# Strengthening Weak Identities Through Inter-Domain Trust Transfer

Giridhari Venkatadri
MPI-SWS, Germany

Oana Goga
MPI-SWS, Germany

Changtao Zhong
King's College London, UK

Bimal Viswanath
Nokia Bell Labs, Germany

Krishna P. Gummadi
MPI-SWS, Germany

Nishanth Sastry
King's College London, UK

## ABSTRACT

On most current websites untrustworthy or spammy identities are easily created. Existing proposals to detect untrustworthy identities rely on reputation signals obtained by observing the activities of identities *over time* within a *single site* or domain; thus, there is a time lag before which websites cannot easily distinguish attackers and legitimate users. In this paper, we investigate the feasibility of leveraging information about identities that is aggregated across multiple domains to reason about their trustworthiness. Our key insight is that while honest users naturally maintain identities across multiple domains (where they have proven their trustworthiness and have acquired reputation over time), attackers are discouraged by the additional effort and costs to do the same. We propose a flexible framework to transfer trust between domains that can be implemented in today's systems without significant loss of privacy or significant implementation overheads. We demonstrate the potential for inter-domain trust assessment using extensive data collected from Pinterest, Facebook, and Twitter. Our results show that newer domains such as Pinterest can benefit by transferring trust from more established domains such as Facebook and Twitter by being able to declare *more* users as likely to be trustworthy much *earlier on* (approx. one year earlier).

## 1. INTRODUCTION

In this paper, we focus on the fundamental challenge of assessing the trustworthiness of identities or accounts in online services ranging from Web service providers like Gmail and social networking / media sites like Facebook and Twitter to crowdsourcing sites like Amazon's Mechanical Turk and e-Commerce sites like eBay. By trustworthiness of an identity, we mean the likelihood that the identity behaves properly in the future (e.g., conforming to the site's Terms of Service).

Determining the trustworthiness of online identities is particularly difficult today because most domains allow users to operate behind *weak identities* – identities that can be created by a user without providing any proof of the user's real (offline) identity (e.g., by the way of passports or credit cards). The lack of trusted references for weak identities makes it hard to hold their owners accountable for misbehavior. For instance, a misbehaving Twitter user, whose identity has been detected and suspended, can evade the system by creating a new identity, with a clean-slate reputation, and continue misbehaving. While *strong identities*, which are verified or certified by trusted authorities (e.g., by way of passport verification), offer greater accountability, they are rarely used in practice because of $(i)$ the privacy concerns they raise for users and $(ii)$ the high sign-on overheads involved in creating new identities. As a result, many online domains (and recent research works) today focus on developing methods to assess the trustworthiness of weak identities [4, 27, 32, 33, 35, 38, 40, 44].

Since weak identities in a domain are created without any *external* references of trust, all trust assessment methods rely *solely* on analyzing the past activity of each identity within the domain. So, while these methods are effective at assessing the trustworthiness of older and active identities with longer historical records of activity within the domain, they are not effective at determining the trustworthiness of newly created or occasionally used identities with limited or no past activity history. This limitation of existing methods causes two key problems: $(i)$ malicious users can exploit newly created identities to launch attacks that are effective until the identities' misbehavior is eventually discovered [24]; $(ii)$ honest users have to patiently groom their newly created identities by exhibiting good behavior over an extended period of time in order to acquire reputation, influence, and access to various resources on the domain.[1]

Against this background, this paper investigates a previously unexploited opportunity to better assess the trustworthiness of a weak identity in an online domain using the weak identities that the owner of the identity has previously created on *other domains*. Specifically, the paper asks: *can trust assessments of weak identities on one domain be strengthened by users providing references of weak identities they created on other domains?*

This question is worth investigating for the following reasons: $(i)$ many honest users already maintain weak identities on multiple online domains (e.g., Gmail, Facebook, Twitter, YouTube, and Pinterest), and acquire different levels of reputation on different domains over time based on their activity on those domains; $(ii)$ many honest users are already beginning to interconnect these identities during sign-ons (e.g., providing their Gmail identities when logging in to Facebook or using social logins like Facebook Login to log in to Pinterest); and $(iii)$ malicious users do not naturally create and groom fake identities on multiple domains over a long period of time [42]. So, unlike honest users, malicious users would incur additional effort and costs if they had to present weak identities with high reputation on other domains. Crucially, at the same time,

[1] For example, Reddit has strict posting quotas for new users which get relaxed as they exhibit good behavior and gain reputation [12].

*WWW 2016,* April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883015.

by allowing weak identities to be used as trust references rather than strong identities such as passports, honest users can still keep their real (offline) identities anonymous (i.e., unlinked to their online identities). $(iv)$ Lastly, there are more established domains such as Facebook, Twitter and Google with sophisticated malicious identity detection systems that could provide good trust references for identities in newer or emerging domains.

In this paper, we argue that online domains should move beyond their current practice of assessing the trustworthiness of identities within their own domain separately and independently of other domains. Instead, we propose an *inter-domain trust transfer framework* that focuses on enabling trust transfer between a source domain and a target domain. Our approach is practical, can be deployed by domains today with low overheads, and offers incentives for both source and target domains. In our framework, source domains transfer reputation or trustworthiness signals about their identities. The signals are computed by source domains independently, based on past activities of identities within their own domain. They serve as signals for the degree of trust the source domain places in its identities, but reveal little about the actual activities of the identities within the domain. Note that such limited information disclosure is also in the self-interests of the domains and also helps to protect user privacy. Each target domain must then re-interpret or re-calibrate the exported signals based on their ability to predict the trustworthiness of identities within the target domain. We introduce relevant terminology in §2 and discuss the framework in detail in §3.

To evaluate our framework, we first study the potential for inter-domain trust transfer by investigating whether reputation and trustworthiness signals on Facebook and Twitter can help to estimate trustworthiness of identities on Pinterest. To this end, we gathered extensive data about 1.7 million identities from Pinterest, a popular social bookmarking domain, and their matching (linked) identities from Facebook and Twitter. Our analysis of this data, presented in §5.2, shows that there is significant potential for leveraging even very simple signals from Facebook or Twitter (computed using the activities of identities within Facebook and Twitter, respectively) to infer the trustworthiness of the matching identities in Pinterest.

Secondly, we demonstrate a practical application of our framework – to *curate* (or deem trustworthy) identities in a domain (§5.3). Identity curation is useful because domains can grant curated identities access to elevated privileges within the domain without worrying about any potential service abuse. We quantify the quality of a curated set of identities using a metric called *purity* which measures the fraction of trustworthy identities in the set. Using our framework, we show that compared to creating a curated set with a certain purity level (0.975) using only information available on Pinterest, we can *double* the number of curated identities by leveraging simple inter-domain trust signals drawn from Twitter and Facebook. Furthermore, relying only on the intra-domain signals in Pinterest would require waiting for up to 15 months for new identities to acquire sufficient reputation to be curated (with a purity level of 0.975); in contrast, new identities can be curated more than a year in advance using inter-domain trust transfer.

## 2. CONCEPTS AND BACKGROUND

In this section, we first discuss the terminology used in the rest of the paper and then provide background on the most widely used approach for estimating trustworthiness of identities today.

### *Terminology.*

**User:** A person/entity $\mathcal{U}$ who uses a service on the Web.

**Domain:** An independent administrative entity $\mathcal{D}$ providing some service on the Web. Examples include webmail, e-commerce, and social media systems. We call the domain from which we transfer trust the *source domain*, and the domain to which we transfer trust the *target domain*.

**Identity:** An account $\mathcal{I}_\mathcal{D}$ created in a domain $\mathcal{D}$ and managed by a user $\mathcal{U}$ to access services offered by the domain. An identity is always associated with a domain and a user.[2] We call the identities user $\mathcal{U}$ has across various domains $\mathcal{D}_i$ the *matching identities* of $\mathcal{U}$.

**Reputation of an identity:** A function of *past* activities of the identity within the domain. Past activities include content posted by the identity, interaction with other identities, and even network-level information (e.g., IP address) associated with web requests. The reputation of an identity can be characterized by several *reputation signals*, $\mathcal{R}_j(\mathcal{I}_\mathcal{D}), j = 1..m$. Reputation signals can be simple functions of activities (e.g., age of the identity since creation) or complex functions of activities (e.g., influence of the identity in the social graph) and they can reflect good behavior (e.g., number of followers), or misbehavior (e.g., number of posted links to malicious sites).

