# Can You Spot the Fakes? On the Limitations of User Feedback in Online Social Networks

David Mandell Freeman
LinkedIn Corporation, USA
dfreeman@linkedin.com

## ABSTRACT

Online social networks (OSNs) are appealing platforms for spammers and fraudsters, who typically use fake or compromised accounts to connect with and defraud real users. To combat such abuse, OSNs allow users to report fraudulent profiles or activity. The OSN can then use reporting data to review and/or limit activity of reported accounts.

Previous authors have suggested that an OSN can augment its takedown algorithms by identifying a "trusted set" of users whose reports are weighted more heavily in the disposition of flagged accounts. Such identification would allow the OSN to improve both speed and accuracy of fake account detection and thus reduce the impact of spam on users.

In this work we provide the first public, data-driven assessment of whether the above assumption is true: are some users better at reporting than others? Specifically, is reporting skill both *measurable*, i.e., possible to distinguish from random guessing; and *repeatable*, i.e., persistent over repeated sampling?

Our main contributions are to develop a statistical framework that describes these properties and to apply this framework to data from LinkedIn, the professional social network. Our data includes member reports of fake profiles as well as the more voluminous, albeit weaker, signal of member responses to connection requests. We find that members demonstrating measurable, repeatable skill in identifying fake profiles do exist but are rare: at most 2.4% of those reporting fakes and at most 1.3% of those rejecting connection requests. We conclude that any reliable "trusted set" of members will be too small to have noticeable impact on spam metrics.

**Keywords:** Social networks; spam detection; online trust; fake accounts; reputation systems.

## 1. INTRODUCTION

Online social networks' ubiquity and popularity make them appealing platforms for spammers and fraudsters to execute their dirty deeds. Since most large OSNs (e.g. Facebook [11, §4] and LinkedIn [16, §8]) require user accounts to reflect a real identity, malicious actors use fake or compromised accounts to connect with and defraud real, unsuspecting users of the platform. Combatting fake accounts (also known as *sybils*) and preventing unauthorized access have thus received much attention from both researchers and practitioners, and there is a large body of work on fake account detection [5,6,9,19,24,26,28,29] and account takeover prevention [2–4, 13] in the literature.

To help combat this abuse, OSNs make available to users a variety of mechanisms for reporting fraudulent profiles or activity, such as "flag" or "report spam" interfaces. The OSN can then review and/or limit activity of accounts that have been reported, for example if a user receives too many reports in a given time period. In addition, there may be mechanisms for users to take positive action on real accounts (e.g., accept an invitation request), which can also serve as a form of reporting.

It seems reasonable to assume that some users are more adept at distinguishing fake profiles from real ones; indeed, several previous works on identifying spam or fake accounts have made use of this assumption [6, 7, 25, 30]. If this assumption is true, then the OSN can augment its takedown algorithms by identifying a "trusted set" of users whose feedback is weighted more heavily in the disposition of flagged accounts. Such identification would allow the OSN to improve both speed and accuracy of fake account detection and thus reduce the impact of spam on users.[1]

While the potential for leveraging high-quality reporters seems high, to date there has been no rigorous published study of reporting ability in real-life social networks. The goal of this paper is thus to test the following hypothesis: *There are some social network users who are good at identifying fake accounts.*

### 1.1 Our contribution

In this work we provide the first public, data-driven assessment of user feedback signals in the context of reporting fake accounts in online social networks. In particular:

- We provide a statistical framework for assessing user reporting skill. If flagging is a real skill, it must be *measurable*; that is, we should be able to determine which users are particularly skilled at flagging and quantify how skilled they are. Furthermore, flagging skill must

[1]To prevent manipulation of the flagging signal, any "trusted" label should not be exposed to the end user (either directly or indirectly).


*WWW 2017,* April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052706

be *repeatable*; that is, a user who demonstrates superior flagging ability in one data set should be able to demonstrate the same ability on a different sample or at a different point in time. If flagging is measurable but not repeatable then even once we identify good flaggers, the OSN cannot use their future flagging activity to help catch fake accounts.

- We apply our framework to data from LinkedIn, the professional social network. We consider three different signals: flagging of fake profiles, accepting connection requests, and rejecting connection requests. We find that members demonstrating measurable, repeatable skill in identifying fake profiles exist but are rare: at most 2.4% of members reporting fake accounts over a six-month period, and at most 1.3% of members rejecting connection requests over a one-month period. We also find that up to 3.8% of members accepting connection requests show skill in identifying *real* accounts.

We note that our analysis is concerned with aggregating reporting signals by *reporter*. LinkedIn and many other social networks also aggregate signals by *reportee*; indeed, a primary motivation for this work is to explore whether reporter-based signals might be effectively incorporated into existing reportee-based systems.

We note further that all signals in our data set were collected organically, without specifically instructing users to look out for fake accounts. This leaves open the question of whether targeted user prompting or some other offical spam-finding program would increase the prevalence of skilled reporters.

## 1.2 Related work

Zheleva et al. [30] describe a system for email spam filtering that uses the very kind of reporter reputation whose existence we are trying to establish, and Chen et al. [7] develop a similar system for fighting SMS spam. Both papers propose a framework in which reports of reliable users are weighted more highly in classifying spam, and provide a mechanism for evolving reporter reputation over time as new reports come in. However, the system of Zheleva et al. requires "an initial set of users who have proven to be reliable in the past," and the authors implicitly assume that such a set both (a) can be identified and (b) will continue to be reliable in the future. Our work calls into question this assumption, at least in the domain of social networks.

Wang et al. [25] describe a crowdsourcing study in which workers are shown accounts and asked to label them as real or fake. They find that in this artificial setting, "people can identify differences between Sybil and legitimate profiles, but most individual testers are not accurate enough to be reliable." They quantify this reliability only in terms of accuracy and do not attempt to test repeatability of the workers' labeling.

Moore and Clayton [20] and Chia and Knapskog [8] have studied the "wisdom of crowds" in reporting phishing and web vulnerabilities, respectively. Both studies find a power-law distribution in participation rates, and the former finds that more frequent reporters achieve higher accuracy and recommend that "the views of inexperienced users should perhaps be assigned less weight when compared to highly experienced users." However, Moore and Clayton do not suggest how such a weight should be determined algorithmically, nor do they consider repeatability.

Cao et al. [6] use negative feedback such as invitation rejection and spam reporting to downweight graph edges in the SybilRank algorithm [5], but they do not consider quality or trustworthiness of the reporters.

