Friend or Faux: Graph-Based Early Detection of Fake Accounts on Social Networks

Adam Breuer
breuer@g.harvard.edu
Harvard
Cambridge, MA, USA

Roee Eilat
reilat@fb.com
Facebook
Menlo Park, CA, USA

Udi Weinsberg
udi@fb.com
Facebook
Menlo Park, CA, USA

ABSTRACT

In this paper, we study the problem of early detection of fake user accounts on social networks based solely on their network connectivity with other users. Removing such accounts is a core task for maintaining the integrity of social networks, and early detection helps to reduce the harm that such accounts inflict. However, new fake accounts are notoriously difficult to detect via graph-based algorithms, as their small number of connections are unlikely to reflect a significant structural difference from those of new real accounts. We present the SYBILEDGE algorithm, which determines whether a new user is a fake account ('sybil') by aggregating over (I) her choices of friend request targets and (II) these targets' respective responses. SYBILEDGE performs this aggregation giving more weight to a user's choices of targets to the extent that these targets are preferred by other fakes versus real users, and also to the extent that these targets respond differently to fakes versus real users. We show that SYBILEDGE rapidly detects new fake users at scale on the Facebook network and outperforms state-of-the-art algorithms. We also show that SYBILEDGE is robust to label noise in the training data, to different prevalences of fake accounts in the network, and to several different ways fakes can select targets for their friend requests. To our knowledge, this is the first time a graph-based algorithm has been shown to achieve high performance (AUC>0.9) on new users who have only sent a small number of friend requests.

KEYWORDS

Social network analysis and graph algorithms; security, privacy, and trust; crowdsourcing and human computation; sybil detection.

1 INTRODUCTION

Online social networks are frequently targeted by malicious actors who create 'fake' or 'sybil' accounts for the purpose of carrying out abuse. Broadly, abuse is conducted in three phases: First, malicious actors create accounts. These accounts then need to establish connections with real users (e.g. by sending friend requests on Facebook). Once they establish sufficient connections, fake accounts can expose their social networks to a variety of malicious activities.

According to its latest Community Standards Enforcement Report [8], Facebook disabled over 2.2 billion such accounts in the first quarter of 2019. The vast majority of these accounts were disabled during or within minutes of account creation, and 99.8% were disabled before being reported by a Facebook user. Despite these impressive figures, the fraction of such accounts that survives registration-time classifiers and forms connections on Facebook still constituted roughly 5% of monthly active users in 2019 [8].

In this paper, we focus on social-graph-based detection of *new* fake accounts that manage to evade registration-time classifiers but have not yet made sufficient connections to perpetrate abuse. We define *new* accounts as those that are less than 7 days old or have sent fewer than 50 friend requests.

While the general problem of using the social graph to detect fake accounts is well-studied, existing algorithms typically do not apply to new accounts. This is because mainstream graph-based algorithms use a *structural difference* to detect fake accounts—namely, that fake accounts tend to have lower connectivity to real users. When popularized over a decade ago, this approach exhibited a key advantage: it was assumed that online social network companies only knew the true {fake, real} labels for a handful of users, and this handful was sufficient to seed a graph-based detection algorithm based on this structural difference. Nonetheless, one disadvantage is that real and fake users will only tend to exhibit this structural difference when they have made a reasonable fraction of their connections, so these algorithms tend to exclude new users from effective detection [1, 3, 19, 28].

However, both the resources available to online social networks and the challenges they face have evolved in the 14 years since the popularization of graph-based algorithms. For example, Facebook now possesses high-confidence {fake, real} labels for a majority of its active users—not just the handful assumed by existing graph-based algorithms. It is therefore now possible to use these additional labels to estimate not just structural differences, but also *individual-level differences* in how different users interact with real and fake accounts. Nonetheless, these labels are typically only available for users who have been active for at least several weeks. Thus, it is natural to consider whether today's greater data availability can inform algorithms capable of detecting *new* fake accounts.

Specifically, using Facebook's data on friending activity and known fake accounts, we observe that there are in fact important *individual-level* differences in how fake accounts interact with real
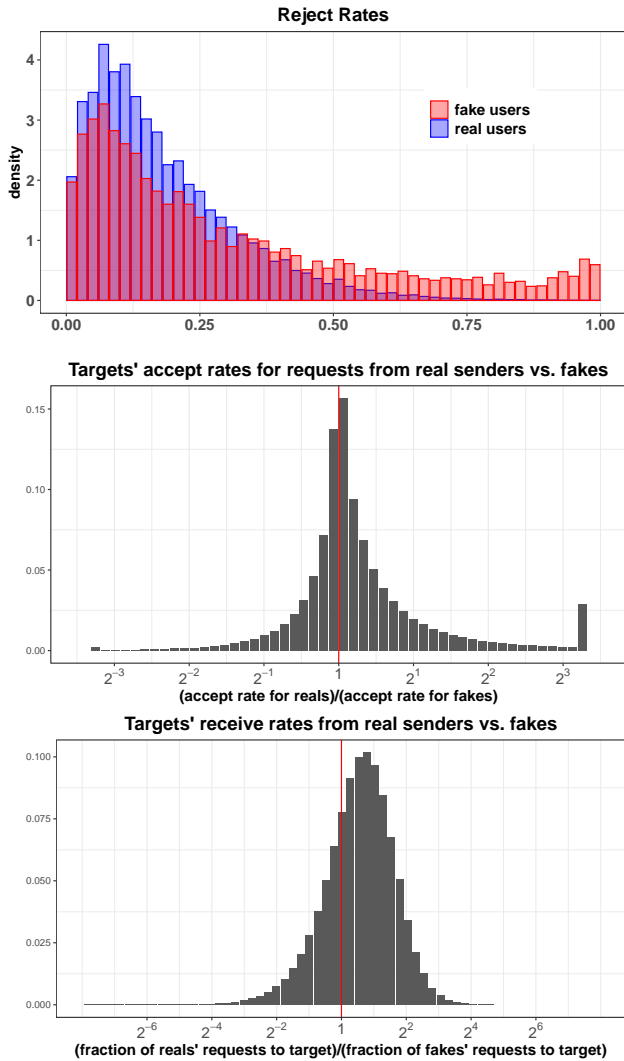
Figure 1: *Top:* **Distribution of rates at which real and fake Facebook users' friend requests are rejected.** *Middle:* **Distribution of Facebook users' ratios of their accept rates for incoming requests from reals and fakes. Users right of the red line at** *x*=1 **are more likely to accept a friend request from a real account than a fake.** *Bottom:* **Distribution of the ratios of the fraction of reals' requests and the fraction of fakes' requests that target each Facebook user. Mass right of the red line at** $x = 1$ **represents users who receive disproportionately more of real users' requests than fakes' requests.**

users, and how real users react to fake accounts. Observing these differences requires looking beyond aggregate statistics such as a user's overall reject rate for the friend requests she sends. For example, Fig. 1 *top* shows the distributions of reject rates for fake and real accounts. As the figure shows, many fakes (like many real accounts) either never or rarely have their friend requests rejected.

Disaggregating this data reveals two key differences: First, *for certain users* (but not others), whether a request comes from a

fake/real account is highly determinative of her decision to accept or reject. Fig. 1 (*middle*) plots the ratios of each Facebook user's rates at which she accepts friend requests from reals and fakes. Mass at $x = 1$ represents Facebook users who accept/reject fakes at the same rate as reals, which provides no information about the sender's {fake, real} label. Mass to the right of $x = 1$ represents users who are more likely to accept a request from a real user than one from a fake user by a factor corresponding to the x-axis. Thus, an unknown user whose friend request is accepted by such a recipient is more likely to be real. For example, the mass at $x = 2^3$ represents users who are 8 times more likely to accept a request from a real user than a fake. In fact, over 1/3 of Facebook users are at least 1.5 times as likely to accept either a real or a fake (i.e., mass outside $(\frac{1}{1.5}, 1.5)$), which provides a strong signal of their senders' labels. Because the tails of this distribution are very wide, we round all users with ratios outside $(\frac{1}{10}, 10)$ to these bounds in the plot.