**Trustworthiness of an identity:** The likelihood that the identity will respect the terms of service (ToS) of its domain in the *future*, denoted by $Trust(\mathcal{I}_\mathcal{D})$. Note that while reputation is a function of past activities of an identity, trustworthiness is a prediction for the future. A *trustworthy* identity is an identity that has a low probability to misbehave (i.e., violate ToS), while an *untrustworthy* identity is an identity that has a high probability to misbehave. A *curated set of identities* is a set of identities that the domain believes are trustworthy.

### *Intra-domain estimation of trustworthiness.*

To estimate trustworthiness of identities today, online domains rely on reputation signals [37] independently sourced from within their own domain. Extensive research has been conducted on leveraging different reputation signals to infer trustworthiness of identities [16, 24, 45]. At a high level, these approaches work by comparing reputation scores of identities that misbehaved (i.e., violated ToS) and identities that did not misbehave, and creating patterns of legitimate and suspicious behavior. Using these patterns, the domain can thereafter compute the probability of an identity to misbehave in the future given its reputation scores, $Trust(\mathcal{I}_\mathcal{D}) = f(\mathcal{R}(\mathcal{I}_\mathcal{D}))$. When an identity has a high probability to misbehave (is untrustworthy), the domain can take appropriate action (serve a CAPTCHA or suspend the identity) to limit the identity's interaction with the service.

There are countless reputation signals that can be associated with an identity. However, a reputation signal is effective at inferring trustworthiness only when it $(i)$ exhibits different scores for trustworthy and untrustworthy identities, and $(ii)$ is *hard to acquire*. The second requirement is important because an easy to acquire reputation score can be easily tampered by an attacker to boost the reputation scores of identities under their control. But, if the reputation signal is hard to acquire, a profit driven attacker might be disincentivized to spend more financial resources to create fake identities and groom them until they achieve a high reputation score. For

[2] A user can create multiple identities within a single domain (e.g., one webmail identity for work and another for interacting with friends and family). However, for the sake of simplicity, we focus on transferring trust from one identity of a user in the source domain to another single identity in the target domain, even though transferring trust between multiple identities is possible.

example, a newly created fake Twitter identity on black markets costs \$0.09, but a fake Twitter identity that is five years old costs \$2! [5].

Lastly, it is important to note that all intra-domain techniques for estimating trustworthiness of individual identities are not effective when analyzing newly created or inactive identities with limited or no past activity history. Our framework, described in the next section, tries to address this limitation and lays out a new direction to infer trustworthiness of identities.

# 3. INTER-DOMAIN TRUST TRANSFER FRAMEWORK

In this section, we present a practical and easy to implement framework that enables inter-domain trust transfer. We first present the design of our framework followed by a discussion on different use cases of our framework.

## 3.1 Framework design

The scenario we consider is of a target domain $\mathcal{T}$, on which a user $\mathcal{U}$ has an identity $\mathcal{I}_{\mathcal{T}}$, and a number of source domains $\mathcal{S}_i$, $i = 1..n$, on which $\mathcal{U}$ has matching identities $\mathcal{I}_{\mathcal{S}_i}$ (see Figure 1). The goal of our framework is to enable the domain $\mathcal{T}$ to estimate the trustworthiness of the identity $\mathcal{I}_{\mathcal{T}}$ by leveraging information $Inf(\mathcal{I}_{\mathcal{I}_{\mathcal{S}_i}})$ about its matching identities $\mathcal{I}_{\mathcal{S}_i}$, along with their reputation $\mathcal{R}(\mathcal{I}_{\mathcal{T}})$ on $\mathcal{T}$:

$$Trust(\mathcal{I}_{\mathcal{T}}) = f(Inf(\mathcal{I}_{\mathcal{S}_1}), ..., Inf(\mathcal{I}_{\mathcal{S}_n}), \mathcal{R}(\mathcal{I}_{\mathcal{T}})). \quad (1)$$

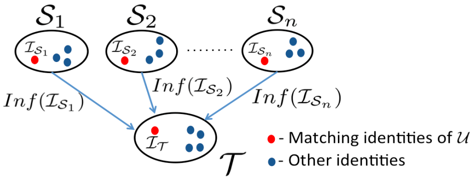

$\mathcal{S}_1$
$\mathcal{S}_2$
$\mathcal{S}_n$
$\mathcal{I}_{\mathcal{S}_1}$
$\mathcal{I}_{\mathcal{S}_2}$
........
$\mathcal{I}_{\mathcal{S}_n}$
$Inf(\mathcal{I}_{\mathcal{S}_1})$
$Inf(\mathcal{I}_{\mathcal{S}_2})$
$Inf(\mathcal{I}_{\mathcal{S}_n})$
$\mathcal{I}_{\mathcal{T}}$
$\mathcal{T}$
● - Matching identities of $\mathcal{U}$
● - Other identities

Figure 1: System model

To enable inter-domain trust transfer: $(i)$ the target domain needs a platform that enables the linkage of $\mathcal{I}_{\mathcal{T}}$ to its matching identities $\mathcal{I}_{\mathcal{S}_i}$; $(ii)$ the source domains need to decide what information to transfer to the target domain about $\mathcal{I}_{\mathcal{S}_i}$ (decide on $Inf(\mathcal{I}_{\mathcal{S}_i})$); and $(iii)$ the target domain needs a strategy to exploit information from multiple $\mathcal{S}_i$ to reason about $\mathcal{I}_{\mathcal{T}}$ (instantiate $f$). We next present solutions for each task.

Note that $\mathcal{T}$ does not need to know all the matching identities $\mathcal{I}_{\mathcal{S}_i}$ of $\mathcal{I}_{\mathcal{T}}$. Our framework can work in an opportunistic manner: if there is information about the matching identities of $\mathcal{I}_{\mathcal{T}}$ (be it one or several), then the framework exploits the available information. Intuitively, as more matching identities $\mathcal{I}_{\mathcal{S}_i}$ are linked to $\mathcal{I}_{\mathcal{T}}$, we can strengthen $\mathcal{I}_{\mathcal{T}}$ further (i.e., prove $\mathcal{I}_{\mathcal{T}}$ is trustworthy).

### *3.1.1 Linking matching identities*

There are different ways $\mathcal{T}$ can enable the linkage of $\mathcal{I}_{\mathcal{T}}$ to its matching identities $\mathcal{I}_{\mathcal{S}_i}$. We split the discussion into four methods based on which entities need to be explicitly involved and give consent. Different techniques involve different privacy and legal risks which we will discuss in §4.2 and §4.3.

**1. $\mathcal{T}$ links identities with involvement of $\mathcal{U}$ and $\mathcal{S}_i$:** The cleanest and simplest scenario is when the user $\mathcal{U}$ herself links her identities $\mathcal{I}_{\mathcal{S}_i}$ to $\mathcal{I}_{\mathcal{T}}$. If the target and the source domain support single sign-on protocols such as OpenID, the linkage is straightforward. Single sign-on is a user authentication protocol that allows a user to use one credential to connect to multiple sites or applications. In the past years, most major social networks such as Facebook, Google+, Twitter, and LinkedIn have adopted such protocols and are allowing users to use their identities in these social networks to authenticate on other domains such as Pinterest, Quora, and Airbnb. When users use such login protocols, they effectively allow their identities in the source domain and in the target domain to be linked. The advantage of this approach is that the technology is already very popular – more than 24% of the top 10,000 websites (according to Alexa) have some form of integration with Facebook [14]. Furthermore, once two identities are linked using this approach, it is possible for the user to further allow the source domain to export information about her identity using an authentication protocol such as OAuth [1]. This approach has the key benefit that both linkage and information transfer are possible today, without the need for new technology, and with little overhead for domains.

**2. $\mathcal{T}$ and $\mathcal{U}$ link identities without the involvement of $\mathcal{S}_i$:** In situations where $\mathcal{S}_i$ does not support single sign-on methods, $\mathcal{U}$ can show proof of ownership of $\mathcal{I}_{\mathcal{S}_i}$ in ways similar to proofs of DNS domain ownership: $\mathcal{U}$ can prove that she owns an identity $\mathcal{I}_{Tw}$ on Twitter by tweeting (as $\mathcal{I}_{Tw}$) a nonce message specified by $\mathcal{T}$. Once the linkage is done, $\mathcal{T}$ can use the publicly available information on $\mathcal{I}_{\mathcal{S}_i}$ to estimate the trustworthiness of $\mathcal{I}_{\mathcal{T}}$.