More generally, our work is related to the wide body of research on peer-to-peer *reputation systems* [17,23], as practiced for example in online auction houses such as eBay. In such systems parties leave publicly visible feedback on each other, which is aggregated per recipient to produce a reputation score. Guha et al. [14] describe how this reputation can propagate through a network and be used to predict trust between two nodes. Our situation is slightly different in that we are aggregating on the user leaving feedback and the feedback is not public.

# 2. EVALUATING REPORTING ABILITY

To assess whether the ability to identify fake accounts is a real skill, we quantify the ability along two axes: whether it is *measurable*, and whether it is *repeatable*. Certainly if we cannot determine which members are better or worse at identifying fakes, then it will be impossible to leverage this ability in those members for whom it exists. Furthermore, even if we can identify some members as being particularly good at identifying fakes, this ability is of no use if it is transient.

We begin by setting some notation to model social network interaction and reporting events. Let $\mathcal{U}$ be the set of users in a social network. We let $u \in \mathcal{U}$ be a (real) user of the social network and let $\mathbf{x}(u) = \{x_1, \ldots, x_n\} \in \mathcal{U}^n$ be a set of users to which user $u$ is exposed during time period $[t_1, t_2]$. For example, these could be users who invite $u$ to connect, whose profile $u$ views, or who appear in $u$'s news feed. Each of the users $x_i$ has a *truth label* $y_i \in \{0, 1\}$ indicating whether this user is real (1) or fake (0).

At any given time $t' \in [t_1, t_2]$, user $u$ may emit a *reporting action* $\mathcal{R}$ for any of the users $x_i \in \mathbf{x}$. The action $\mathcal{R}$ may be *positive*, designed to apply to real accounts (e.g. accept connection request), or *negative*, designed to apply to fake accounts (e.g. flag as spam). We define $\sigma(\mathcal{R})$ to be 1 for positive actions and 0 for negative actions. For each $i$ we define

$$r_i = \begin{cases} 1 & \text{if } u \text{ reports } x_i, \\ 0 & \text{if } u \text{ does not report } x_i. \end{cases}$$

We assume that the reporting action $\mathcal{R}$ is fixed for any given data set (i.e., we do not analyze data sets with mixed actions) and thus $\sigma(\mathcal{R})$ is well-defined for the entire data set.

Given this notation, we now have four possible outcomes for each user $x_i \in \mathbf{x}$: the user can be real or fake, and the user can be reported or ignored. We denote the quantities of each outcome by $a_u, b_u, c_u, d_u$, defined as follows:

| | Reported | Ignored |
|---|---|---|
| Real | $a_u = \sum_i r_i y_i$ | $b_u = \sum_i (1 - r_i) y_i$ |
| Fake | $c_u = \sum_i r_i (1 - y_i)$ | $d_u = \sum_i (1 - r_i)(1 - y_i)$ |

(1)

We now use these quantities to develop scores that measure reporting ability.

## 2.1 Measurability

We start by defining reporter precision; i.e., the probability that a given report will be correct. Concretely, if $r'$ is a reporting event from $u$ on a new member $x'$ with truth label $y'$, then we define the *precision score* to be

$$\tilde{P}(u) = \Pr[y' = \sigma(\mathcal{R}) \mid r' = 1]. \quad (2)$$

We can estimate this probability as follows:

$$P(u) = \frac{a_u \sigma(\mathcal{R}) + c_u(1 - \sigma(\mathcal{R}))}{a_u + c_u}. \quad (3)$$

Our definition guarantees that a score of 1 corresponds to the best reporter regardless of whether the reporting action is designed to identify real or fake accounts.

We observe that the precision score $P$ does not distinguish a reporter who flagged once and was correct from one who flagged 50 times and was always correct. To make this distinction we *smooth* the precision score by adding $\alpha$ correct and $\alpha$ incorrect flags to the user's data, and denote the resulting function by $P_s(u)$. Now for $\alpha = 1$, the user flagging only once correctly has $P_s(u) = 0.67$ while the user flagging 50 times correctly has $P_s(u) = 0.98$. For each data set we analyze, we determine the optimal value of $\alpha$ by evaluating area under the precision-recall curve [10] on a test set.

**Informedness**. The precision score $P$ has an additional drawback in terms of measuring reporting ability: it is insensitive to the relative proportion of real and fake accounts that user $u$ has interacted with, and in particular does not take into account the users that $u$ ignores. To see why this is a problem, consider two users $u$ and $u'$ in Table 1. User $u$ has reported 50% of both the real and the fake accounts he saw, while user $u'$ has reported 50% of the fakes and only 5% of the real accounts she saw. It is clear that in this case $u'$ is the better flagger, but the score $P(u) = P(u') = 0.5$ does not help us draw this conclusion.

| $u$ | Report | Ignore |
|---|---|---|
| Real | 5 | 5 |
| Fake | 5 | 5 |

| $u'$ | Report | Ignore |
|---|---|---|
| Real | 5 | 95 |
| Fake | 5 | 5 |

**Table 1:** If $\mathcal{R}$ is negative (e.g. report spam), $u'$ is more skilled at identifying fakes than $u$.

A more robust metric will take into account a user's baseline propensity for reporting; this is especially relevant for a signal like invitation accept, where some users may accept all or none of their incoming invitations. To measure this we use *informedness*, also known as *Youden's J-statistic* [27], which "quantifies how informed a predictor is for the specified condition, and specifies the probability that a prediction is informed in relation to the condition" [21]. Informedness is defined to be true positive rate minus false positive rate, which in our case is

$$\tilde{I}(u) = \Pr[r' = 1 \mid y' = \sigma(\mathcal{R})] - \Pr[r' = 1 \mid y' = 1 - \sigma(\mathcal{R})]. \quad (4)$$

(Note that $\tilde{I}(u)$ can take values in $[-1, 1]$; negative quanities indicate quality of the signal as a predictor of the inverse classes.) We estimate $\tilde{I}(u)$ by computing

$$I(u) = \left( \frac{a_u}{a_u + b_u} - \frac{c_u}{c_u + d_u} \right) (2\sigma(\mathcal{R}) - 1). \quad (5)$$

If either of the denominators $a_u + b_u$, $c_u + d_u$ is zero, then $u$ has interacted with either only real or only fake accounts and $I(u)$ is undefined. Now our two users $u, u'$ from Table 1 have scores $I(u) = 0$ and $I(u') = 0.45$, respectively, reflecting our intuition that $u'$ is more skilled at reporting.