Second, we observe a key difference in how some fake accounts select targets for their friend requests *differently* than real accounts. Specifically, certain users tend to be more or less frequently targeted by friend requests from fakes compared to real users, such that sending a request to such a target reveals information about the sender's label. Fig. 1 *bottom* plots the ratios of the fraction of reals' requests and the fraction of fakes' requests that target each Facebook user. Here, mass at $x = 1$ represents users who are equally likely to be selected as the recipient for a fake sender's friend request as for a real sender's friend request. Note that some users (mass to the left of $x = 1$) are preferred by fake senders, but many users (mass to the right of $x = 1$) are disproportionately likely to be selected by a real sender. Thus, an unknown user who sends a request to such a target to the right of $x = 1$ is more likely to be real, and vice versa. Here, 65% of users are at least 1.5 times as likely to be selected by a real vs. a fake or vice versa (i.e., mass outside $(\frac{1}{1.5}, 1.5)$), which provides a strong signal of their senders' labels.

These two key individual-level differences suggest a new means to detect *new* fake users despite their sparse connections: existing users are unequal in how their acceptances of friend requests reflect information about senders' real/fake labels, and real/fake senders deliberately target their requests to different sets of recipients.

Main contribution. In this paper we present SYBILEDGE, an algorithm to identify *new* fake accounts on social networks. SYBILEDGE returns the probability that each new user is a fake by aggregating over (I) her choices of friend request targets and (II) these targets' corresponding accepts/rejects. We show that this algorithm rapidly detects new fake users at scale on the Facebook network and outperforms state-of-the-art benchmark algorithms. We also show that SYBILEDGE is robust to label noise in the training data, to greater prevalence of fake accounts in the network, and to several different ways fakes can select targets for their friend requests. To our knowledge, this is the first time a graph-based algorithm has been shown to achieve high performance (AUC>0.9) on new users who have sent only a small number of friend requests.

Technical overview. SYBILEDGE classifies new users by combining three key components: First, SYBILEDGE estimates whether a new user is a fake by aggregating over her *choices of friend request targets*, giving more weight to targets to the extent they are preferred by other fakes vs. real users. Second, SYBILEDGE aggregates

over these *targets' responses* (accept/reject) to the user's friend requests, giving more weight to targets to the extent they respond differently to fakes versus real users. Finally, during these aggregations, SYBILEDGE gives more weight to choices of targets and their responses when we are more *confident* that they distinguish fakes from real users. Together, these three components give SYBILEDGE a natural means to elicit the information about a new user's {fake, real} identity from each of her friendship edges.

1.1 Related work

A variety of work has proposed graph-based algorithms to detect fake accounts [1–3, 5, 7, 10, 13, 14, 16–18, 21–32]. Mainstream graph-based algorithms typically proceed from the *homophily assumption*, which assumes that a pair of connected users shares the same {fake, real} label with high probability, such that fakes tend to be poorly connected to real users overall [1, 13, 24, 28, 30, 31]. Based on this assumption, a variety of graph-based algorithms attempt to propagate trust out from a small known set of trusted real users to unknown ones based on their connectivity to the known set.

More specifically, these algorithms typically propagate trust outwards via either random walks or Markov random fields (i.e. loopy belief propagation methods). Random walk based methods proceed on the basis of the assumption that unknown real users will be reachable in relatively few hops from the known set of real users, whereas reaching fake accounts requires additional hops on average. These algorithms therefore typically proceed via a series of short random walks on the network to partition nodes into real and fake sets on this basis. Random walk based methods include the seminal SYBILGUARD algorithm [31], as well as SYBILLIMIT [30], SYBILINFER [7], SYBILWALK [14], INTEGRO [3], and SYBILRANK [28]. Importantly, while random walk based approaches require either a known set of real users or a known set of fakes, they cannot leverage both at the same time. They are also considered less robust to misclassification (i.e. label noise) in the set of known users [24].

In contrast to random walk based methods, loopy belief propagation methods take a probabilistic view. These methods use Markov random fields to capture network structure and define a joint probability distribution over each node's label, which is iteratively updated to propagate labels from known fake or real nodes to unknown ones. Algorithms of this type include the seminal SYBILBELIEF [13], SYBILFUSE [11], and GANG [22]. Such algorithms are able to incorporate information about both known real and known fake nodes, and they are also robust to some noise in this set of known labels. Recently, Wang et al. proposed a hybrid algorithm, SYBILSCAR [24], based on this approach. SYBILSCAR iteratively propagates probabilistic estimates of unknown nodes' labels based on a known set of users of each type.

Importantly, both types of algorithm require that all users have had sufficient 'stabilization time' to make the majority of their connections such that they will exhibit the homophily assumption [1, 3, 19, 28]. Due to these requirements, evaluations of fake detection algorithms have often excluded users with less than e.g., 1 to 6 months of tenure on the social network [3, 28], which provides an ample 'grace period' for fake accounts to perpetrate abuse.

One partial exception is VOTETRUST [26]. VOTETRUST assumes that a majority of users (including those with known real labels) will be long-tenured, but this long-tenured set can be leveraged to classify a new user. For VOTETRUST, however, this advantage comes at a cost: VOTETRUST requires the additional dataset of (ideally) all historical friendship requests in the history of the social network, or at very least, sufficient historical requests such that the directed graph of requests is connected [26]. We note that data on old friendship requests is typically not among the datasets considered to be readily accessible for analysis in the current generation of online social networks.

The homophily assumption may also cause these algorithms to misclassify when some 'successful' fake accounts succeed in connecting to many real accounts. Recent research [9, 12] suggests that this phenomenon is relatively prevalent on social networks. For the same reason, the homophily assumption renders these algorithms vulnerable to *sampling attacks* whereby a malicious user defeats these algorithms by instructing some of her fake accounts to send many friend requests to real users (knowing that many of these accounts may be detected), then instructing her remaining fake accounts to send requests only to the subset of real users who were willing to accept requests from fakes. By generating fake users who are densely connected to real accounts, the attacker may succeed in convincing an algorithm that fake users are real [6, 26].

Paper organization. We present the SYBILEDGE algorithm in Section 2. We evaluate the performance of SYBILEDGE on the Facebook network in Section 3. We study SYBILEDGE's robustness to label noise in Section 4 and its robustness to the prevalence of fake accounts in Section 5. We conclude the paper in Section 6.

2 THE SYBILEDGE ALGORITHM

In this section we derive the SYBILEDGE (*Expert Decision Given Edges*) algorithm and its three key components: *target selection, target response,* and *confidence weighting.*

Preliminaries. Our goal is to determine the posterior probability p_i that a new user i is fake as a function of the set T_i of targets (friend request recipients) to whom she sends friend requests and their respective responses. Let $\delta_i \in \{S, B\}$ represent user i's label as a fake/sybil (S) or real/benign (B) account—that is, the label we want to learn. Let $x_{ij} \in \{0, 1\}$ denote target j's response to i's friend request (i.e. accept or reject), where $x_{ij} = 1$ denotes that j accepted i's request, and let $X_i \in \{0, 1\}^{|T_i|}$ denote the binary vector of all responses x_{ij} to i's requests from her set of targets T_i.

We denote by r_j^S (and r_j^B) an arbitrary fake (and real) sender's probability of choosing user j as the target when she sends her first friend *r*equest. We denote by R_i^S the vector of probabilities r_j^S for all of i's targets T_i, and by R_i^B the corresponding vector of probabilities r_j^B for all of i's targets. We denote the probability that j *a*ccepts a request from a fake or a real sender as a_j^S and a_j^B, respectively. Similarly, we denote by A_i^S the vector of accept probabilities a_j^S for all of i's targets, and by A_i^B the vector of accept probabilities a_j^B for i's targets.