**3. $\mathcal{T}$ and $\mathcal{S}_i$ link identities without involvement of $\mathcal{U}$:** $\mathcal{T}$ and $\mathcal{S}_i$ might decide to collaborate and merge their user databases without the consent or knowledge of $\mathcal{U}$. Linking identities across databases can be relatively straightforward: users normally need an email address when they create accounts on different domains; thus, the domains could use the email addresses as keys to link the identities of the same users. Furthermore, additional information such as the IP addresses from which users log in can be used to increase the accuracy of linking identities. Note that domains are not required to exchange actual email or IP addresses, as that may violate both the privacy of users and domains. Instead, domains can use standard replacements such as hashes of the identity strings to check whether there is a match. Information about identities can subsequently be transferred only in case of a match.

**4. $\mathcal{T}$ links identities without involvement of $\mathcal{U}$ or $\mathcal{S}_i$:** When $\mathcal{T}$ wants to link identity $\mathcal{I}_{\mathcal{T}}$ with $\mathcal{I}_{\mathcal{S}_i}$, it may be possible to search whether $\mathcal{U}$ has identities on other domains without the involvement of either $\mathcal{U}$ or $\mathcal{S}_i$. For instance, it has been shown that it is possible to find the matching identities of users in an automatic way by exploiting what users reveal about themselves [23] (e.g., the name, the profile photo, the location). Once the matching identities are obtained, $\mathcal{T}$ can use the publicly available information of the matching identities to estimate the trustworthiness of $\mathcal{I}_{\mathcal{T}}$.

Method 1 is easy to implement using existing technologies and requires consent from all entities, and thus could be a practical way of linking identities today.

### *3.1.2 Choosing what information to transfer*

The source domain needs to transfer information, $Inf(\mathcal{I}_{\mathcal{S}_i})$, that will be useful to estimate the trustworthiness of identities in other domains. There are two types of signals that could be useful: $(i)$ *reputation signals* - $\mathcal{R}_j(\mathcal{I}_{\mathcal{S}_i})$ or $(ii)$ *trustworthiness signals* - $Trust(\mathcal{I}_{\mathcal{S}_i})$. The difference between reputation and trustworthiness signals from an operational point of view is that reputation is purely a function of past activities of an identity, while trustworthiness is a function of both past activities of an identity and the knowledge that $\mathcal{S}_i$ has from previous attacks. The trustworthiness of an identity can be seen as the source domain's own interpretation of what a reputation of the identity means. For example, if $\mathcal{I}_{\mathcal{S}_i}$ is 3 days old and posted 100 posts in the past 10 minutes, then $\mathcal{I}_{\mathcal{S}_i}$ will misbehave in the future with a likelihood of 0.99.

The choice between reputation signals and trustworthiness signals is not straightforward. If the trustworthiness of identities correlates well across domains (i.e., if a user misbehaves in $\mathcal{S}_i$, she will also misbehave in $\mathcal{T}$), then a trustworthiness signal is the ideal information for $\mathcal{T}$ because it can take advantage of the well tuned trust reasoning system of the source domain. However, in many cases, the correlation might not be perfect. For example, Facebook might assign a low trustworthiness to identities with fake names (since it has a strict real name policy). Such a trustworthiness signal computed by Facebook may not be useful to Twitter, which does not require the use of real names. In this case, reputation signals might be more useful because they allow $\mathcal{T}$ to make its own interpretation of what a particular reputation means in terms of trustworthiness in $\mathcal{T}$. The source domain can freely choose to transfer either reputation or trustworthiness signals or even both; however, it also needs to consider in its decision the risks involved by sharing such data (which we will discuss in §4.2).

### *3.1.3 Exploiting the information received*

Since we let each source domain decide for itself what reputation or trustworthiness signals to compute and export, it is important that $\mathcal{T}$ is able to interpret the transferred information in the context of trustworthiness in $\mathcal{T}$. There are two main challenges in doing this. Firstly, how does $\mathcal{T}$ estimate the usefulness of a reputation (or trustworthiness) signal transferred by $\mathcal{S}_i$ for inferring the trustworthiness of identities on $\mathcal{T}$? For example, what does the information that a user has an identity with 100 friends on Facebook indicate about the trustworthiness of the matching identity on Pinterest? Secondly, how does $\mathcal{T}$ compare a reputation (or trustworthiness) signal from one source domain with a signal from another source domain? For example, is a user more likely to be trustworthy on Pinterest if she has 1000 followers on Twitter or 100 friends on Facebook?

We propose a *calibration* step that enables $\mathcal{T}$ to solve these challenges and to exploit information $Inf(\mathcal{I}_{\mathcal{S}_i})$ from multiple source domains, in addition to information from its own domain $\mathcal{R}(\mathcal{I}_{\mathcal{T}})$ to reason about the trustworthiness of $\mathcal{I}_{\mathcal{T}}$. Concretely, the calibration step instantiates $f$ from Equation (1). Technically, the process of calibration is similar to the process of estimating trustworthiness from reputation signals within the same domain (see §2); the difference is that we additionally use signals from other source domains. Thus, in the calibration step, $\mathcal{T}$ takes identities that misbehaved in $\mathcal{T}$ (negative examples) and identities that did not misbehave (positive examples) and trains a classifier with their reputation $\mathcal{R}(\mathcal{I}_{\mathcal{T}})$ in $\mathcal{T}$ as well as information about their matching identities $Inf(\mathcal{I}_{\mathcal{S}_i})$ in other domains.[3] Note that $\mathcal{T}$ does not need to know the meaning of the signals exported by $\mathcal{S}_i$ to perform the calibration step. Additionally, the calibration step can easily handle the fact that target domains may have limited information about the matching identities of a user. If $\mathcal{T}$ does not know the matching identity of a user in a given source domain $\mathcal{S}_i$, then $\mathcal{T}$ can treat the corresponding signals as missing values when building the classifier.[4] Thus, the resulting classifier will allow $\mathcal{T}$ to estimate the trustworthiness of $\mathcal{I}_{\mathcal{T}}$ based on *any* information available to $\mathcal{T}$ about $\mathcal{I}_{\mathcal{T}}$.

## 3.2 Usage scenarios

We discuss the different applications of our trust transfer framework. Our framework is inherently useful for detecting trustworthy identities and raising barriers for attackers.

### *3.2.1 Applications of identity curation*

We define a curated identity as one which the domain believes is trustworthy (i.e., that will not misbehave in the future). A domain can leverage reputation or trustworthiness signals transferred from other domains to better curate its identities. Identity curation has different applications; we discuss two useful applications of identity curation, and explain how they can benefit from our trust transfer framework.

**Early access to elevated privileges for identities:** Domains typically curate identities that have spent considerable time in the domain building their reputation, and provide elevated privileges to the curated identities. For example, to combat spam and trolling, Reddit strictly rate-limits posts of new identities until they have been in the system for a period of time and shown markers of good behavior [12]. However, our trust transfer framework allows a user that already has a good record at writing Wikipedia articles to link her Wikipedia identity to her Reddit identity; the reputation signals exported by Wikipedia could then potentially let Reddit curate the user's identity and lift the rate-limits earlier on.

**From identity curation to content curation:** Any social media system has a certain amount of spammy or unwanted content, and it is important that this content does not get promoted by the domain's recommendation systems. To address this problem, previous works have proposed only promoting content from a curated set of identities [22]. In order to provide a good coverage of trustworthy content, it is essential for the domain to be able to curate a large fraction of identities in the system. However, any approach to curation that relies only on activities of identities within the domain requires analyzing the behavior of identities over a long period of time, hence making it hard for the domain to curate newer identities. This can be a significant limitation, especially for newer domains. However, our trust transfer framework allows domains to additionally curate new identities that have matching identities with long activity histories on other domains. We discuss this further in §5.3.