**Hypothesis Testing**. While the informedness $I(u)$ takes into account all the information we have about user $u$, it does not do a good job of distinguishing skilled users from lucky ones. Consider for example the two users $v$ and $v'$ of Table 2, again with a negative reporting signal $\mathcal{R}$. We have $I(v) = I(v') = 0.5$, but it may be the case that $v$ reports half of *all* users, whether real or fake; for $v'$ the difference between actions on real and fake users is unambiguous.

| $v$ | Report | Ignore |
|---|---|---|
| Real | 2 | 2 |
| Fake | 1 | 0 |

| $v'$ | Report | Ignore |
|---|---|---|
| Real | 20 | 20 |
| Fake | 10 | 0 |

**Table 2:** If $\mathcal{R}$ is negative (e.g. report spam), $v'$ shows skill at identifying fakes, while $v$ may have gotten lucky.

To distinguish these two cases we undertake a statistical hypothesis test, with the null hypothesis being that the user is equally likely to report real and fake accounts, i.e.:

$$H_0 : \frac{\Pr[r' = 1 \mid y' = \sigma(\mathcal{R})]}{\Pr[r' = 1 \mid y' = 1 - \sigma(\mathcal{R})]} = 1. \quad (6)$$

For a good flagger the odds ratio in (6) will be greater than 1, so we wish to compute a one-sided $p$-value that gives the probability of obtaining data at least as extreme as the observed data, conditioned on $H_0$. Our test of choice is *Fisher's exact test* on $2 \times 2$ contingency tables [12], as implemented in the R statistical computing program [22]. If we let $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, Fisher's test computes a $p$-value $p_F(M)$ which is defined to be the proportion of $2 \times 2$ matrices that have the same row and column sums as $M$ and are "less extreme" than $M$ as defined by the Wald statistic [18]. The advantage of the Fisher test is that it is accurate even with small sample sizes, as opposed to, e.g., a $\chi^2$-test.[2] We thus define the *Fisher score* to be

$$F(u) = 1 - p_F \begin{pmatrix} a_u & b_u \\ c_u & d_u \end{pmatrix}, \quad (7)$$

where we subtract from 1 so that good scores are close to 1.

Using this metric, the two users of Table 2 have $F(v) = 0.4$ and $F(v') = 0.997$, reflecting our intuition that $v'$ truly does have different flagging behaviors on real vs. fake accounts.

One drawback to the Fisher score is that it rewards a statistically significant difference between reports on real and fake accounts even if precision and recall are low. For example, consider the following user $w$:

| $w$ | Report | Ignore |
|---|---|---|
| Real | 20 | 80 |
| Fake | 5 | 5 |

[2]Mehta and Senchaudhuri [18] suggest that Barnard's test [1] may be more appropriate for this situation, but we could not find an implementation that would compute the test statistic on thousands of samples in a reasonable amount of time.

This user has $P(w) = 0.2$ and $I(w) = 0.3$ but $F(w) = 0.95$ — she acts differently on real and fake accounts but is not particularly good at identifying either.

## 2.2 Repeatability

In the previous section we developed several metrics to measure reporting ability of social network users. However, if true ability exists then it must persist upon repeated sampling — knowing that a user has been skilled at reporting in the past is of no use if that user will not continue to be skilled in the future. We now develop metrics to measure this property.

We start with some notation. Let $\mathcal{U}$ be a set of users as above and let $\mathcal{D}$ be the set of all possible observations about a user $u \in \mathcal{U}$ (e.g., the profiles viewed and profiles flagged by the user $u$). Let $m\colon \mathcal{D} \to \mathbb{R}$ be a scoring function on the observed data for a user $u$. Suppose that for each user $u_1, \ldots, u_k$, we have two sets of observations $d_1, \ldots, d_k$ and $d'_1, \ldots, d'_k$. We wish to determine how the two sets of scores $\mathbf{s} = \{m(d_1), \ldots, m(d_k)\}$ and $\mathbf{s}' = \{m(d'_1), \ldots, m(d'_k)\}$ are related, and in particular to assess how much information one set can give us about the other.

**Correlation**. The most straightforward measure of this relation is the *Pearson correlation coefficient* [15], which measures linear correlation between the two vectors $\mathbf{s}$ and $\mathbf{s}'$. If the correlation is close to zero then we can conclude that flagging ability is not repeatable; however a score close to 1 does not necessarily indicate the opposite. To see this, consider ten users $u_1, \ldots, u_{10}$, where $m(d_i) = i/10$ and $m(d'_i) = i/20$. These scores are perfectly linearly correlated but clearly the second set shows much poorer ability than the first set, and we would not want to claim that these users demonstrated repeatability.

We also can compute the *Spearman correlation coefficient*, which is the Pearson coefficient of the two vectors of ranks computed from $\mathbf{s}$ and $\mathbf{s}'$. The Spearman coefficient is more robust to nonlinear effects [15], but the example above still gets a perfect score.

**Persistence**. Correlation gives a single measure of whether scores "match up" between two different samples. However, to identify which users are skilled at any given score threshold we need a *continuous* measure. Specifically, we want to determine the following: suppose that a score of $\beta$ indicates a "good" flagger. If user $u$ has a good score on one set of observations, what is the probability that $u$ also has a good score on a second set of observations? We estimate this probability by defining the *persistence at score* $\beta$ to be

$$\pi(\beta) = \frac{|\{u_i : m(d_i) \geq \beta \wedge m(d'_i) \geq \beta\}|}{|\{u_i : m(d_i) \geq \beta \vee m(d'_i) \geq \beta\}|}. \quad (8)$$

The symmetry of this definition makes it suitable for situations such as A/B testing where observations are placed randomly into one of two buckets.

The persistence score clearly shows that our sample of ten users described above does not demonstrate good repeatability despite the correlation. If we assume 0.5 is a "good" score, the sample has $\pi(0.5) = 0.17$ since out of the six users with good scores in either set, only $u_{10}$ has a good score in both sets.

## 2.3 Evaluating Scores

The measurements discussed in Section 2.1 output real-numbered scores between 0 and 1, which we can then use to define an ordering on reporters that reflects their relative flagging ability. If we want to use these scores to label "skilled reporters" then we must choose a score threshold. This choice is necessarily a business decision to be made by weighting the relative costs of false positives and false negatives and picking an operating point. Estimating such costs is outside the scope of this work; therefore when possible we present complete curves so the reader can view the data across the full range of possibilities.

However, we also wish to offer a concrete assessment of our findings, rather than only presenting curves, which necessitates picking a specific score cutoff. We choose our cutoffs as follows: let $\rho$ be the average flagging precision for the entire data set. For each of our metrics we divide the score range into buckets of width 0.05 and choose the threshold $t$ to be the smallest bucket lower bound that satisfies the following:

a) The cumulative precision of all buckets with scores $\geq t$ is at least $(1 + \rho)/2$, and
b) The bucket with scores in $[t, t + .05)$ has precision at least $(1 + \rho)/2$.