Finally, suppose we know a *l*abelled set $L = L^S \cup L^B$ of known fake and real users, where L^S is the set of known fakes and L^B is the set of known real users, and suppose we have prior knowledge π_i of i's label.

Notation	Description
i	The user whose label we infer;
$\delta_i \in \{S, B\}$	User i's label: fake (S) i.e. sybil, or real (B) i.e. benign;
π_i	Prior on user i's probability of being a fake;
p_i	Posterior probability that i is a fake;
T_i	The set of targets i sends friend requests to;
r_j^S, r_j^B	Probabilities that user j is the target of an arbitrary fake and real sender's first friend request, resp.;
R_i^S, R_i^B	The vectors of probabilities r_j^S and r_j^B, respectively for all targets j whom i sends requests to;
$x_{ij} \in \{0, 1\}$	Target j's response to i's request (1 = accept);
X_i	Observed responses $[x_{i1}, \ldots, x_{i|T_i|}]$ of all i's targets;
a_j^S, a_j^B	Probabilities that target j accepts a friend request from a fake and from a real sender, respectively;
A_i^S, A_i^B	Vectors of probabilities a_j^S and a_j^B, respectively for all targets j whom i sends requests to;
L, L^S, L^B	Sets of known (users, fakes, reals);
$\rho_j, \rho_j^S, \rho_j^B$	Counts of known (users, fakes, reals) who sent requests to j;
f_j, f_j^S, f_j^B	Counts of known (users, fakes, reals) whose requests j accepted;
σ_j, ϕ_j	Priors on target j's quality as a classifier of users who send requests to j and are accepted by j, resp.

Table 1: Notation used in SYBILEDGE

2.1 Component I: a user's selection of targets

Here, we derive the first component of SYBILEDGE, which updates our estimate of whether user i is a fake based on whether she selects targets for her friend requests that are preferred by known fake versus known real senders. Specifically, we model that each new user selects a target for her first friend request via a draw from a multinomial distribution corresponding to her {fake, real} label: Fake users select each target j with probability r_j^S, but real users select j with probability r_j^B. We can then estimate the posterior probability that a sender i is fake based on the relative probabilities that a fake/real user would have selected i's set T_i of targets:

$$Pr\left[\delta_i{=}S|T_i, R_i^S, R_i^B, \pi_i\right] = \frac{\pi_i Pr\left[T_i|\delta_i{=}S, R_i^S\right]}{\pi_i Pr\left[T_i|\delta_i{=}S, R_i^S\right] + (1{-}\pi_i)Pr\left[T_i|\delta_i{=}B, R_i^B\right]} \tag{1}$$

Where R_i^S and R_i^B denote the vector of all probabilities r_j^S and r_j^B, respectively, for the targets $j \in T_i$ to whom user i sends requests.

We assume that conditional on the sender's label δ_i, the relative probability that a sender selects any target j is conditionally independent[1] of everything else, and that the count $|T_i|$ of friend requests the sender sends is independent of her label.[2] Technically, as a user i sends more friend requests, she reduces the remaining set of possible targets for her next friend request, making each

of them slightly more probable for the next request. However, because the network is very large compared to any user's number of friend requests, sampling targets with replacement is a very good approximation of sampling without replacement.[3]

Thus, we can then compute this **target selection component** via:

$$p_i = \frac{\overbrace{\pi_i}^{\text{prior}} \prod_{j \in T_i} r_j^S}{\pi_i \prod_{j \in T_i} r_j^S + (1 - \pi_i) \prod_{j \in T_i} r_j^B} \tag{2}$$

Here, the numerator is the joint probability of sender i's selections of friend request targets T_i given these targets' probabilities at which they are selected by *fake* accounts. The denominator then gives the total probability of i's selections of targets, which we compute by adding the probability of these selections given that the sender i was *fake* plus the probability that they occurred given that sender i was *real*. Therefore, the entire expression gives the *relative* probability that i is fake given her selections of targets, scaled by π_i, the prior probability that i is fake (for example, we might set this the overall fraction of fake accounts at Facebook).

The key intuition is that eq. 2 only updates our posterior estimate that i is fake to the extent her targets are selected by fake and real users at different rates (i.e. to the extent that i sends requests to targets who are further from $x{=}1$ in Fig. 1, *bottom*). In section 2.4, we show how to estimate targets' selection rates r_j^S and r_j^B.

2.2 Component II: targets' responses

Here, we derive the second component of SYBILEDGE, which updates our estimate of whether user i is fake based on her targets' *responses* to her friend requests. Suppose (unlike above) that a target is equally likely to receive a friend request from an arbitrary real or fake account, such that receiving a friend request from a user reveals no information about that user's label.[4] However, suppose we observe the targets' responses (acceptances/rejections) x_{ij} of i's friend requests, and targets may *accept* fake senders' requests at different rates than real senders' requests. If we know each target's probabilities a_j^S and a_j^B of accepting a request from a fake sender and from a real sender, respectively, then we can use the sequence of observed responses X_i to each of user i's friend requests to estimate the probability that she is fake. Denote by A_i^S and A_i^B the vectors of probabilities a_j^S and a_j^B, respectively, for all targets $j \in T_i$ to whom i sends requests. Assume that conditional on the sender's label δ_i, targets' responses are conditionally independent of everything else. We estimate the probability i is fake via:

$$Pr\left[\delta_i{=}S|X_i, A_i^S, A_i^B, \pi_i\right] = \frac{\pi_i Pr\left[X_i|\delta_i{=}S, A_i^S\right]}{\pi_i Pr\left[X_i|\delta_i{=}S, A_i^S\right] + (1{-}\pi_i)Pr\left[X_i|\delta_i{=}B, A_i^B\right]} \tag{3}$$

We now show how to compute this probability. Because a target may accept or reject a request, we first simplify notation by defining

---

[1] This is a standard assumption (see e.g. [20]). While not true in general (e.g. some targets are more popular), this assumption is advantageous as it may limit the effect any one observation has on model predictions, rendering it more adversarially robust.
[2] The assumption that a user's count $|T_i|$ of friend requests is independent of her label is advantageous because is allows SYBILEDGE to apply equally to accounts that are e.g. 1 and 7 days old—that is, accounts that have sent fewer/more friend requests.

---

[3] Absent this approximation, we would re-normalize r_j^S and r_j^B after each subsequent request, so e.g. the numerator in eq. 2 would become $\prod_{j \in T_i} r_j^S / (\sum_{k \in L \setminus T_i[:j]} r_k^S)$, where $T_i[:j]$ denotes the targets to whom she sent requests before j.
[4] In this case, $r_j^S = r_j^B, \forall j$, so eq. 2 factors to the prior π_i.

a function $\mathcal{A}(x_{ij}, \delta_i)$ that takes two inputs: target j's accept or reject x_{ij} of i's friend request, and the indicator δ_i of whether the source i is fake or real. $\mathcal{A}(x_{ij}, \delta_i)$ returns the probability that target j accepts i conditional on her {fake, real} label if we observe that i's friend request was accepted by j, or the complement of this probability if we observe that i was rejected by j:

$$\mathcal{A}(x_{ij}, \delta_i) = \begin{cases} a_j^{\delta} & if \ x_{ij} = 1; \\ 1 - a_j^{\delta} & if \ x_{ij} = 0 \end{cases}$$

Now we can compute this **target response component** via:

$$p_i = \frac{\overbrace{\pi_i}^{prior} \prod_{j \in T_i} \mathcal{A}(x_{ij}, \delta_i=S)}{\pi_i \prod_{j \in T_i} \mathcal{A}(x_{ij}, \delta_i=S) + (1 - \pi_i) \prod_{j \in T_i} \mathcal{A}(x_{ij}, \delta_i=B)} \quad (4)$$

Here, the product in the numerator is the probability of observing sender i's accepts and rejects conditional on her targets using the probabilities at which they accept and reject *fake* accounts. The denominator then gives the total probability of observing these accepts and rejects. At a high level, the entire expression captures the question 'did source i's accepts/rejects appear to be due to her targets treating her as they treated fakes or as they treated reals?'.