### *3.2.2 Outsourcing Sybil defenses*

Newly deployed domains may find it hard to afford a dedicated security team to build a Sybil detection system. Trust transfer could be especially beneficial in such a scenario because it allows the target domain to take advantage of the potentially sophisticated Sybil detection systems that established source domains employ. For example, a target domain could *require* its users to use social logins and login with a valid Facebook identity. Potential spammers would be forced to create an identity on Facebook first and would be subject to its anti-spam systems, which are presumably more sophisticated. Note that if presenting identities on other domains is *optional*, than attackers could hide their misbehavior by simply not linking their identities. If the target domain, however, only grants limited privileges to newly created identities of users that do not prove they have been trustworthy (for a reasonable amount of time) in other domains, then not presenting an identity on a source domain (or presenting a newly created identity) would result in less effective attacks.

[3] To calibrate, a target $\mathcal{T}$ needs examples of misbehaving identities on its own domain. However, if $\mathcal{T}$ is new, it might not have such information. In this case, instead of the calibration step, $\mathcal{T}$ could simply consider that if a user shows he has an identity on $\mathcal{S}_i$ he is trustworthy – $\mathcal{T}$ simply relies on the Sybil detection system of $\mathcal{S}_i$.

[4] Handling missing values is a standard task in machine learning.

# 4. DISCUSSION

In this section, we discuss the incentives for different parties to participate in our framework and further discuss certain risks associated with our framework and techniques to mitigate the risks.

## 4.1 Incentives

**Incentives to source domains:** Many established domains such as Facebook and Google are in stiff competition for a greater share of the social login market, in a bid to control a greater share of users' social activities across the web [13]. By exporting reputation signals or trustworthiness signals, a source domain can increase the value offered by its social login platform to users and target domains, leading to more users and target domains opting to use its social login platform over that of other source domains.

**Incentives to users:** Trustworthy users would be incentivized to participate in the framework by linking their identities because this would help them gain early access to elevated privileges on the target domain (as discussed in section §3.2).

**Incentives to target domains:** Target domains clearly stand to benefit from our framework as it provides them with external references of trust for identities, thereby enabling them to reason about trustworthiness of identities better and earlier. As discussed in section §3.2, emerging domains (with mostly newly created identities) would be particularly incentivized to participate as target domains since they might have limited resources for fighting Sybils.

## 4.2 Privacy risks

Whenever a system enables data sharing it inevitably faces risks related to what can be learned from the data aside from its main purpose. In this section, we discuss privacy risks incurred by our trust transfer framework and techniques to mitigate the risks.

When a source domain $\mathcal{S}_{good}$ transfers information to a malicious target domain $\mathcal{T}_{bad}$, the information could allow $\mathcal{T}_{bad}$ to learn potentially sensitive information about $\mathcal{U}$ which would violate his privacy. One approach to limit the privacy risk would be to allow users to only share attributes they consider to be non-sensitive; this can easily be implemented as part of OAuth. However, despite this, users may still be vulnerable to privacy loss due to possible correlations between the transferred signals and sensitive attributes. Clearly, such correlation is possible if the exported signals $\mathcal{R}(\mathcal{I}_\mathcal{S})$ and $Trust(\mathcal{I}_\mathcal{S})$ have been computed using sensitive attributes. However, there can be a correlation even if $\mathcal{R}(\mathcal{I}_\mathcal{S})$ and $Trust(\mathcal{I}_\mathcal{S})$ are not directly computed using sensitive attributes. For example, if $\mathcal{S}_{good}$ exported the frequency of photo uploads by an identity as a reputation signal, this could correlate with the user's age because teenagers might upload photos with relatively high frequency.[5]

In order to limit the extent to which $\mathcal{T}_{bad}$ can infer sensitive attributes from $\mathcal{R}(\mathcal{I}_\mathcal{S})$ and $Trust(\mathcal{I}_\mathcal{S})$, $\mathcal{S}_{good}$ could transform $\mathcal{R}(\mathcal{I}_\mathcal{S})$ and $Trust(\mathcal{I}_\mathcal{S})$ to limit their correlation with the sensitive attributes.[6] Fortunately, there is a lot of literature in the database community tackling similar problems in the context of sharing anonymized databases. Notions such as *t-closeness* [29] have been proposed to limit the extent to which sensitive attributes of individuals can be inferred from their non-sensitive attributes that are released in the anonymized database. By applying t-closeness on a database comprising of $\mathcal{R}(\mathcal{I}_\mathcal{S})$ and $Trust(\mathcal{I}_\mathcal{S})$ for all identities $\mathcal{I}_\mathcal{S}$, $\mathcal{S}_{good}$, can transform $\mathcal{R}(\mathcal{I}_\mathcal{S})$ and $Trust(\mathcal{I}_\mathcal{S})$ to coarser-grained variants, $\mathcal{R}'(\mathcal{I}_\mathcal{S})$ or $Trust'(\mathcal{I}_\mathcal{S})$. This step would lead to the property that for any given values of $\mathcal{R}'(\mathcal{I}_\mathcal{S})$ or $Trust'(\mathcal{I}_\mathcal{S})$, the set of identities with those values have a similar distribution of sensitive attributes as the global distribution of sensitive attributes across all identities in $\mathcal{S}_{good}$. This implies that the amount of information that an attacker can learn about the sensitive attributes of any identity from $\mathcal{R}'(\mathcal{I}_\mathcal{S})$ or $Trust'(\mathcal{I}_\mathcal{S})$ is limited.

An algorithm has been proposed in [29] that, given $\mathcal{R}(\mathcal{I}_\mathcal{S})$ or $Trust(\mathcal{I}_\mathcal{S})$, can efficiently compute $\mathcal{R}'(\mathcal{I}_\mathcal{S})$ or $Trust'(\mathcal{I}_\mathcal{S})$ that leads to the minimal loss in utility (i.e., limiting privacy risk while minimizing the loss in utility). $\mathcal{S}_{good}$ can then export $\mathcal{R}'(\mathcal{I}_\mathcal{S})$ or $Trust'(\mathcal{I}_\mathcal{S})$. The t-closeness approach allows $\mathcal{S}_{good}$ to pick a particular bound for the privacy loss depending on the privacy vs. utility tradeoff it wants to achieve. The utility loss given a particular bound on the privacy loss depends on the strength of correlation between $\mathcal{R}(\mathcal{I}_\mathcal{S})$ or $Trust(\mathcal{I}_\mathcal{S})$, and the sensitive attributes.

Besides privacy risks resulting from the information transferred by our framework, the mere linking of identities can result in a privacy risk because $\mathcal{T}_{bad}$ could learn sensitive attributes from the public data shared by users in $\mathcal{S}_{good}$ [26]. This is a general problem arising from the linkage of identities for any purpose. Nevertheless, one way to limit this risk is to anonymously link identities (e.g., Facebook anonymous login [9]), where $\mathcal{T}_{bad}$ only receives an anonymous identifier from $\mathcal{S}_{good}$ that cannot be used to identify the public profile of the matching identity on $\mathcal{S}_{good}$. However, we acknowledge the limitation that a determined attacker could still identify the matching identity in $\mathcal{S}_{good}$ using other channels (e.g., general search engines or search interfaces provided by $\mathcal{S}_{good}$) by leveraging any information (about the identity) available on $\mathcal{T}_{bad}$.

## 4.3 Ethical and legal risks

Ethical and legal aspects are especially important when transferring information about identities between different domains. In §3.1, we proposed various methods by which trust transfer can happen between domains. Methods that do not require the involvement of the users to link their identities risk violating user privacy, and running afoul of new regulations such as the "right to be forgotten" in the EU. Also, methods that involve scraping public information from source domains might violate the ToS of the source domains. One of the methods proposed reduces these legal risks by requiring the involvement and consent of all parties involved: when linking identities using social logins, the users themselves initiate the linkage; when information is transferred through OAuth, both the source domain and users agree on the information that is being transferred.

From an ethical point of view, it is questionable whether a user should be treated differently in a target domain if he misbehaved in another (source) domain. However, the calibration step from §3.1.3 allows the target domain to not blindly use the transferred information but to interpret it in the context of its own domain.

# 5. EVALUATION

We now evaluate the potential for inter-domain trust transfer to reason about the trustworthiness of identities. In the rest of this section, we consider the popular image sharing site, Pinterest, to be the target domain and evaluate whether it can benefit from trustworthiness or reputation signals exported by source domains, Facebook and Twitter. Such a study is possible because Pinterest allows users to link their Pinterest identities to their Facebook, Twitter or Email identities via social logins (§3.1.1).