Condition (a) requires us to choose a threshold that decreases the error rate (i.e., $1 -$ precision) by at least half over an average flagger. We include condition (b) to maintain high quality across all scores within the "skilled" range. For example, suppose the following hold over our data set: $\rho = 0.6$, reporters with scores in $[0.75, 1]$ have precision 0.9, and reporters with scores in $[0.7, 0.75)$ have precision 0.5 but are not numerous enough to bring the cumulative average for the bucket $[0.7, 1]$ below $(1+\rho)/2 = 0.8$. In this case, we would not want to label the $[0.7, 0.75)$ reporters as skilled, so we choose $t \geq 0.75$.

**Combining scores**. Given the limitations of each of our metrics, we do not feel confident labeling a user as "skilled" based on only a single metric. We thus define a *skilled reporter* $u$ to be one that has at least two of the three measures $P_s(u)$, $I(u)$, $F(u)$ greater than the appropriate threshold $t_P$, $t_I$, $t_F$ (computed as described above) on two different data sets. In other words, we define skill to mean that two of the three following characterstics can be repeated over time:

- The user flags with sufficient precision;
- The user flags real and fake accounts in different proportion;
- The difference between flagging behavior on real and fake accounts is statistically significant.

# 3. USER FLAGGING

We now apply the framework of Section 2 to real data, beginning with user flagging. Most social networks contain an option to flag profiles or content for being inappropriate and/or violating the terms of service. This flagging data then feeds into the back end, where it can be used to take down content or remove offending members from the site. It can also provide a measure of recall for classifiers or be used to label training data.

In this section we investigate whether some members show a measurable, repeatable skill in flagging fake profiles. If such a skill existed, the social network could, for example,

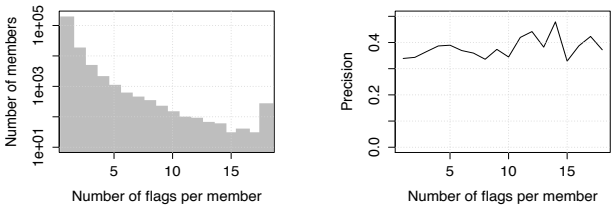


**Figure 1:** Left: histogram of flags per member, with logarithmic $y$-axis. The rightmost bar consists of all buckets with fewer than 20 members (i.e. members with at least 18 flags). Right: flagger precision for each bucket in the histogram.

identify skilled flaggers and use their flagging signals to take down fake accounts automatically.

In the following discussion we will denote by the *reporter* the member initiating the flag action, and by the *reportee* the member being flagged.

## 3.1 Data Collection

We obtained a subset of the flagging data from LinkedIn for the period from February 15 to August 15, 2016. In order to focus on fake profiles (as opposed to non-fake, objectionable content), we consider only user flags in which one of the following reasons was selected: "Fake Identity," "Impersonation," "Misrepresentation," or "Account Hacked."[3] Since the flags we are looking at are at the profile level, we don't want to give one member credit for flagging the same profile multiple times; in the extreme, this would allow a bad actor to take down any profile simply by flagging it a lot. We thus deduplicated the data on (reporter, reportee) pair, leaving us with a total of 293,114 flags from 227,550 unique reporters, flagging 237,638 unique reportees.

For ground truth, we obtained labels from LinkedIn's Anti-Abuse Team indicating whether each reportee was real or fake. The number of flags of fake members in this data set was 101,632, or 35% of the total, comprising 65,992 unique fake accounts.[4] A histogram of the number of members with each flag count and the precision for each bucket appears in Figure 1 (buckets with fewer than 20 reporters (i.e. $\geq$18 flags) are grouped together in the last bucket). Interestingly, when taken in aggregate, precision does not appear to increase with the number of flags; in fact, the precision for the 1-flag bucket is 34% and that for the $\geq$18-flag bucket is 37%. So certainly there is opportunity for some reporters to be particularly good at identifying fake accounts.

In order to compute flagging recall and false positive rate, we need some estimate of a member's overall exposure to real and fake accounts. While there are many ways two members can interact with each other (e.g. messaging, articles, recommendations, search results), we use profile views as a simple proxy. We obtained logs of member profile views over the same six-month period for each member that flagged at least once, and labeled the viewee as real or fake.

## 3.2 Measurability

We consider the three measures of reporter ability defined in Section 2.1: precision, informedness, and Fisher score. The cumulative distribution functions of the three scores are shown in Figure 2.

We began by computing smoothed precision scores for smoothing parameters $\alpha \in \{0, 2^{-5}, 2^{-4}, \ldots, 2^5\}$. To choose the best smoothing parameter, we held out one third of the data as a test set and computed the area under the precision-recall curve for each choice of $\alpha$. We found that $\alpha = 0.5$ gave greatest area under the curve, so we used this value to compute $P_s$.

We next note that 57% of all reporters flag only once and are incorrect, while 29% flag once and are correct. These correspond to the large vertical lines at 0.25 and 0.75 in the smoothed precision plot. We find that $P_s$ achieves our target of 67% precision (halfway between 1 and the global average of 35%) on each score bucket above 0.65 (see Figure 4); we thus choose $t_P = 0.65$ as the threshold for a skilled reporter by this metric. We find that 33% of all reporters are skilled by this definition. However this statistic includes those that reported once correctly, so only at most 3% of reporters (7,763) could possibly demonstrate repeatability.

Before computing the informedness score, we observe that some 73,438 members (32%) saw no fake profiles, and thus have undefined informedness score. For purposes of plotting the CDF of Figure 2 we assign these members a score of zero. For informedness we use the procedure defined in Section 2.3 to obtain threshold of $t_I = 0.05$ for skilled reporters. Indeed, 0.05 is the leftmost point in the Informedness plot of Figure 4 for which both curves are above 0.67. We find that 34% of all reporters are skilled by this definition; however, the number of such reporters with more than one flag is only 6% (12,914), which again bodes ill for repeatability.

Finally we turn to the Fisher score. We obatin a threshold $t_F = 0.2$, with 36% of reporters achieving this score or higher. Taking out those that flagged only once leaves 6% (14,132 members) with more than one flag.