The key aspect to note is that eq. 4 only updates the posterior estimate that i is fake to the extent her targets respond differently to requests from fakes vs. reals (i.e. to the extent that i's targets are further from $x=1$ in Fig. 1, *middle*). In section 2.4, we show how to estimate targets' accept rates a_j^S and a_j^B for each class of senders.

2.3 The SybilEdge equation

Here we show how to compute the key equation in the SYBILEDGE algorithm, which combines these target selection and target response components to aggregate the information about a user's {fake, real} label contained in each of her friendship edges. Specifically, we say that the probability of observing each of i's accepted or rejected edges can be decomposed as (I) the probability that i would select the edge's target conditional on i's {fake, real} label, and (II) the target's response conditional on i's selection of the target and i's label. We thus determine the posterior probability i is a fake by aggregating over i's edges via **the SYBILEDGE equation**:

$$p_i = \frac{\overbrace{\pi_i}^{prior} \prod_{j \in T_i} \overbrace{\mathcal{A}(x_{ij}, \delta_i=S)}^{target\ response} \cdot \overbrace{r_j^S}^{selection}}{\pi_i \prod_{j \in T_i} \mathcal{A}(x_{ij}, \delta_i=S) \cdot r_j^S + (1-\pi_i) \prod_{j \in T_i} \mathcal{A}(x_{ij}, \delta_i=B) \cdot r_j^B} \quad (5)$$

Here, the products in the numerator give us the joint probability that (I) i selects the set of targets to whom she sends friend requests *as a fake user would select targets*; and (II) these targets respond with the accepts and rejects we observe given that they treat i as a fake when accepting/rejecting her. The products in the denominator then give the total probability that i selects these targets and

they respond with the accepts/rejects we observe. The SYBILEDGE equation therefore gives us the relative probability that the i's set of requests, accepts, and rejects are those of a fake user.

The SYBILEDGE equation thus captures our intuitions that a user i is more likely to be a fake to the extent that she selects targets who are preferred by fakes (for whom $r_j^S > r_j^B$), and also to the extent her targets respond differently to her requests than they usually respond to requests from reals (for whom $\mathcal{A}(x_{ij}, \delta_i=S) > \mathcal{A}(x_{ij}, \delta_i=B)$).

2.4 Component III: weighting target confidence

The discussion above assumes we know the true probabilities at which fakes and reals each select each target (r_j^S and r_j^B), and the probabilities at which each target accepts a request from either class (a_j^S and a_j^B). In practice, we must estimate these parameters from observed social graph data. Therefore, we introduce the final component of the SYBILEDGE algorithm: SYBILEDGE gives more weight to selections of targets and targets' responses not only as a function of the magnitude of the difference of targets' request and accept probabilities for fakes vs. real users (as above), but also as a function of our *confidence* in these differences.

SYBILEDGE accomplishes this confidence weighting as follows. First, consider how to compute a_j^S, a_j^B, that is, each target's probability of accepting a friend request from an arbitrary fake or real user. Suppose we know a set L^S of existing fakes and a set L^B of existing real users. The maximum-likelihood estimate of a_j^S is just target j's count of accepts of the requests she received from known fakes divided by the total count of these requests. However, if we used this approach for all targets, then the SYBILEDGE equation would give equal weight to a target who responded to only a few requests (i.e. a target whose accept rates we know with low confidence) and a target had responded to thousands of requests (whose accept rates we know with high confidence).

Therefore, we instead use estimators for these rates that, *in the absence of data to the contrary*, shrink a_j^S and a_j^B towards each other. This is because in the case where $a_j^S = a_j^B$, we say target j is equally likely to accept i's friend request regardless of whether i is a fake or a real, so the fact that j accepts i does *not* update i's probability of being a fake according to the *target response* component of the SYBILEDGE equation above.

Specifically, let f_j denote the count of j's *acceptances* of friend requests from users with known labels, and let f_j^S and f_j^B denote the counts accepted from known fakes and known reals, respectively. Let ρ_j denote the count of all friend requests that known users sent to target j, and let ρ_j^S and ρ_j^B denote j's count of friend requests from just known fake senders and just known real senders, respectively. We use estimators that **reweight target accept rates** based on our confidence via:

$$\hat{a}_j^{overall} = \frac{f_j}{\rho_j}; \ \hat{a}_j^S = \frac{f_j^S + \phi_j \cdot \hat{a}_j^{overall}}{\rho_j^S + \phi_j}; \ \hat{a}_j^B = \frac{f_j^B + \phi_j \cdot \hat{a}_j^{overall}}{\rho_j^B + \phi_j}$$

$$(6)$$

Where $\phi_j : \phi_j \geq 0$ is a 'confidence' prior on target j for the *target response* component of SYBILEDGE. Setting $\phi_j = 0, \forall j$ recovers the maximum likelihood estimators for $\hat{a}_j^S$ and $\hat{a}_j^B$, which compel the SYBILEDGE equation to place equal weight on targets for whom we

have observed more or less acceptance data.[5] In contrast, as we increase ϕ_j, we shrink $\hat{a}_j^S$ and $\hat{a}_j^B$ together *to a degree that is inversely proportional to the count of friend requests j responded to*, which compels SYBILEDGE to place less weight on targets who have only accepted/rejected a small number of reals or fakes in the past. In this case, the SYBILEDGE equation will tend to learn only from targets whom we have *repeatedly* observed accepting reals at a different rate than fakes (i.e. targets whose acceptance rates are known with high confidence). Similarly, by increasing ϕ_j for a particular target j but not others, we can selectively downweight the influence of target j's accepts/rejects on the model's predictions, which may be advantageous if we suspect target j of being a malicious or adversarial user.

SYBILEDGE uses a similar approach to place more weight in the *target selection* component on targets when we are more confident (i.e. have observed more data) about how they are selected by fakes vs. reals. Similarly to above, we could imagine maximum likelihood estimators for the probability r_j^S (or r_j^B) that a fake (or real) user will send her first friend request to target j by computing target j's count of requests received from known fakes (or reals) divided by the count of all requests sent by known fakes (or reals). But, as above, this approach would cause SYBILEDGE to give equal weight to a target who received only a few requests (i.e. a target whose rates we know with low confidence) and a target who received thousands of requests (whose rates we know with high confidence).

Therefore, we instead reweight target selection rates based on our confidence. Specifically, let ρ_L denote the count of all friend requests sent by known users, and let ρ_{L^S} and ρ_{L^B} denote the counts sent by known fakes and reals, respectively. Instead of the maximum likelihood estimators described above, we use the following to **reweight target selection rates:**

$$\hat{r}_j^{overall} = \frac{\rho_j}{\rho_L}; \; \hat{r}_j^S = \frac{\rho_j^S + \sigma_j \cdot \hat{r}_j^{overall}}{\rho_{L^S} + \sigma_j}; \; \hat{r}_j^B = \frac{\rho_j^B + \sigma_j \cdot \hat{r}_j^{overall}}{\rho_{L^B} + \sigma_j} \tag{7}$$

Here, $\sigma_j : \sigma_j \geq 0$ is our 'confidence' prior on target j for the *target selection* component of the SYBILEDGE equation: if we set $\sigma_j = 0, \forall j$, the SYBILEDGE equation places *equal* weight on friend requests sent to targets for whom we have observed more/less data; increasing $\sigma_j > 0, \forall j$ causes the SYBILEDGE equation to place *more weight* on targets for whom we have observed more data. More specifically $\sigma_j = 0$ recovers the maximum likelihood estimators for r_j^S and r_j^B, whereas increasing $\sigma_j > 0$ shrinks $\hat{r}_j^S$ and $\hat{r}_j^B$ towards each other *to a degree that is inversely proportional to the overall count of friend requests j received from fakes or reals*. This in turn causes the SYBILEDGE equation to place less weight on learning from targets for whom we have observed fewer friend requests (recall that the *target selection* component only updates the probability that a user is fake to the extent that fake users send requests to her targets at different rates than real users). Similarly, by setting σ_j higher for a particular target j compared to others, we downweight the influence of the selection of target j compared to other targets in the SYBILEDGE equation.