[5] We assume that the attacker is able to create a number of fake identities on $\mathcal{S}_{good}$ (and link them to identities on $\mathcal{T}_{bad}$), and to trick some honest users to link their identities on $\mathcal{S}_{good}$ with their identities on $\mathcal{T}_{bad}$. The attacker can thereafter use these identities as *training examples* to find correlations between exported signals and sensitive attributes.

[6] To simplify, we assume that the sensitive attributes are known in advance.

| | Facebook | Twitter | Email |
|---|---|---|---|
| Random | 55% | 7% | 38% |
| Suspended | 25% | 13% | 62% |
| Black market | 31% | 21% | 48% |

Table 1: Source domain distribution for random, suspended and black market identities in Pinterest.

## 5.1 Data

We collected three types of data for our analysis: $(i)$ Information about identities owned by users across multiple domains, Pinterest, Twitter, Facebook, and email domains (i.e., the *matching identities* of the user). Given an identity created by a user on Pinterest, this information enables us to find the matching identities of the user on Twitter, Facebook, or email domains. $(ii)$ Reputation and trustworthiness signals associated with identities in the source domains (Twitter and Facebook). We leverage these signals to reason about the trustworthiness of the matching identities in Pinterest. $(iii)$ Ground truth information about untrustworthy identities in Pinterest, which enables us to evaluate how well we can reason about trustworthiness of identities in the target domain.

### *5.1.1 Information about matching identities*

Users are allowed to log in to Pinterest with their Twitter or Facebook identities through social logins [10], or with their email accounts. When a user uses his Twitter or Facebook identity to log in to Pinterest, a link to his Twitter or Facebook profile appears on his public Pinterest profile by default. We look for such links to the social domains to find matching identities of Pinterest users on Twitter or Facebook. If there is no link to the above two social domains, we assume that the user has only used their email account to create the account on Pinterest.[7]

Before finding matching identities in the source domains, we first need a sample of identities in the target domain (Pinterest). We collected information about a random sample of 1.7M Pinterest identities. In the rest of the paper, we refer to these identities as *random identities*.[8] Using the methodology described in the previous paragraph, we find the matching identities of these random identities on Twitter and Facebook. The first row of Table 1 shows the proportion of Pinterest users who logged in with their Twitter, Facebook or email identities. We do not have information about the exact domains of the email accounts, so we group all the email identities together. We observe that a majority (62%) of Pinterest users linked their Facebook or Twitter identities with their Pinterest identity, with Facebook being the more popular source domain.

### *5.1.2 Reputation and trustworthiness signals*

For all the matching identities in the source domains, we collected data that captures reputation and trustworthiness of the identities. While we are limited to computing reputation signals that can only be derived from publicly available data, in practice, the source domain could potentially use any information associated with identities to compute reputation signals. However, this gives us the opportunity to *evaluate the potential for transferring trust, using some of the simplest and most basic (easy to compute) reputation signals* derived from publicly visible data.

**Reputation signals:** We collected data about reputation signals based on the following two types of information associated with the identities:

(1) *Domain*: The simplest signal that can be associated with an identity is its domain (e.g., Email, Twitter, or Facebook). The intuition here is that identities in some domains take more effort to create than others, and attackers might choose to link their Pinterest identities to identities in source domains where it is easier to create them. This simple signal reflects the fact that a user had dedicated the time and effort required to create an account in a particular source domain.

(2) *Activity*: Here, we consider a number of simple reputation signals reflecting the users' activities in Twitter and Facebook. For each Facebook or Twitter identity, we collected data about the *number of followers* (*friends* for Facebook) and the *age* of the account (number of months since creation). Age can be a powerful signal because identities who have survived (without getting suspended) for a long time in established source domains like Facebook or Twitter (withstanding numerous Sybil detection checks) might be less likely to misbehave in the target domain compared to newly created identities. On Twitter, we also collected data about the *number of expert lists* [21] where the identity appears and the *ratio of followers/followings*. Lastly, we consider the *age difference* between the creation of the Facebook or Twitter identity and the creation of the Pinterest identity. A small age difference might be indicative of a Sybil attack, since Sybil attackers are known to create accounts in bulk [42] within a short period of time. For example, they might create multiple accounts on Twitter or Facebook and immediately use them for creating accounts on Pinterest. Note that many of the above reputation signals have already been used in the literature to estimate the trustworthiness of identities in a single domain [16]. We now have the opportunity to investigate whether such signals have predictive value across domains as well.

**Trustworthiness signals:** To collect information about the trustworthiness of identities, we rely on information released by the source domains when they take action against an identity for misbehavior. This is a simple binary signal indicating whether an identity has been suspended or not, by the domain.[9] Intuitively, users who misbehave in one domain are more likely to misbehave in other domains. We check for account suspensions on Facebook and Twitter. In Twitter, out of the 132,535 matching identities, 8,614 were suspended, while on Facebook, out of the 968,230 matching identities, 36,132 were suspended (or deleted by the user).[10]

### *5.1.3 Ground truth for untrustworthy identities*

To evaluate the utility of our framework for reasoning about trustworthiness in the target domain, we identify a diverse set of untrustworthy identities in Pinterest that serves as valuable ground truth information (for trustworthiness). We collected the following four datasets of untrustworthy identities on Pinterest:

**Suspended identities:** The easiest way to obtain data about untrustworthy identities is to identify the identities suspended by Pinterest for violation of ToS. When we try to fetch the profile page of a suspended identity, Pinterest returns a 404 HTTP error message.

[7] Users have the option to hide the link to the two social source domains in their privacy settings. Thus, our approach has the limitation that we wrongly classify the source domain as an email domain.

[8] We drew the random sample of identities from a near complete Pinterest dataset collected as part of a previous study [50] in 2013 by authors of our paper.

[9] While today the only trustworthiness signal we have is account suspension and it is a binary signal, in general a trustworthiness signal could take any number of values.

[10] We cannot distinguish between suspended and deleted identities on Facebook.

We use this signal to identify suspended identities on Pinterest.[11] Out of the 1.7M Pinterest identities, we found that 74,549 have been suspended.

**Identities with black-market association:** In this dataset, we collect information about identities associated with a black-market service. Images posted by identities on Pinterest are called *pins*. Other identities can *repin* (or reshare) and also *like* (or express interest in) existing pins [11]. Today, there is a strong incentive to increase the popularity of content on social media sites. This has led to the emergence of a variety of black-market sites where one can buy services that help to artificially boost the popularity of their content. There are multiple black-market sites where Pinterest repins or likes can be fraudulently obtained [6, 7, 8]. We focus on a popular black-market site called addmefast.com. Each day, subscribers of this site receive a list of pins they should like or repin. We subscribed to the site and collected a list of 135 such pins in one day. On Pinterest, we then collected all identities that had liked or repinned two or more (to improve the likelihood of discovering untrustworthy identities) of the 135 collected pins. We gathered 1,706 such untrustworthy identities.

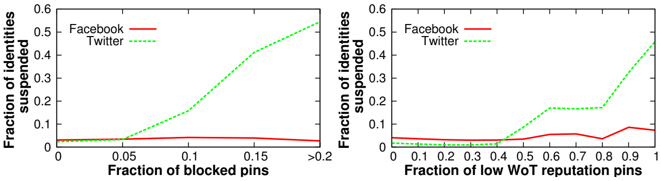
**Identities with pins that are blocked:** Each pin on Pinterest has an associated URL that redirects the user to the page hosting the image. Pinterest is known to block URLs to domains that redirect, or contain spammy, misleading, or inappropriate content, or otherwise violate its ToS. We refer to pins with blocked URLs as blocked pins. There are 724,672 Pinterest identities with at least one blocked pin, which includes 43% of all Pinterest identities. We assume that a vast majority of the random Pinterest identities are indeed trustworthy, and hence, we do not consider all identities that posted a single blocked pin to be untrustworthy. Instead, we assume that identities with higher fractions of blocked pins are more likely to be untrustworthy. For example, a small fraction (1%) of Pinterest identities have a vast majority (65%) of their pins blocked, and these identities are more likely to be untrustworthy.