## 3.3 Repeatability

To measure repeatability we split our data set approximately in half and compute the three measures on each half. To avoid time-based effects (e.g., LinkedIn might have been better at catching fake accounts in June than in May), we split according to whether the timestamp (in milliseconds)

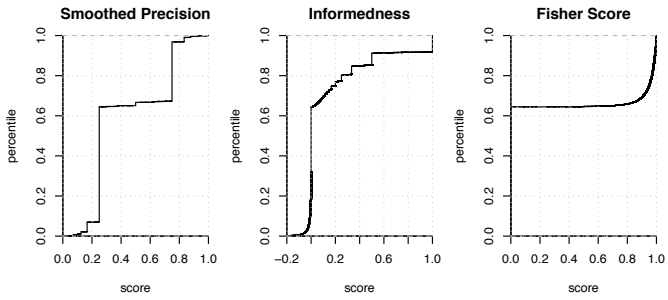


**Figure 2:** Cumulative distribution functions for the three measures of reporter ability on user flagging data.

[3]We include the "Account Hacked" option because many members, upon encountering unwanted activity from a reasonable-looking account, assume the account was hacked rather than fake.

[4]We note that these labels are our best knowledge rather than incontrovertible truth, especially with regard to false negagtives (i.e., fake accounts not yet labeled as such).

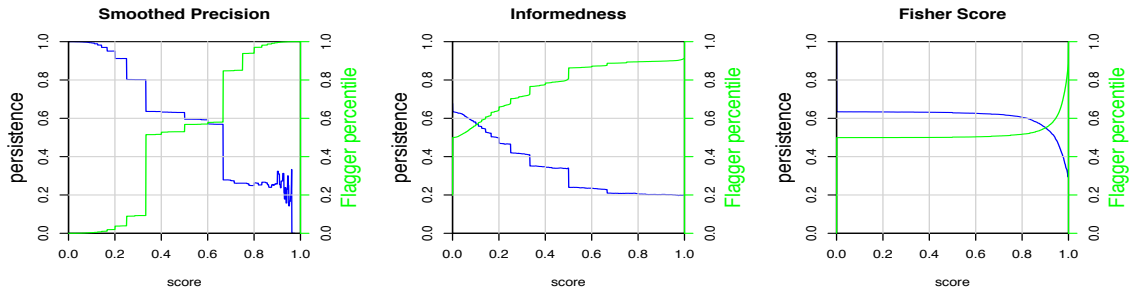


**Figure 3:** Blue curves: persistence of the three measures of reporter ability on the user flagging data. Green curves: cumulative distribution functions for each persistence measure.

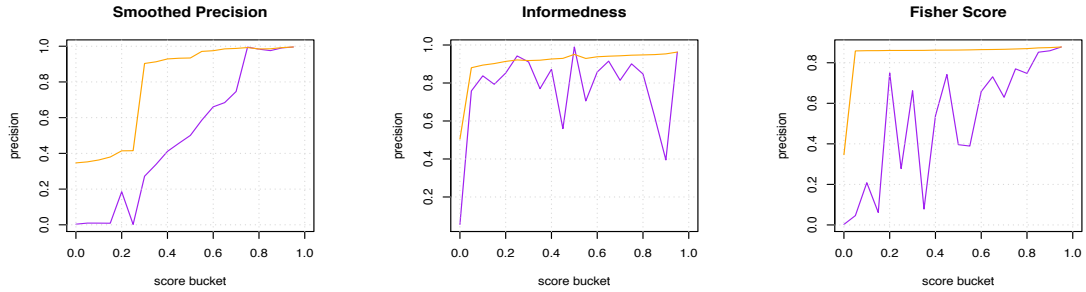


**Figure 4:** Purple curves: precision of reporters in each score bucket $[x, x + 0.05)$. Orange curves: cumulative precision of all reporters with score at least $x$.

of each observed event is 0 or 1 mod 2. This left 222,366 members (98%) with events in both data sets; however, only 18,138 (8%) have at least one reporting event in each half of the data. This observation already bodes ill for repeatability, as it assigns no predictive value to 92% of our set of members. The rest of this section treats only the remaining 8%.

We start by computing the correlation between scores on the two halves of the data set; results are as follows:

| Score | Pearson | Spearman |
|---|---|---|
| Smoothed Precision | 0.69 | 0.66 |
| Informedness | 0.52 | 0.49 |
| Fisher Score | 0.62 | 0.63 |

Figure 3 shows persistence of each of our three metrics as a function of the score. (As above, undefined informedness is assigned a score of zero.) If we let the thresholds $t_P = 0.65$, $t_I = 0.05$ and $t_F = 0.2$ be the cutoffs for "good" precision, informedness, and Fisher scores, respectively, we obtain the following:

| | Cutoff | Persist. | Pct. | Members |
|---|---|---|---|---|
| Smoothed Precision | 0.65 | 0.57 | 0.42 | 4333 |
| Informedness | 0.05 | 0.62 | 0.48 | 5344 |
| Fisher Score | 0.20 | 0.63 | 0.50 | 5748 |

We see that with our chosen cutoffs, all three scores are "sticky": if a member shows skill on one half of the data, the odds are around 60% that the member will show skill on the other half. However, even the 50% of members who show persistence in Fisher score corresponds to only 2.5% of our initial set of reporters — there are simply not enough members that report often enough to demonstrate persistence.

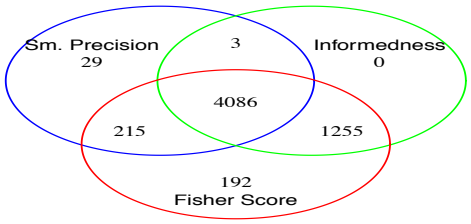


**Figure 5:** Venn diagram of persistent flagging scores. Labels indicate the number of members with scores above the appropriate threshold(s) on both halves of the data.

To quantify who is a skilled reporter, we look for members exhibiting two of the three scores above the appropriate threshold $t_P, t_I, t_F$ on both halves of the data. Figure 5 shows the Venn diagram of overlaps. We see that skill according to all three metrics overlaps significantly, and a total of 5,559 members satisfy our criteria for being skilled. This number represents 31% of all those who flagged once in each data set, and 2.4% of our initial set of flaggers.

Our analysis thus shows a small set of members who demonstrate skill at reporting spam. These 5,559 members flagged a total of 25,636 times at 82% cumulative precision, covering 17,919 unique fake accounts over the six months of data collection. This precision is unlikely to be high enough for a single flag to be cause for removing an account from the site. However, if we look at only the 4,304 flaggers in the high-precision band (blue circle in Figure 5), these members flagged 13,940 accounts at 97% precision, which is high enough that one could reasonably take action on a single flag.

# 4. INVITATIONS

One of the core components of a social network is the ability for members to "connect" with other members. These connections usually reflect relationships that exist in the offline world (e.g. friend or coworker). Members that are connected have additional privileges that do not apply to non-connected members. These privileges may include seeing each other's detailed profile information and/or contact details, having each other's updates show up in a "news feed," being recommended for connections, jobs, or products based on each other's data, and sending private messages to each other.