---

[5]There is a mathematical equivalence between these estimators and the Beta conjugate model in Bayesian inference.

2.5 The SYBILEDGE algorithm

These *target selection, target response*, and *confidence weighting* components form the SYBILEDGE algorithm:

Algorithm 1 SYBILEDGE
---
input $G_{requests}(V, E), G_{accepts}(V, E'), L, \pi, \sigma, \phi$
    for known user $j \in L$
        compute weighted request rates $\hat{r}_j^S$ and $\hat{r}_j^B$ per *eq. 7*
        compute weighted accept rates $\hat{a}_j^S$ and $\hat{a}_j^B$ per *eq. 6*
    for new user $i \in V \backslash L$
        compute p_i per *eq. 5*
    return p_i for all i

2.6 Choosing tuning parameters ϕ and σ

A key property of tuning parameters ϕ_j and σ_j is that, by increasing one relative to the other, we can tune SYBILEDGE to place more emphasis on learning from the set of targets a user chooses to send requests to relative to learning from whether those targets accept or reject. Specifically, as we increase $\sigma_j \to \infty, \forall j$, SYBILEDGE sets $\hat{r}_j^S \approx \hat{r}_j^B, \; \forall j$. The algorithm then ceases to update its estimate of i's label based on the set of targets i chooses, and we recover the *target response component* from the full SYBILEDGE algorithm. This in turn makes SYBILEDGE more robust to attack, as a fake user cannot 'appear real' by sending requests to recipients who typically are not targeted by fakes. However, this robustness comes at a cost in terms of SYBILEDGE's recall. Consider, for example, that when all σ_j are large, we will be less likely to detect a fake account that sends requests to targets who receive proportionally many requests from fakes, but who accept fakes at the same rate they accept reals.

2.7 SYBILEDGE properties

In addition to its strong performance on real and simulated Facebook data, SYBILEDGE exhibits six advantageous properties:

Rapid classification of new users. Previous methods typically require a lengthy 'stabilization period' before a new account can be classified, and are generally *less* likely to correctly classify a fake account that succeeds in making many friends with real users (even if those users are not discriminating). In contrast, SYBILEDGE becomes increasingly likely to identify a fake as she (1) sends more friend requests; (2) sends requests to more discriminating targets who accept fakes at a different rate than they accept reals (increasing the difference between a_j^S and a_j^B for i's targets); (3) sends requests to targets who are more often victimized by requests from fake accounts (increasing the difference between r_j^S and r_j^B); and (4) sends requests to targets who are more active users (for whom we have greater confidence in a_j^S and a_j^B).

Robustness to sampling attacks. A key property of SYBILEDGE is that targets only carry weight in the model to the extent that they receive and accept friend requests from real and fake users at different rates. Thus, a fake account cannot improve the SYBILEDGE's estimate of her probability of being fake even if she identifies and connects to many real users who accept e.g. *all* requests indiscriminately. Note that an indiscriminately accepting target has $a_j^S = a_j^B$,

which causes the target's accept or reject to appear on both the numerator and denominator of the *target response* component of the SybilEdge equation. This target's response then factors out and has no effect on our posterior estimate p_i of i's label.

Low complexity. SybilEdge has complexity $O(|E|)$ where E is the set of friend requests. Because social networks are typically sparse [13, 15], we have $O(|E|) = O(|V|)$. This compares favorably to state-of-the-art algorithms such as SybilBelief and SybilSCAR, which require $O(k|E'|)$, where k is the number of iterations (at least $O(\log(|V|))$ and E' is the set of accepted friend requests [13, 24].

Interpretability. Unlike mainstream sybil detection algorithms, SybilEdge is interpretable. For example, SybilEdge might classify a user as fake with high probability *because* her friend requests were rejected by specific users who tend to accept all requests from real users and reject those from fakes, and *because* she also sent requests to other users who are preferred targets of fakes. Such interpretability enables researchers to audit the model's classifications—an important precondition for disabling fake accounts.

Probabilistically labelled training data. SybilEdge accepts probabilistically labelled training data rather than binary $\{fake, real\}$ labelled data if desired. For example, an acceptance of a request from a user that data suggests is fake with probability 0.25 can be input as an acceptance of 0.25 fake users and 0.75 real users.

Robustness to label noise in the training data. In Section 4 below, we show that SybilEdge is robust to the presence of misclassified users in the training dataset L of known fake/real users.

3 EVALUATIONS

Our goal in this section is to show that SybilEdge achieves high performance (AUC>0.9) on new users at scale on the Facebook network, and that it significantly outperforms state-of-the-art benchmark algorithms. In subsequent sections we also show that SybilEdge is robust (i) to label noise in the training data, (ii) to greater prevalence of fake accounts in the network, and (iii) to several different ways fakes can select targets for their friend requests.

3.1 Evaluation on the Facebook network

We implemented SybilEdge at scale at Facebook, and we ran it in an offline evaluation setting on the global Facebook network. Specifically, we trained SybilEdge using just a three-month period of historical friending data from the last year. To train the model, we also used the historical set of real/fake labels from Facebook's internal fake classifiers from these three months. These labels include a highly calibrated real/fake label for all accounts that are >30 days old, which provided a label for all users in our three months of training data. We then tested SybilEdge by attempting to classify new users who joined Facebook anytime in the week immediately following these three months *using only this one week of data on their friending activity*. That is, we test SybilEdge's ability to detect new accounts who are each between 0 and 7 days old.[6] Because significant additional time has now passed since these users joined Facebook, they have since been labeled via our same set of fake classifiers. We compare SybilEdge's output to these known labels.

---

[6]To ensure fairness in this evaluation, for all new users i we set a prior π_i equal to the overall fraction of fakes among new Facebook users, and for all known targets j we set confidence priors $\sigma_j = \phi_j = const.$
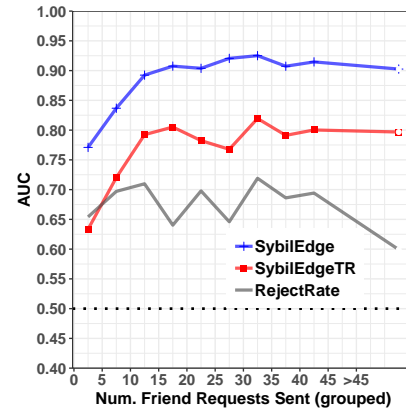
Figure 2: Performance of SybilEdge (blue) and SybilEdgeTR (red) on new global Facebook users partitioned by the number of friend requests they sent: $[[0, 5], [6, 10], \ldots, [41, 45], [46, \infty]]$.

Comparison metrics and benchmarks. Due to imbalance in the classes of fakes and real nodes (guessing 'all real' yields 95% accuracy), we adopt the standard approach and use ROC AUC to measure SybilEdge's performance [4, 24]. Recall that an AUC of 0.5 means a classifier is no better than random on the test set.

For comparison, we also include two benchmarks: RejectRate and SybilEdgeTR (Section 3.2 below adds additional benchmarks).

RejectRate. RejectRate just computes the AUC of each new user's fraction of sent friend requests that are rejected by her targets.

SybilEdgeTR. SybilEdgeTR is a simplified version of SybilEdge that uses only the *target response* component (eq. 4), and not the *target selection* component, so a new user's choice of targets does not affect the posterior probability she is fake (i.e., SybilEdgeTR is SybilEdge with $\sigma \rightarrow \infty$, see Section 2.6). SybilEdgeTR probes how much of SybilEdge's performance is due to *target response* alone.