**Identities with pins that have low reputation URLs:** We measure reputation of URLs associated with pins using the Web of Trust (WoT) [15] website; which computes the reputation of different websites based on various metrics, and input from the community. The WoT reputation scores vary between 0 (lowest reputation) and 100 (highest reputation), and WoT suggests that reputation scores less than 59 can be considered low. If the domain is very new or very unpopular, the WoT database might not have a score for that domain. We consider not having a WoT score also to be a signal for low reputation URLs because spammers are known to create new domains when their current domains get blacklisted. There are 1,083,951 such Pinterest identities that have at least 1 pin with a low WoT reputation score or no pin with a WoT score. Since this includes a significant fraction of all identities, in the same way as for identities with blocked pins, we assume that identities with higher fractions of low WoT reputation pins are more likely to be untrustworthy.

**Ethical concerns:** We believe that our data collection process does not raise ethical concerns because we have only collected publicly available data. Moreover, our data collection processes adhered to the API rate limits of the different domains. We consulted a local privacy lawyer, who confirmed that our research is in accordance with the Max Planck Society's Ethics Guidelines as well as with the applicable German data protection legislation (§28 BDSG).

| | Twitter | Facebook |
|---|---|---|
| Random | 4% | 3.4% |
| Suspended | 36.1% | 9.3% |
| Black market | 8.4% | 14.5% |

Table 2: Percentage of identities suspended in Twitter and Facebook corresponding to random, suspended and black market identities in Pinterest.



(a) Blocked pins

(b) Low WoT reputation pins

Figure 2: Fraction of Pinterest identities with different fractions of malicious pins that have their matching identities on Facebook or Twitter suspended.

## 5.2 Potential for inter-domain trust transfer

In this section, we study the potential for inter-domain trust transfer by investigating whether the previously introduced reputation signals and trustworthiness signals on Facebook and Twitter are indicative of trustworthiness on Pinterest (are able to discriminate between trustworthy and untrustworthy Pinterest identities). More precisely, we analyze whether a random set of Pinterest identities (a majority of which would be expected to be trustworthy) have different reputation or trustworthiness scores than a set of untrustworthy Pinterest identities.

Recall from §3.1.3 that to estimate the trustworthiness of identities, the target domain has to do a calibration step and check the correlation between signals from the source domain and trustworthiness in the target domain. The analysis we do in this section can be seen as the exploratory part of the calibration step. Later, in §5.3, we will show how we can actually leverage these signals together to curate identities on Pinterest.

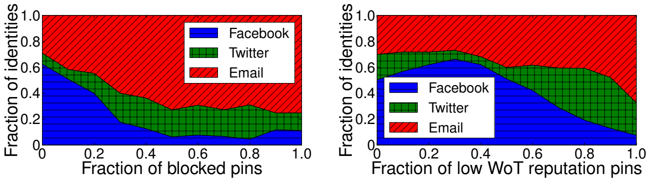
### 5.2.1 *Leveraging trustworthiness signals*

Intuitively, if a user has misbehaved in a domain, he is more likely to misbehave in other domains. We investigate whether the binary trustworthiness signal of whether an identity has been suspended on a source domain is indicative of trustworthiness of the matching identity on the target domain.

Table 2 shows the percentage of suspended matching identities on Twitter and Facebook for random, suspended and black-market Pinterest identities. We can see that suspended and black-market identities on Pinterest are 9 times (2 respectively) more likely to have their matching identities suspended on Twitter and 2 times (4 respectively) more likely to have their matching identities suspended on Facebook than random Pinterest identities.

For untrustworthy identities with malicious pins, Figure 2 plots how the fraction of suspended matching identities on Twitter and Facebook varies with an increasing fraction of malicious pins on Pinterest. The figure shows that as the fraction of malicious pins increases, the likelihood of matching identities on Twitter to be suspended increases. On Facebook, there is no apparent increasing trend potentially because Facebook is less aggressive at suspending identities than Twitter [41] (and instead focuses on removing malicious activities associated with the identities), or because these untrustworthy identities have not misbehaved sufficiently on Facebook. Thus, generally, suspension on the source domain can be indicative of trustworthiness on the target domain.

[11] One limitation of this approach is that Pinterest returns the same HTTP error message for identities that are deleted (by the users themselves).



(a) Blocked pins

(b) Low WoT reputation pins

Figure 3: Correlation between the source domain distribution and the fraction of malicious pins posted by Pinterest identities.

### 5.2.2 *Leveraging reputation signals*

**Domain of matching identity:** Not all source domains are equal: intuitively, using an identity from certain source domains may be more indicative of trustworthiness than using an identity from others. In this section, we study whether the user's choice of linking his Pinterest identities to either his email, Twitter or Facebook identity is indicative of the trustworthiness of his Pinterest identity.

More precisely, we analyze whether untrustworthy identities prefer to link their identities to different source domains compared to random identities. Table 1 shows the source domain distribution of the random, suspended, and black-market Pinterest identities. We can see that compared to random identities, untrustworthy identities are more likely to use email addresses to log in to Pinterest and less likely to use Facebook identities. 62% of suspended Pinterest identities use email addresses compared to 38% of random Pinterest identities; while 55% of random Pinterest identities choose to use Facebook identities, only about 25% of suspended Pinterest identities use Facebook identities.

For untrustworthy identities with malicious pins (blocked or low WoT reputation), we bin identities according to their fraction of malicious pins, and we check the source domain distribution in each bin as shown in Figures 3a and 3b.[12] We observe that identities with a higher fraction of malicious pins (blocked or low WoT reputation) are less likely to link their Facebook identities and more likely to link their email identities. Thus, untrustworthy identities prefer different source domains than trustworthy identities.

One explanation for this could be that attackers prefer source domains where it is easier to create identities. To verify our hypothesis, we checked whether creating an email identity is easier than creating an identity on Facebook. We tried to create multiple Facebook and Hotmail identities from a single IP address over a period of two days. We were able to create two identities on Facebook after which we were asked for phone verification, and eight identities on Hotmail (with no phone verification). We also manually collected price information for identities belonging to different domains from eight black-market services, which included fiverr.com and blackhatworld.com. The median price for Facebook identities was the highest ($0.51), followed by Twitter identities ($0.09), and with Hotmail identities having the lowest price ($0.01). This can explain why attackers prefer email addresses over Facebook identities to log in to Pinterest: they are easier to create and cheaper to buy.

**Activity information associated with matching identity:** In this section, we investigate whether reputation signals reflecting users' activities on Facebook and Twitter are indicative of trustworthiness on Pinterest. Figure 4 compares the CDFs of various reputation signals for random Pinterest identities and four different kinds of untrustworthy Pinterest identities. For identities that posted malicious pins, we use a threshold for the fraction of malicious pins posted, which corresponds to the top (1%) most untrustworthy identities. This corresponds to thresholds of 0.10 and 0.65, respectively for the fraction of blocked pins posted and for the fraction of low WoT reputation pins posted. While we do not show results for other threshold fractions (for malicious pins) due to space constraints, we briefly describe how the results vary for other threshold fractions.

Figure 4 shows that some reputation signals reflecting users' activities on Twitter and Facebook can indeed help in distinguishing random Pinterest identities from untrustworthy Pinterest identities. While different reputation signals are indicative to different extents, we can see that reputation signals based on age particularly seem to help distinguish random Pinterest identities from untrustworthy Pinterest identities.

For untrustworthy identities corresponding to higher threshold fractions of blocked pins than the threshold shown, the reputation signals from Twitter are more strongly indicative than signals from Facebook. For untrustworthy identities corresponding to higher threshold fractions of low WoT reputation pins, reputation signals from both Facebook and Twitter are strongly indicative.

**Takeaway:** Overall, our results indicate that there is potential for transferring reputation signals to benefit Pinterest, and highlights the importance of Pinterest calibrating the various reputation signals it receives, since they correlate to different extents with trustworthiness on Pinterest. The next section investigates the power of combining all reputation and trustworthiness signals together to curate identities on Pinterest.

## 5.3 Leveraging inter-domain trust transfer

The goal of this section is to show the benefits of inter-domain trust transfer in a practical scenario. For this, we consider the task of curating identities in the target domain Pinterest. We first describe the process of curating identities on Pinterest. We then estimate the extent to which inter-domain reputation and trustworthiness signals from Twitter and Facebook can help the task of curating identities on Pinterest.