The additional privileges afforded to connections are enticing to spammers and scammers: most platforms do not allow random members to send messages to each other, so to get information to their targets scammers must connect. To do this they send *invitations* to connect, which may appear to the recipient as email, push notifications, or in-app badges.

Upon receiving an invitation, the member can do any of three actions:

- **accept**: A member clicks "Connect" or '✓' on an inbound invitation (request to connect in the network).
- **reject**: A member clicks "Decline" or an '×' on an inbound invitation.
- **ignore**: The member does not respond to the invitation, either due to not receiving the notification or not bothering to respond.

A member who is good at identifying spammers should accept invitations from real accounts and reject (or at least ignore) invitations from fake accounts. Reject signals from such members could be used to take automatic action on fakes, while accept signals could be used as evidence that the sender is real and can thus be allowed more privileges (e.g., higher activity limits).

## 4.1 Data collection

Our invitations data set consists of a subset of invitations sent by LinkedIn members during June 2016. We took all invitations sent during that month and labeled the sender as real or fake using labels provided by LinkedIn's Anti-Abuse Team. We call invitations sent by fake accounts "spam" and invitations sent by real accounts "non-spam."

We then labeled each invitation as accepted, rejected, or ignored, based on the recipient's response. However, the time dimension adds some complications to the task of providing meaningful labels. In particular:

- We know the disposition of the invitation at the present time, but this means that an invitation sent on June 1 has had 29 more days to receive a response than one sent on June 30. We solve this issue by treating all invitations receiving an accept or reject more than $h$ hours after sending as ignored; this places all invitations on equal footing.
- LinkedIn's fake account defenses were constantly running during this period. After a fake account is found, its invitations can no longer be responded to; if this happens within $h$ hours of sending and before the recipient could respond, then this invitation's lifetime is not comparable to others recieved by the same member. We thus remove from our data set all invitations where the sender was restricted less than $h$ hours after sending the invitation and the recipient did not respond before the sender was restricted.[5]

We collected data as described above for $h \in \{4, 24, 168\}$. We required that each recipient in our data set receive at least two spam and three non-spam invitations during the sample period. For each value of $h$ we then sampled 500,000 members. We present our full analysis as above for $h = 24$, i.e., invitations with actions in the first 24 hours after sending, and then discuss in Section 4.4 how the other time periods compare.

## 4.2 Measurability

If a member truly has the ability to identify fake accounts from invitations, the accept rate will be lower for invitations from fake accounts than from real accounts and the reject rate will be higher. In the terminology of Section 2, invitation accept is a positive reporting action, and invitation reject is a negative reporting action.

We began by computing smoothed precision scores for smoothing parameters $\alpha \in \{0, 2^{-5}, 2^{-4}, \ldots, 2^{5}\}$. To choose the best smoothing parameter, we held out one half of the data as a test set and computed the area under the precision-recall curve for each choice of $\alpha$. We found that for invitation accept $\alpha = 1$ gave greatest area under the curve, while for invitation reject $\alpha = 4$ was optimal. We will use these values to compute $P_s$ in the sequel.

We apply our three measures of flagging ability to both the invitation accept and reject data. Plots of the cumulative distribution function are shown in Figure 6. We notice right away that some members act identically on all incoming invitations regardless of whether they are from real or fake accounts; these members correspond to the vertical jumps at zero in the plots of $I(u)$.

The precision of invitation acceptance in our entire data set was 92%, while that for invitation rejection was 22%. We used the method of Section 2.3 to determine thresholds for skilled flaggers by each measure:

No bucket of width 0.05 achieved the required precision of 61% for the Fisher score on rejections, so in this case we applied the procedure of Section 2.3 with buckets $[10^{1-t}, 10^{-t})$

[5]This choice does bias our dataset in the sense that "hard-to-catch" fakes will be overrepresented, but we feel that assuming "easy-to-catch" fakes will in fact be caught is reasonable for a study of flagging in real life.

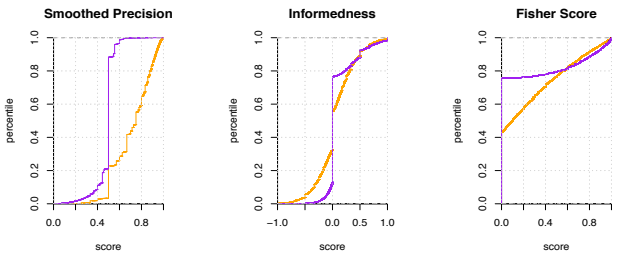


**Figure 6:** Cumulative distribution functions for the three measures of reporter ability on the invitation accept signal (orange) and invitation reject signal (purple).

| | Accept | | Reject | |
|---|---|---|---|---|
| | Cutoff | Pct. | Cutoff | Pct. |
| Smoothed Precision | 0.95 | 0.05 | 0.55 | 0.10 |
| Informedness | 0.20 | 0.27 | 0.90 | 0.02 |
| Fisher | 0.65 | 0.15 | 0.99999 | 0.0006 |

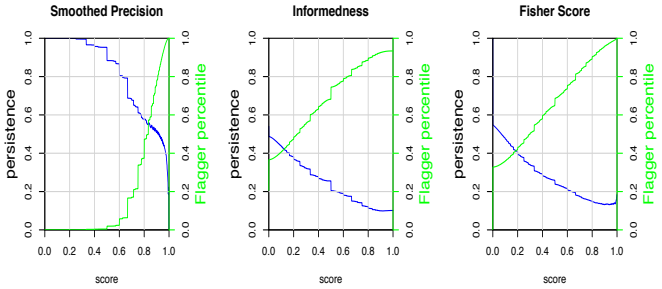


**Figure 7:** Blue curves: persistence of the three measures of reporter ability on the **invitation accept data**. Green curves: cumulative distribution functions for each persistence measure.

and found that the smallest $t$ satisfying the two conditions is $t = 5$, giving us a threshold of 0.99999 for the Fisher score.

We see from the data that in general more users are skilled at accepting invitations than rejecting them; however this effect could simply be due to the prevalence of non-spam in the data set.

## 4.3 Repeatability

To test repeatability we look at a slightly different data set: we split our June 2016 data in half according to whether the invitation timestamp is 0 or 1 mod 2, and we sample 500,000 members who (a) received invitations in both halves and (b) received at least one spam and two non-spam invitations in each half. (We note that of all members receiving spam invitations in the month, condition (a) excludes more than 75%.) In this data set, 78% accepted at least one invitation in each half and 25% rejected at least one invitation in each half. As before, we exclude members not reporting in both halves from our subsequent analysis.