Results. Fig. 2 plots AUC for groups of these new users partitioned by the number of friend requests they sent. SybilEdge and SybilEdgeTR improve in AUC as new users send more friend requests and converge to AUC's of 0.91 and 0.80, respectively, for all users who send more than 15 friend requests. We note that SybilEdge's high AUC values here mean that it successfully detected even those new users who joined Facebook on the last of the 7 days in the test set (i.e. who were only 1-day old at detection time). This evaluation is (to our knowledge) the first demonstration that a graph-based algorithm can detect fakes given just the small set of friend requests they attempt in their first days of activity.

We also manually inspected SybilEdge's errors, and we found that similarly to [26], the class of 'false positives' among new users who sent more than 15 requests reveals many 'real-but-spammy' users who abused friend recommendations by sending many unwanted requests. Thus, as in [26], we conclude that SybilEdge's 'false positives' can actually be desirable outputs.

We also note that, in contrast to some previous evaluations of graph-based algorithms on other social networks, the class of new fake Facebook accounts detected by SybilEdge cannot easily be
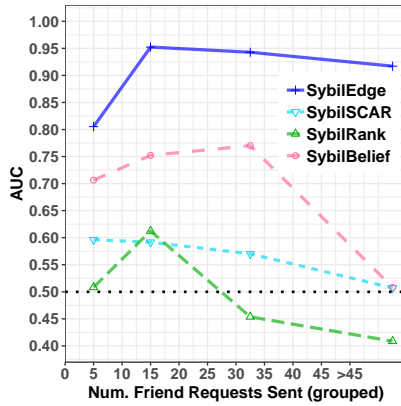
Figure 3: Performance of SYBILEDGE (dark blue) vs. state-of-the-art benchmarks on new Facebook users partitioned by the number of friend requests they sent: $[[0, 10], [11, 20], [21, 45], [46, \infty]]$.

Figure 4: SYBILEDGE performance under $0\%, 10\%, 20\%,$ and 30% label noise on new global Facebook users partitioned by the number of friend requests they sent: $[[0, 5], [6, 10], \ldots, [41, 45], [46, \infty]]$.

distinguished by basic network statistics such as reject rates. For example, the authors of VOTETRUST note that during their evaluation on the Renren network, fakes were distinguishable by their low average acceptance rate of 0.2 versus 0.8 for real users [26]. In contrast, reject rates yield AUC generally under 0.7 for the class of new users on Facebook. Thus, we conclude that SYBILEDGE was able to elicit much more information from a new user's sparse friendships by leveraging the differences in targets' selections and responses.

## 3.2 SybilEdge vs. state-of-the-art algorithms

We also compare SYBILEDGE to state-of-the-art benchmark algorithms on a Facebook network. Because benchmark algorithms have greater computational complexity than SYBILEDGE (see Section 2.7), we restrict the Facebook network in this evaluation to all users in a single country with roughly 1 million users. This restriction improves computational feasibility of the benchmarks, and it enables us to use their authors' publicly available code implementations for the sake of experimental transparency (see [13, 23, 24]).

We compare SYBILEDGE to:

SybilRank. SYBILRANK [28] is a state-of-the-art random walk based algorithm. Unlike SYBILEDGE, SYBILRANK uses only the graph of accepted friend requests and a set of known real users (nodes). As in [28], we run SYBILRANK for $\log 2(|V|)$ iterations.

SybilBelief. SYBILBELIEF [13] is a state-of-the-art loopy belief propagation algorithm. SYBILBELIEF uses the friendship graph of accepted friend requests and both known real users and known fakes. As in [13], we run SYBILBELIEF with edge weights set to 0.9.

SybilSCAR. SYBILSCAR [23, 24] is a recent probabilistic algorithm. SYBILSCAR uses the graph of accepted friend requests and both known real users and known fakes. We run both versions of this algorithm: SYBILSCAR-C with all weights equal to half the inverse of the average degree as in [23], and user-degree weighted SYBILSCAR-D. Each point in Fig. 3 reports the higher of their two AUC's.
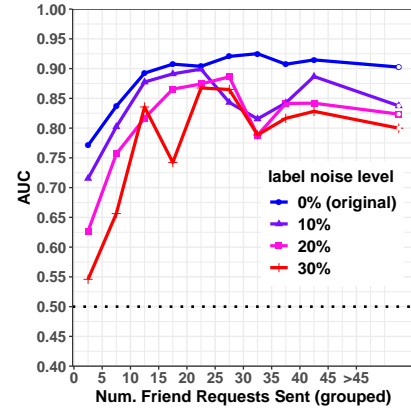
Results. Fig. 3 plots each algorithm's AUC for groups of new users partitioned by the number of friend requests they sent.[7] Overall, SYBILEDGE consistently outperforms all benchmarks regardless of the number of friend requests new users sent. Specifically, whereas SYBILEDGE achieves AUC>0.91 on all new users who have sent more than 10 friend requests, the best performing benchmark, SYBILBELIEF, achieves a maximum AUC of 0.77, and its performance degrades to no-better-than-random for new users who send >45 friend requests. Further investigation suggests that benchmarks' poor performance is largely due to the fact that some new fake users violate the homophily assumption and connect to many indiscriminately accepting real users, and the subset of new fake users who send the most friend requests (for whom the benchmarks' performance is lowest—see rightmost points in Fig. 3) are particularly likely to do so. In these cases, SYBILRANK tends to rank these fake users in particular as more likely to be real than low-degree real users (resulting in a low or even negative AUC), and SYBILBELIEF and SYBILSCAR tend to 'over-propagate' known real users' labels via these connections such that the majority of new users converge to identical '100% real' posteriors, resulting in AUC of 0.5.

4 ROBUSTNESS: LABEL NOISE

Robustness to label noise in the training data is a desirable and well-studied property of sybil detection algorithms [13, 24]. To test the robustness of SYBILEDGE to noise in a realistic setting, we repeat the evaluation of SYBILEDGE on the global Facebook network dataset (Fig. 2), but randomly flip up to 30% of known real and fake users' {fake, real} labels in the training data we use to compute each target's rates. Fig. 4 plots SYBILEDGE's performance on the global Facebook network with various levels of added label noise. Note that even with 30% of added label noise, SYBILEDGE still converges to >0.80 AUC on new users who have sent more than 20 friend requests. We therefore conclude that SYBILEDGE applies well even to

---

[7]We note that Fig. 3 uses fewer partitions than Fig. 2 to ensure each partition still has sufficient new fake accounts for evaluation on this one-country Facebook network.

social networks where training labels are known with significantly less confidence than they are at Facebook.

5 ROBUSTNESS: BEHAVIORS & PREVALENCE

Our goal in this section is to show that SYBILEDGE's performance advantage is also robust to conditions that differ from the current Facebook network—specifically, to (i) several different ways fakes can select targets for their friend requests, and (ii) greater prevalence of fake accounts in the network. To accomplish this, we designed a variety of synthetic friend request networks to capture a variety of ways fake users can choose targets for their requests. For each synthetic network, we then used real Facebook user data to realistically simulate how Facebook users would respond (accept/reject). Across these simulations, SYBILEDGE still rapidly converged to detect fakes after they sent only a small number of friend requests regardless of how they selected targets for these requests or their overall prevalence in the network. In all cases, SYBILEDGE outperformed state-of-the-art graph-based algorithms, whose performance changed markedly depending on how fake users chose targets, and who struggled to detect both low-degree fakes and fakes who succeed in friending less discriminating real users.

