), 18:1–18:39.

[17] Jason Jones, Jaime Settle, Robert Bond, Christopher Fariss, Cameron Marlow, and James Fowler. 2013. Inferring tie strength from online directed behavior. *PloS one* 8, 1 (2013), e52168.

[18] Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.

[19] Matthieu Latapy, Clémence Magnien, and Nathalie Vecchio. 2008. Basic notions for the analysis of large two-mode networks. *Social networks* 30, 1 (2008), 31–48.

[20] A. Mohaisen, H. Ttran, N. Hopeer, and Y. Kim. 2012. On the mixing time of directed social graphs and security implications. In *AsiaCCS*. Seoul, Korea.

[21] Seth Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. 2014. Information network or social network?: the structure of the twitter follow graph. In *WWW Companion*. Seoul, Korea.

[22] Aditya Pal and Scott Counts. 2011. Identifying Topical Authorities in Microblogs. In *WSDM*. Hong Kong, China.

[23] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *WWW*. Hyderabad, India.

[24] Nguyen Tran, Min Bonan, Jinyang Li, and Lakshminarayanan Subramanian. 2009. Sybil-resilient online content voting. In *NSDI*. Boston, MA.

[25] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna Gummadi. 2009. On the evolution of user interaction in Facebook. In *WOSN*. Barcelona, Spain.

[26] Wei Wei, Fengyuan Xu, Chiu Tan, and Qun Li. 2012. SybilDefender: Defend against sybil attacks in large social networks. In *INFOCOM*. Orlando, FL.

[27] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna Puttaswamy, and Ben Zhao. 2009. User interactions in social networks and their implications. In *EuroSys*. Nuremberg, Germany.

[28] Rongjing Xiang, Jennifer Neville, and Monica Rogati. 2010. Modeling relationship strength in online social networks. In *WWW*. ACM.

[29] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Zhao, and Yafei Dai. 2014. Uncovering Social Network Sybils in the Wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8, 1 (2014), 2.

[30] Zhi Yang, Jilong Xue, Christo Wilson, Ben Zhao, and Yafei Dai. 2015. Uncovering User Interaction Dynamics in Online Social Networks. In *ICWSM*. 698–701.

[31] Haifeng Yu, Phillip Gibbons, Michael Kaminsky, and Feng Xiao. 2010. SybilLimit: a near-optimal social network defense against sybil attacks. *IEEE/ACM Transactions on Networking* 18 (June 2010), 885–898. Issue 3.

[32] Haifeng Yu, Michael Kaminsky, Phillip Gibbons, and Abraham Flaxman. 2006. SybilGuard: defending against sybil attacks via social networks. In *SIGCOMM*. Pisa, Italy.

[33] Jinxue Zhang, Xia Hu, Yanchao Zhang, and Huan Liu. 2016. Your age is no secret: Inferring microbloggers' ages via content and interaction analysis. In *ICWSM*. Cologne, Germany.

[34] Jinxue Zhang, Jingchao Sun, Rui Zhang, and Yanchao Zhang. 2015. Your actions tell where you are: Uncovering Twitter users in a metropolitan area. In *IEEE CNS*. Florence, Italy.

[35] Jinxue Zhang, Rui Zhang, Jingchao Sun, Yanchao Zhang, and Chi Zhang. 2016. TrueTop: a sybil-resilient system for user influence measurement on Twitter. *IEEE/ACM Transactions on Networking* 24, 5 (October 2016), 2834–2846.

[36] Jinxue Zhang, Rui Zhang, Yanchao Zhang, and Guanhua Yan. 2013. On the impact of social botnets for spam distribution and digital-influence manipulation. In *IEEE CNS*. Washington, D.C.

[37] Jinxue Zhang, Rui Zhang, Yanchao Zhang, and Guanhua Yan. 2017. The rise of social botnets: attacks and countermeasures. *IEEE Transactions on Dependable and Secure Computing* (2017). in press.