### 5.3.1 *Identity curation*

A curated set of identities, as defined in §2, is a set of identities that the domain considers trustworthy. In practice, depending on how conservative the curated set is, it might contain some untrustworthy identities. To measure the quality of a curated set, we define two metrics: *purity* and *coverage*. We define the purity level as the fraction of curated identities that the domain believes are trustworthy in a curated set (e.g., a purity of 99% means that the domain believes 99% of the identities in a curated set will not misbehave). The coverage is the fraction of all identities in the system that are included in the curated set. There is a trade-off between purity and coverage: a curated set with high purity will contain fewer identities than a curated set with low purity.

In this section, we investigate whether Pinterest can achieve a higher coverage for a given target purity level if it augments the *intra-domain* reputation signals (signals that characterize the activities of identities on Pinterest: age of the account, number of followers and ratio of followers and followings) with *inter-domain* reputation and trustworthiness signals obtained through the trust transfer framework (signals that characterize the activities of matching identities on Twitter and Facebook: matching identity domain, suspension on Twitter, suspension on Facebook, number of Twitter followers, ratio of Twitter followers and followings, listed count

[12] We filter out identities that have insufficient activities (less than 20 pins) because they could accidentally have a high fraction of malicious pins. This filtering reduced the overall number of identities with blocked pins to 676,321 (from 724,672) and identities with pins with low WoT reputation to 819,677 (from 1,083,951).

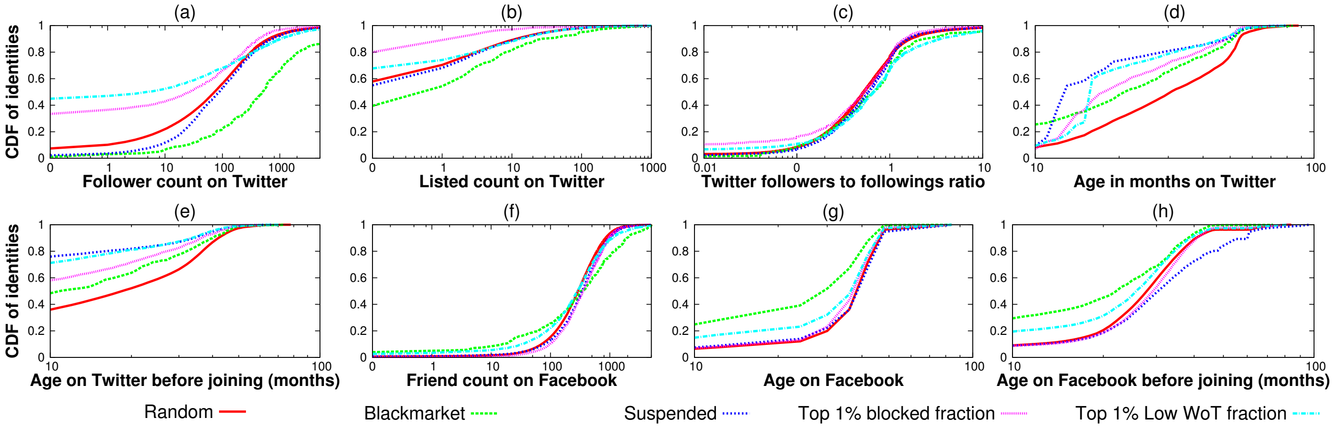


Figure 4: CDFs for several reputation signals reflecting users activities on Twitter and Facebook for different kinds of untrustworthy Pinterest identities vs. random Pinterest identities.

on Twitter, age of Twitter account, age on Twitter before joining Pinterest, number of friends on Facebook, age of the account on Facebook and age on Facebook before joining Pinterest).

**Methodology:** To obtain a curated set of Pinterest identities, we first rank all Pinterest identities according to the probability that they will misbehave in the future (the lower the probability, the higher the trustworthiness). We then pick the top $n$ trustworthy identities such that the purity remains within a target purity level.

We compute the probability of Pinterest identities to misbehave in the future in two ways: first, we only use intra-domain reputation signals, and then we use both intra-domain and inter-domain reputation signals. As we described in §2 and §3.1.3, we can use a binary classifier to compute the probability of Pinterest identities to misbehave in the future. For all Pinterest, Twitter, and Facebook identities, we first collected their reputation scores in July 2013 and further conducted an additional crawl 11 months later to determine which Pinterest identities misbehaved. We build two classifiers: one that takes as input intra-domain reputation scores and a second that takes both intra-domain and inter-domain reputation scores to predict misbehavior. For building the classifier, we consider all types of untrustworthy Pinterest identities together because we want the curated set of identities to be pure with respect to as many kinds of misbehavior as possible. For identities that post malicious pins, we consider the top 17,000 (which corresponds to the 1% most untrustworthy Pinterest identities) identities to be untrustworthy, as ranked by their fraction of malicious pins.

In total we have 107,372 untrustworthy identities (the negative examples) and slightly less than 1.6 million Pinterest identities that are not untrustworthy (the positive examples). We split the dataset in 60% for training and 40% for testing. We use logistic regression as our binary classifier. In order to handle class imbalance when training the classifier, we set the cost of misclassifying negative examples as positive proportionally higher.

### 5.3.2 Curation quality

We first evaluate the benefit in terms of coverage at a given target purity level. We then study what kinds of identities benefit the most from inter-domain reputation signals.

**Coverage:** To estimate and compare the coverage for a given target purity level, we apply the two classifiers on the test data and we rank all the Pinterest identities according to the probabilities returned by the classifier. Figure 5 shows the coverage vs. the level of purity when curating identities using all signals (intra-domain and inter-domain reputation signals), only Pinterest signals (only intra-domain reputation signals), and by simply randomly picking identities. The plot shows that by using both intra-domain and inter-domain reputation signals, we can obtain a curated set of identities that has a much larger coverage. For example, given a target purity level of 0.975 (fraction 0.025 of untrustworthy users), we get a coverage of about 20.2% using only Pinterest reputation signals, while we get a coverage of around 47.6% by additionally using reputation signals from Twitter and Facebook.

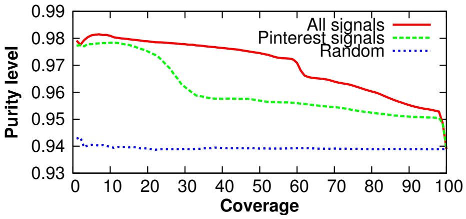


Figure 5: Coverage of the curated set for a given purity level for identities curated using all signals, only Pinterest signals, and identities curated at random.

**Which identities benefit the most?** To understand what kind of identities benefit the most from inter-domain reputation signals, we study the properties of identities that do not get curated using only intra-domain reputation signals, but that get curated using both intra-domain and inter-domain reputation signals – we call these identities the *additionally curated* identities. Figure 6 compares the CDF of the age of the accounts on Pinterest for identities that are curated using intra-domain reputation signals, and the additionally curated identities.[13] The plot shows that the additionally curated identities are young identities that have probably not had time to gain sufficient reputation on Pinterest, confirming our expectations. Figure 6 also shows that to curate identities using Pinterest reputation signals alone, we have to wait at least 15 months, but by exploiting inter-domain reputation signals, we can curate identities more than 10 months in advance. Thus, using inter-domain reputation signals allows us to curate *more* identities and enables us to do it *faster*.

[13] To curate identities, we set a target purity level of 0.975.

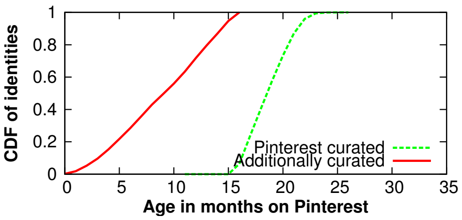


Figure 6: Age of curated Pinterest identities: identities curated using Pinterest reputation signals vs additionally curated identities using all signals.

## 6. RELATED WORK

There are two main lines of related work:

**Reasoning about trustworthiness of identities within online domains:** There has been a lot of work on reasoning about the trustworthiness of identities within a particular domain. Most of it focuses on detecting fake identities in a number of different ways including using the social network structure [19, 36, 39, 47, 48], using crowdsourcing from AMT workers [44]; or using supervised learning techniques [16, 30, 45] or unsupervised learning techniques [41, 43] that exploit different reputation signals. The primary drawback of these works is that identities have to build up a history of activities on these domains before the domains can reason about their trustworthiness. A few recent studies started to focus on how quickly we can detect malicious identities, and proposed techniques to detect malicious identities earlier on [28, 38]. However, many attackers can still evade the proposed defenses; therefore, domains still need to observe the activities of identities over a period of time to certify that they are trustworthy. Our framework, however, allows domains to identify trustworthy identities from the first day they join the system.