5.1 Robustness simulations setup

In each simulation, we set $n = 10000$ nodes (users) and randomly select 5% of them to be fakes, which matches Facebook's global fraction [8] of fake users (we later increase this to 10% to probe robustness to a greater prevalence of fake accounts). We randomly select 80% of nodes to have known fake/real labels and 20% to have unknown labels. This reflects a realistic 'difficult case' of a community where a full 20% of users are new. We then generate synthetic digraphs of friend requests using a variety of random graph models parameterized by Facebook data. This set of synthetic digraphs is selected to encompass a variety of possible strategies that fakes may deploy ranging from randomly targeting real users to preferentially targeting high-degree users or even users who have previously accepted friend requests from other fakes. For each friend request, we then draw an 'accept' or 'reject' based on mapping the simulated recipient to an actual Facebook user's accept rates for fakes/reals, which ensures that our simulated users' behaviors are consistent with actual Facebook users.

Benchmark algorithms We run SYBILEDGE and each benchmark algorithm from Section 3 on these graphs to classify the 'unknown' 20% of users (test set). We also include an additional benchmark:

VoteTrust. While we did not run VOTETRUST [26] on the Facebook network (Section 3.2) because it requires significant additional data[8], we include it in our simulations. VOTETRUST is an interesting benchmark because it is a random walk based algorithm, but like SYBILEDGE, VOTETRUST uses the directed graph of friend requests, accepts, and rejects. VOTETRUST detects fakes by propagating trust from known real nodes via random walks, then aggregating accepts/rejects of unknown users' requests weighted by their targets' trust scores.

5.2 Generating friend request graphs

First, we generate synthetic friend request graphs using various models, each parameterized by Facebook data, which capture various ways fakes can choose their targets. For each graph model, we vary the input parameters to produce a range of graphs with various average out-degrees (number of friend requests sent) from 1 to 50.

Erdős Rényi (n=10000). We generate friend request graphs using the directed Erdős Rényi model. We vary the probability p of an edge to yield a range of graphs where nodes' expected number of friend requests varies from 1 to 50 (i.e. $\frac{1}{n-1} \leq p \leq \frac{50}{n-1}$). These graphs capture a scenario where nodes send friend requests to targets chosen uniformly at random, but targets accept requests as in observed Facebook behavior (see Section 5.3 below).

FB-parameterized configuration model (n=10000). In practice, some users receive many more friend requests than others. To capture this in a realistic manner, we design directed configuration graphs by mapping each node uniformly at random to an observed Facebook user's count of actual friend requests. We then use each user's count as both her in-degree and out-degree distributions. The resulting graphs capture the scenario where we see a realistic distribution of friend requests, but fakes are careful not to betray their identities by sending many more requests than they receive.

FB-parameterized stochastic block model (n=10000). In practice, real users are much more likely to send requests to other real users than to fakes. We capture this by generating directed SBM graphs of friend requests with two clusters: one of fakes and one of reals. We set the probability of a friend request within- or across-clusters (the edge probability matrix $P_{2,2}$) to the observed ratios at which fakes/reals send requests to fakes/reals on Facebook.

FB-parameterized preferential attachment (n=10000). In practice, we observe that many fakes preferentially target users who have already been targeted by other fakes (see Fig. 1). To capture this, we design preferential attachment graphs of friend requests. First, we randomly map each simulated fake user uniformly at random to an actual observed fake Facebook user's receive counts from fake/real senders, and we map each simulated real user to corresponding data from a real Facebook user. We these counts as the preferential attachment process weights $\alpha_{fake,j}$ and $\alpha_{real,j}$, i.e., the *a priori* probability that each fake or real user, respectively, will *send* a friend request to target j. We then run the classic k-out preferential attachment algorithm until all nodes send k friend requests, and we generate a range of graphs with $k = 1$ to 50.

5.3 Modeling request acceptances/rejections

After generating a friend request digraph in each simulation, we generate the corresponding 'accept' or 'reject' for each request as follows: First, we map each simulated target node to a tuple of Facebook data describing a randomly selected Facebook user's historical rates at which she accepted requests from real users and fakes, respectively.[9] Here, we are careful to map each simulated fake to an actual fake Facebook user's rates and each simulated real

---

[8]Specifically, as described in Section 1.1, aside from the friendship graph and 3-month sample of users' friend requests we use to train SYBILEDGE and other benchmarks, VOTETRUST also requires complete older historical data on friend requests, which is not among the data that is considered readily accessible for analysis.

---

[9]We note that, due to the fact that millions of users have identical rates, this information is not identifying.
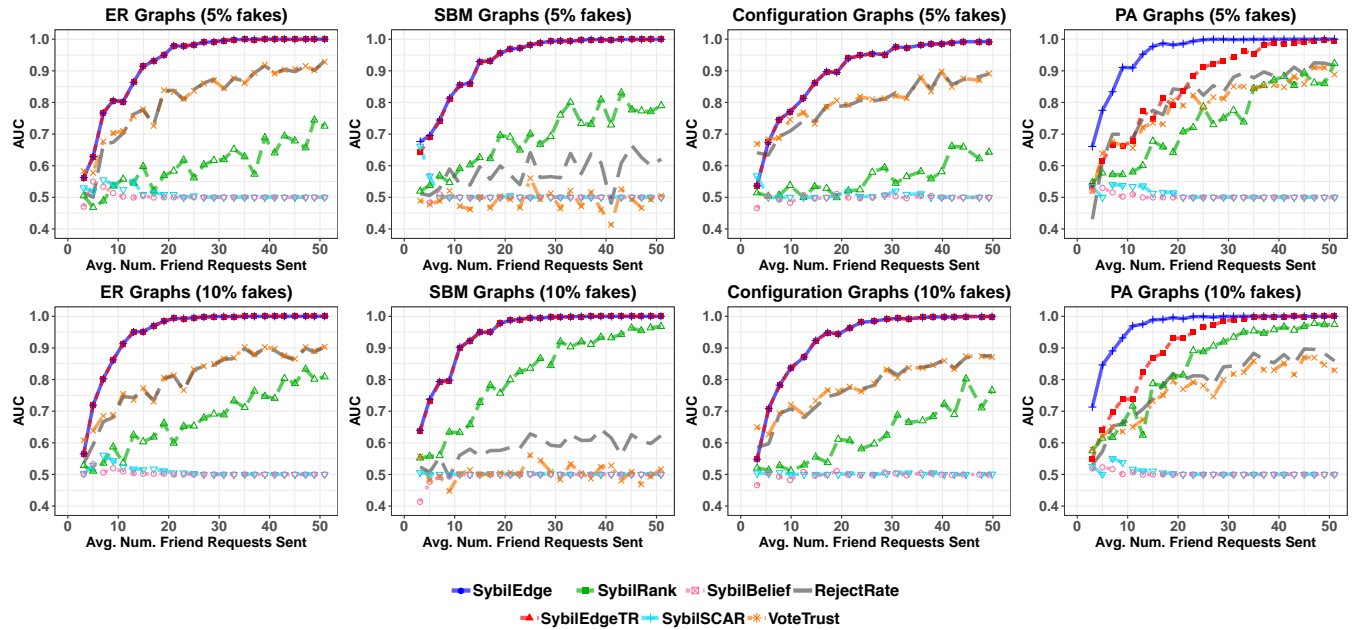
Figure 5: Performance of SYBILEDGE (blue) and SYBILEDGETR (red) versus benchmarks on random graphs with 5% fakes (top row) or 10% fakes (bottom row). Each point represents the AUC of one algorithm on a graph generated with one parameter combination. Different combinations yield graphs with different avg. number of friend requests sent by each user.

user to an actual real Facebook users' rates.[10] We use these rates as Bernoulli weights to draw 'accepts' or 'rejects' for her simulated friend requests from real users and fakes, respectively. This process synthesizes realistic 'accepts' and 'rejects' that match actual Facebook user-level distributions from fake and real accounts.