Correlation of scores between the two halves is high for the precision metric, but for the other two metrics is noticeably less than for the user flagging data:

| | Accept | | Reject | |
|---|---|---|---|---|
| Score | Pearson | Spearman | Pearson | Spearman |
| Sm. Prec. | 0.59 | 0.66 | 0.75 | 0.65 |
| Informed | 0.23 | 0.24 | 0.29 | 0.28 |
| Fisher Score | 0.31 | 0.33 | 0.30 | 0.30 |

Figures 7 and 8 show persistence of each of our three metrics as a function of the score. (As before, undefined informedness is assigned a score of zero.) We observe that persistence is very low for all measures except precision on acceptances; very few members can demonstrate repeatability on either the invitation accept or the invitation reject signal.

Using the thresholds for skill calculated in Section 4.2, we obtain the following:

The Venn diagrams of the overlaps of reporters skilled on the two halves (Figure 9) show 19,073 members who are

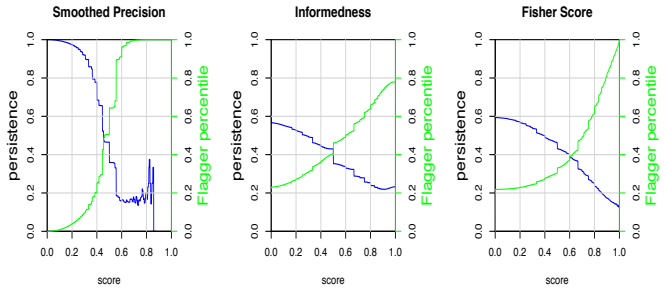


**Figure 8:** Blue curves: persistence of the three measures of reporter ability on the **invitation reject data**. Green curves: cumulative distribution functions for each persistence measure.

| **Accept** | Cutoff | Persist. | Pct. | Members |
|---|---|---|---|---|
| Smoothed Prec. | 0.95 | 0.46 | 0.08 | 14095 |
| Informedness | 0.20 | 0.38 | 0.53 | 78290 |
| Fisher Score | 0.65 | 0.20 | 0.22 | 17011 |
| **Reject** | Cutoff | Persist. | Pct. | Members |
| Smoothed Prec. | 0.55 | 0.33 | 0.32 | 12960 |
| Informedness | 0.90 | 0.22 | 0.27 | 7171 |
| Fisher Score | 0.99999 | 0.14 | 0.01 | 124 |

skilled at accepting invitations, and 6,320 skilled at rejection. These numbers represent 3.8% and 1.3% of all members in our data set, respectively. However, since we required members in our data set to receive at least two spam invitations in the month, a condition which excludes more than 75% of all members receiving spam, the proportion of all members seeing spam whom we can identify as skilled is no more than 1.0% for acceptance and 0.3% for rejection.

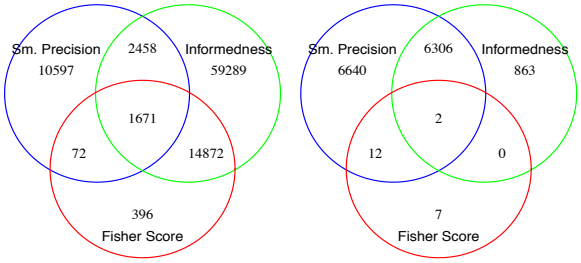


**Figure 9:** Venn diagrams of persistent abilities on invitation accept (left) and reject (right). Labels indicate the number of users with scores above the appropriate threshold(s) on both halves of the data.

Our skilled accepting members have, cumulatively, 98% precision on real members, while the skilled rejecting members have 99% precision on fake accounts. We conclude that for skilled flaggers, the invitation accept signal can be strong enough to definitively place a sender into the "real" category, while the invitation reject signal is strong enough to label the rejected member as a spammer. We note that since the Fisher score cutoff for rejection is nearly impossible to achieve, most members skilled at rejection have high precision and informedness, which translates in practice to flagging few times and almost always correctly.

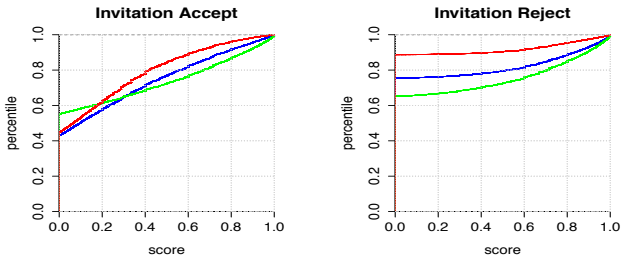


**Figure 10:** Fisher scores for invitation response signal based on different response times. Blue curves plot responses within 24 hours. Green curves plot responses within 4 hours. Red curves plot responses within 7 days.

## 4.4 Some Experiments

**Multiple invitation windows**. Recall that our data to this point is only counting invitation responses within 24 hours of the send time; acceptances and rejections after this time are counted as ignores. We use a time bound because utility of the reputation signal decreases over time: most fake accounts sending invitation spam are caught within 24 hours, and we wish to use the reputation signal to increase the speed of takedown.

We now consider what happens as we vary the time bound $h$. Specifically, we consider the same data with three different time bounds: 4 hours, 24 hours, and 7 days. Plots of the Fisher score appear in Figure 10. We see that as expected, allowing the signal to "bake in" for a longer period gives recipients more opportunity to distinguish themselves as skilled reporters.

**Simulating response to spam invitations**. Our results show that *some* members are good at detecting spam — but very few. One natural question is whether this result is an artifact of multiple hypothesis testing; e.g., if you look for an effect with 95% significance you'll find it 5% of the time in random data. To answer this question, we run the following experiment: we simulate each member's response to invitations from fake accounts using that member's responses to invitations from real accounts as a prior.

Specifically, we wish to replace $c_u$ and $d_u$ in each member's matrix (1) with $c_u + d_u$ samples from a binomial distribution with reporting probability $\frac{a_u}{a_u+b_u}$. However, we cannot produce such a sample in the case where $a_u$ or $b_u = 0$, so we smooth by adding $p$ events to $a_u$ and $(1-p)$ events to $b_u$, where $p$ is the global prior probability of a reporting event (i.e. accept or reject) on a real invitation. After generating the samples in this manner, we then ran the same Fisher test. The results (green curves in Figure 11) show a slight distinction between real and simulated data in terms of the score distibution, but persistence drops to zero for random data while it bottoms out around 15% for real data. This simulation thus finds almost no members skilled at flagging, suggesting that the small number of skilled members we found in the real data is not due to random noise.