5.4 Robustness simulation results

Fig. 5 plots each algorithm's AUC versus the average user's count of friend requests sent (i.e. out-degree) for each graph model. Note that regardless of how fake users selected their targets, SYBILEDGE consistently achieved near-perfect classification after observing an *average* of 20 friend requests from each user (of which ~4 were sent by unknown users and thus excluded from training). Thus, after training on ~16 edges per known user, SYBILEDGE classified new fakes almost perfectly, *including those who sent only a couple of requests*, across all graph models. This suggests that SYBILEDGE's strong performance on the real Facebook network (Section 3.1) is quite robust to different ways fakes can select targets.

SYBILEDGE also reaped an additional performance advantage over SYBILEDGETR in preferential attachment graphs, as in these graphs fakes chose targets differently than real users. Per Section 3.1 and Fig. 2, this is consistent with SYBILEDGE's performance advantage on the real Facebook network.

In contrast, the performance of all benchmark algorithms was markedly inconsistent across the different graph models, and none matched the performance of SYBILEDGE on any graph model. We inspected their errors and found that, as with evaluations on real

data (Section 3.2), benchmarks' poor performance was largely due to the fact that new (simulated) users' sparse connections were insufficient to realize the homophily assumption. Also, as in the evaluations on real data, some real users accepted requests indiscriminately from many fakes, causing SYBILSCAR and SYBILBELIEF to 'over-propagate' known real users' labels out to other fakes, which resulted in many misclassifications. Additionally, all benchmarks struggled to distinguish fakes from low-degree real users.

Finally, SYBILEDGE's performance actually *improved* slightly when we increased the fraction of fake accounts in the data from 5% (*Fig. 5 top row*) to 10% (*bottom row*). This is because the increase in the fraction of fake users improves balance such that a greater fraction of targets in the training data receive requests from known fake users, so SYBILEDGE can better estimate targets' receive rates and accept rates for fakes when there have been fewer requests overall. This suggests that SYBILEDGE's performance is quite robust to even a marked increase in the current fraction of fake accounts.

6 CONCLUSION

We presented SYBILEDGE, a social-graph-based algorithm for the detection of new fake accounts on social networks. The class of new fakes has traditionally been overlooked by social-graph-based algorithms, which leverage network-structural differences to identify long-tenured fakes. However, we have shown it is possible to detect new fakes by leveraging small individual-level differences in how new fakes interact with other users, and how these users in turn react to new fakes. Because early detection limits the harm that such accounts can inflict, the development of such techniques is a promising new area for impactful research.

---

[10]For the preferential attachment graphs, we are careful to maintain the same mapping as during graph synthesis.

REFERENCES

[1] Muhammad Al-Qurishi, Mabrook Al-Rakhami, Atif Alamri, Majed Alrubaian, Sk Md Mizanur Rahman, and M Shamim Hossain. 2017. Sybil defense techniques in online social networks: a survey. *IEEE Access* 5 (2017), 1200–1219.

[2] Lorenzo Alvisi, Allen Clement, Alessandro Epasto, Silvio Lattanzi, and Alessandro Panconesi. 2013. Sok: The evolution of sybil defense via social networks. In *2013 ieee symposium on security and privacy*. IEEE, 382–396.

[3] Yazan Boshmaf, Dionysios Logothetis, Georgos Siganos, Jorge Lería, Jose Lorenzo, Matei Ripeanu, and Konstantin Beznosov. 2015. Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs.. In *NDSS*, Vol. 15. 8–11.

[4] Yazan Boshmaf, Matei Ripeanu, Konstantin Beznosov, and Elizeu Santos-Neto. 2015. Thwarting fake OSN accounts by predicting their victims. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*. ACM, 81–89.

[5] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Kamesh Munagala. 2015. Combating friend spam using social rejections. In *2015 IEEE 35th International Conference on Distributed Computing Systems*. IEEE, 235–244.

[6] Qiang Cao and Xiaowei Yang. 2013. SybilFence: Improving social-graph-based sybil defenses with user negative feedback. *arXiv preprint arXiv:1304.3819* (2013).

[7] George Danezis and Prateek Mittal. 2009. SybilInfer: Detecting Sybil Nodes using Social Networks.. In *NDSS*. San Diego, CA, 1–15.

[8] Facebook. 2019. *Community Standards*. https://www.facebook.com/communitystandards

[9] David Mandell Freeman. 2017. Can you spot the fakes?: On the limitations of user feedback in online social networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1093–1102.

[10] Hao Fu, Xing Xie, Yong Rui, Neil Zhenqiang Gong, Guangzhong Sun, and Enhong Chen. 2017. Robust spammer detection in microblogs: Leveraging user carefulness. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 6 (2017), 83.

[11] Peng Gao, Binghui Wang, Neil Zhenqiang Gong, Sanjeev R Kulkarni, Kurt Thomas, and Prateek Mittal. 2018. Sybilfuse: Combining local attributes with global structure to perform robust sybil detection. *arXiv preprint arXiv:1803.06772* (2018).

[12] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. 2012. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 61–70.

[13] Neil Zhenqiang Gong, Mario Frank, and Prateek Mittal. 2014. Sybilbelief: A semi-supervised learning approach for structure-based sybil detection. *IEEE Transactions on Information Forensics and Security* 9, 6 (2014), 976–987.

[14] Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. 2017. Random walk based fake account detection in online social networks. In *Dependable Systems and Networks (DSN), 2017 47th Annual IEEE/IFIP International Conference on*. IEEE, 273–284.

[15] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 29–42.

[16] Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. 2010. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 251–260.

[17] Abedelaziz Mohaisen, Nicholas Hopper, and Yongdae Kim. 2011. Keep your friends close: Incorporating trust into social network-based sybil defenses. In *2011 Proceedings IEEE INFOCOM*. IEEE, 1943–1951.

[18] SiHua Qi, Lulwah AlKulaib, and David A Broniatowski. 2018. Detecting and characterizing bot-like behavior on Twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 228–232.

[19] Devakunchari Ramalingam and Valliyammai Chinnaiah. 2018. Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering* 65 (2018), 165–177.

[20] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research* 11, Apr (2010), 1297–1322.

[21] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. The DARPA Twitter bot challenge. *Computer* 49, 6 (2016), 38–46.

[22] Binghui Wang, Neil Zhenqiang Gong, and Hao Fu. 2017. GANG: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs. In *Data Mining (ICDM), 2017 IEEE International Conference on*. IEEE, 465–474.

[23] Binghui Wang, Jinyuan Jia, Le Zhang, and Neil Zhenqiang Gong. 2018. Structure-based sybil detection in social networks via local rule-based propagation. *IEEE Transactions on Network Science and Engineering* (2018).

[24] Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. 2017. SybilSCAR: Sybil detection in online social networks via local rule based propagation. In *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*. IEEE, 1–9.

[25] Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. 2018. Sybilblind: Detecting fake users in online social networks without manual labels. In *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 228–249.

[26] Jilong Xue, Zhi Yang, Xiaoyong Yang, Xiao Wang, Lijiang Chen, and Yafei Dai. 2013. Votetrust: Leveraging friend invitation graph to defend against social network sybils. In *2013 Proceedings IEEE INFOCOM*. IEEE, 2400–2408.

[27] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 71–80.

[28] X Yang, Q Cao, and M Sirivianos. 2012. SybilRank: Aiding the detection of fake accounts in large scale social online services.

[29] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y 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] Haifeng Yu, Phillip B Gibbons, Michael Kaminsky, and Feng Xiao. 2008. Sybillimit: A near-optimal social network defense against sybil attacks. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 3–17.

[31] Haifeng Yu, Michael Kaminsky, Phillip B Gibbons, and Abraham Flaxman. 2006. Sybilguard: defending against sybil attacks via social networks. In *ACM SIGCOMM Computer Communication Review*, Vol. 36. ACM, 267–278.

[32] Jinxue Zhang, Rui Zhang, Jingchao Sun, Yanchao Zhang, and Chi Zhang. 2015. Truetop: A sybil-resilient system for user influence measurement on twitter. *IEEE/ACM Transactions on Networking* 24, 5 (2015), 2834–2846.