# 5. CONCLUSIONS AND OPEN QUESTIONS

Recall from the introduction that we set out to test the following hypothesis: *There are some social network users who are good at identifying fake accounts.*

Through our analysis we have found that such a skill does exist, but it applies to only a small minority of all users: at most 2.4% of members who reported a fake profile through the flagging interface over a six-month period and at most 1.3% of members who rejected multiple invitation requests in a one-month period. We also found at most 3.8% of members accepting invitation requests to be skilled at identifying *real* accounts. On the basis of our analysis, we cannot recommend that online social networks build a "reporter-based reputation system" as described in [30], since few reporters are both good at flagging and can repeat their performance upon multiple sampling.

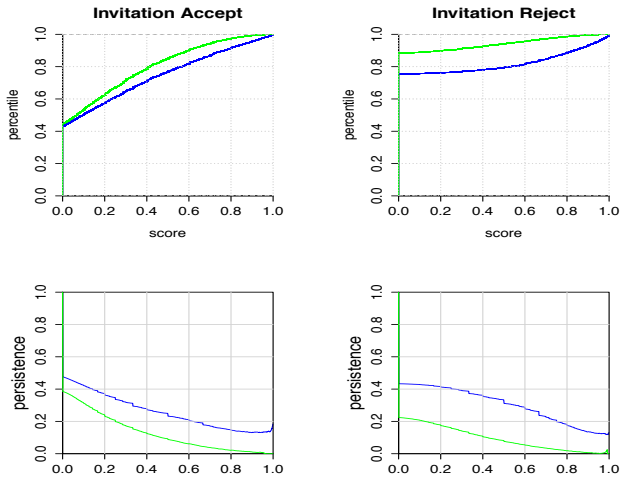


**Figure 11:** Fisher score persistence for real and simulated responses to spam invitations. Blue curves plot real data. Green curves plot simulated data sampled from a binomial distribution with prior equal to the smoothed mean on responses to real invitations.

One important area for future study is to explore how cues in the user interface could affect people's ability to identify fake accounts. Our data was collected "organically," i.e., without any particular prompting or instructions. It is possible that surfacing additional feedback to reporters and/or providing incentives for correctly reporting fakes could encourage users to act more often and more accurately and thus boost the impact of a reporter reputation system.

A second avenue for further work is to explore characteristics of good reporters. The observation by Wang et al. [25] that some geographic subgroups of people are more accurate than others when asked to identify fakes suggests that we could understand and exploit the differences that make these reporters more accurate.

Finally, we note that our study is based on data from a single large social network, and we encourage the research community to reproduce the study using data from other social networks to see if the same conclusions hold.

**Acknowledgments**. The author thanks Mary-Katharine Juric, Krishnaram Kenthapadi, Kun Liu, Paul Rockwell, Ya Xu, and the anonymous reviewers for helpful feedback on earler versions of this work.

# 6. REFERENCES

[1] G. Barnard. A new test for $2 \times 2$ tables. *Nature*, 156:177, 1945.

[2] J. Bonneau, E. Bursztein, I. Caron, R. Jackson, and M. Williamson. Secrets, lies, and account recovery:

Lessons from the use of personal knowledge questions at Google. In *WWW*, pages 141–150. ACM, 2015.

[3] J. Bonneau, C. Herley, P. van Oorschot, and F. Stajano. The past, present, and future of password-based authentication on the web. *Commun. ACM*, 2015.

[4] E. Bursztein, B. Benko, D. Margolis, T. Pietraszek, A. Archer, A. Aquino, A. Pitsillidis, and S. Savage. Handcrafted fraud and extortion: Manual account hijacking in the wild. In *IMC*, pages 347–358. ACM, 2014.

[5] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *NDSI*, 2012.

[6] Q. Cao and X. Yang. SybilFence: Improving social-graph-based sybil defenses with user negative feedback. Duke CS Technical Report: CS-TR-2012-05, available at `http://arxiv.org/abs/1304.3819`, 2013.

[7] L. Chen, Z. Yan, W. Zhang, and R. Kantola. TruSMS: A trustworthy SMS spam control system based on trust management. *Future Generation Computer Systems*, 49:77–93, 2015.

[8] P. H. Chia and S. J. Knapskog. Re-evaluating the wisdom of crowds in assessing web security. In *International Conference on Financial Cryptography and Data Security*, pages 299–314. Springer, 2011.

[9] G. Danezis and P. Mittal. SybilInfer: Detecting sybil nodes using social networks. In *NDSS*, 2009.

[10] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[11] Facebook Inc. Statement of rights and responsibilities. `http://www.facebook.com/legal/terms`, accessed 19 Feb 2017.

[12] R. A. Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[13] D. M. Freeman, S. Jain, M. Dürmuth, B. Biggio, and G. Giacinto. Who are you? A statistical approach to measuring user authenticity. In *NDSS*, 2016.

[14] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW*, pages 403–412. ACM, 2004.

[15] J. Hauke and T. Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93, 2011.

[16] LinkedIn Corporation. User agreement. `http://www.linkedin.com/legal/user-agreement`, accessed 19 Feb 2017.

[17] S. Marti and H. Garcia-Molina. Taxonomy of trust: Categorizing P2P reputation systems. *Computer Networks*, 50(4):472–484, 2006.

[18] C. R. Mehta and P. Senchaudhuri. Conditional versus unconditional exact tests for comparing two binomials. Cytel Software corporation, 2003. Available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.572.632&rep=rep1&type=pdf`.

[19] A. Mohaisen, N. Hopper, and Y. Kim. Keep your friends close: Incorporating trust into social network-based sybil defenses. In *IEEE INFOCOM 2011*, pages 1943–1951. IEEE, 2011.

[20] T. Moore and R. Clayton. Evaluating the wisdom of crowds in assessing phishing websites. In *International Conference on Financial Cryptography and Data Security*, pages 16–30. Springer, 2008.

[21] D. Powers. Evaluation: From precision, recall and $F$-factor to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2:37–63, 2011.

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[23] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

[24] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao. You are how you click: Clickstream analysis for sybil detection. In *Proceedings of The 22nd USENIX Security Symposium*, pages 241–256, 2013.

[25] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Y. Zhao. Social Turing tests: Crowdsourcing sybil detection. 2013.

[26] C. Xiao, D. M. Freeman, and T. Hwa. Detecting clusters of fake accounts in online social networks. In *AISec*, pages 91–101. ACM, 2015.

[27] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[28] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. Sybillimit: A near-optimal social network defense against sybil attacks. In *IEEE Symposium on Security and Privacy*, pages 3–17. IEEE, 2008.

[29] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. Sybilguard: Defending against sybil attacks via social networks. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 267–278. ACM, 2006.

[30] E. Zheleva, A. Kolcz, and L. Getoor. Trusting spam reporters: A reporter-based reputation system for email filtering. *ACM Transactions on Information Systems (TOIS)*, 27(1):3, 2008.