Finally, there have been a number of works that focus on identifying tampered crowd computations (e.g., content whose popularity or rating has been artificially boosted by a group of untrustworthy identities) [17, 18, 20, 42, 46]. At a high level, these techniques look at the characteristics or behavior of a group of participants in a crowd computation and check whether they exhibit anomalous patterns or known malicious patterns. Some techniques, in particular the one proposed by Viswanath et al. [42], can detect tampered crowd computations even if the participating identities have short activity histories on the domain. However, these techniques can only be used to reason about the trustworthiness of crowd computations, and cannot be used to reason about the trustworthiness of individual identities.

**Inter-domain knowledge and reputation transfer:** There are a couple of works that exploit transfer learning [34] techniques to handle the problem of data sparsity in the target domain by transferring knowledge (e.g., the exact activities of an identity, all the reviews posted about a business etc.) from other related source domains. Transferring knowledge has been proposed to enhance the efficiency of various tasks such as estimating the quality of merchants from reviews [31], predicting user behavior [51], and finding the right crowd workers in a crowdsourcing site to solve a given task [49]; however, it has not been studied in the context of transferring trust. All these works assume a collaborative environment where the target domain trusts the source domain, and they do not consider the risks of violating users' privacy.

Similar to our work, Grinshpoun et al. [25] proposed a high-level architecture to combine the reputation of entities, such as users or businesses, across different domains. However, they do not provide any experimental evaluation to validate their architecture. Moreover, their architecture is targeted to aggregate reputation and not to estimate the trustworthiness of identities.

Finally, several startups, such as Klout [2] or TrustCloud [3], are aggregating reputation of users' identities across domains to estimate the influence of users or to estimate their trustworthiness. However, it is not clear what data they aggregate, whether the data aggregation preserves the privacy of users, and how these metrics should be used by domains. Instead, we propose a flexible framework that does not incur a significant loss of privacy for users and where such reputation metrics could be calibrated to meet the goals of any target domain.

## 7. CONCLUDING DISCUSSION

This paper argues for a fundamental shift in the way we handle identities online: Because each domain operates independently of other domains, each domain has to create a separate identity for a user, and each user has to create a separate identity on each domain. Whereas it has been recognized that maintaining multiple identities rapidly becomes a maintenance nightmare for users, not enough attention has been paid to the fact that separate weak online identities creates problems for domains as well: each domain has to separately reason about the trustworthiness of a user, based on past actions within that domain. Young domains may not have enough data about identities or the resources needed to develop sophisticated Sybil detection systems; and, the actions of new identities need to be examined for a period of time before the identities can be deemed trustworthy. This may result in many legitimate (and sometimes even expert) users being unable to fully participate in the domain (e.g., they are not allowed to edit web documents, or may only be given limited content voting powers).

To address this problem, we propose a framework for transferring trust *across* domains. Our insight is that although users may be new on a particular domain, most honest users would have long histories and established reputations on other domains they have been using before. Using extensive data from Facebook, Twitter and Pinterest, we establish that inter-domain trust transfer is feasible, practical, and beneficial. On average, users have three years more history on Facebook and two years more on Twitter than on Pinterest. Our analysis shows that by using the reputation established by the users on these websites through their long history of past actions and recalibrating such signals to its own context, a younger domain such as Pinterest can whitelist users as trustworthy (with a probability of misbehavior lower than 2.5%) with approximately one year less history on Pinterest itself.

## 8. ACKNOWLEDGEMENTS

This work was partially supported by the 'Space for Sharing (S4S)' project on empathy and trust, with funding from the cross-council programmes for Connected Communities (led by the AHRC), the Digital Economy (led by the EPSRC) and the Partnership for Conflict, Crime and Security Research (led by the ESRC), in partnership with the Defence Science and Technology Laboratory (Dstl) and the Centre for the Protection of National Infrastructure (CPNI).

## 9. REFERENCES

[1] http://oauth.net.

[2] http://klout.com/home.

[3] https://trustcloud.com.

[4] Advogato trust metric.
http://www.advogato.org/trust-metric.html.

[5] Black-market for old twitter accounts.
http://buybulkaccounts.blogspot.de/p/5-years-old.html.

[6] Black-market website. http://addmefast.com.

[7] Black-market website. http://www.purchasesocial.com.

[8] Black-market website. https://socioblend.com.

[9] Facebook anonymous login. http://newsroom.fb.com/news/2014/04/f8-introducing-anonymous-login-and-an-updated-facebook-login.

[10] Facebook login. https://developers.facebook.com/docs/facebook-login/v2.3.

[11] Pinterest introductory guide. https://help.pinterest.com/en/guide/all-about-pinterest.

[12] Reddit faq. http://www.reddit.com/r/help/wiki/faq.

[13] Social login market. http://www.gigya.com/blog/the-landscape-of-social-login-facebook-makes-a-comeback.

[14] Tech blog post about facebook integration. http://royal.pingdom.com/2012/06/18/how-many-sites-have-facebook-integration-youd-be-surprised.

[15] Web of trust. https://www.mywot.com.

[16] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *CEAS'10*.

[17] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *WWW'13*.

[18] Q. Cao, X. Yang, J. Yu, and C. Palow. Uncovering large groups of active malicious accounts in online social networks. In *ACM CCS'14*.

[19] G. Danezis and P. Mittal. SybilInfer: Detecting Sybil Nodes Using Social Networks. In *NDSS'09*.

[20] S. Feng, L. Xing, A. Gogar, and Y. Choi. Distributional footprints of deceptive product reviews. In *AAAI ICWSM'12*.

[21] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: Crowdsourcing search for topic experts in microblogs. In *ACM SIGIR'12*.

[22] S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly, and K. Gummadi. On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In *ACM CIKM'13*.

[23] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi. On the reliability of profile matching across large online social networks. In *ACM KDD'15*.

[24] C. Grier, K. Thomas, V. Paxson, and C. M. Zhang. @spam: the underground on 140 characters or less. In *ACM CCS'10*.

[25] T. Grinshpoun, N. Gal-Oz, A. Meisels, and E. Gudes. Ccr: A model for sharing reputation knowledge across virtual communities. In *WI-IAT'09*.

[26] R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Preventing private information inference attacks on social networks. *IEEE Trans. on Knowl. and Data Eng.*, 2013.

[27] A. M. Kakhki, C. Kliman-Silver, and A. Mislove. Iolaus: Securing online content rating systems. In *WWW'13*.

[28] A. Leontjeva, M. Goldszmidt, Y. Xie, F. Yu, and M. Abadi. Early security classification of skype users via machine learning. In *AISec'13*.

[29] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE ICDE'07*.

[30] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *ACM CIKM'10*.

[31] M. McGlohon, N. S. Glance, and Z. Reiter. Star quality: Aggregating reviews to rank products and merchants. In *AAAI ICWSM'10*.

[32] A. Mislove, A. Post, K. P. Gummadi, and P. Druschel. Ostra: Leveraging trust to thwart unwanted communication. In *NSDI'08*.

[33] M. Mondal, B. Viswanath, A. Clement, P. Druschel, K. P. Gummadi, A. Mislove, and A. Post. Defending against large-scale crawls in online social networks. In *ACM CoNEXT'12*.

[34] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 2010.

[35] A. Post, V. Shah, and A. Mislove. Bazaar: Strengthening user reputations in online marketplaces. In *NSDI'11*.

[36] D. Quercia and S. Hailes. Sybil Attacks Against Mobile Users: Friends and Foes to the Rescue. In *IEEE INFOCOM'10*.

[37] T. Stein, E. Chen, and K. Mangla. Facebook immune system. In *SNS'11*.

[38] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *USENIX Security'13*.

[39] N. Tran, J. Li, L. Subramanian, and S. S. Chow. Optimal Sybil-resilient Node Admission Control. In *IEEE INFOCOM'11*.

[40] N. Tran, B. Min, J. Li, and L. Subramanian. Sybil-resilient online content voting. In *USENIX NSDI '09*.

[41] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In *USENIX Security'14*.

[42] B. Viswanath, M. A. Bashir, M. B. Zafar, S. Bouget, S. Guha, K. P. Gummadi, A. Kate, and A. Mislove. Strength in numbers: Robust tamper detection in crowd computations. In *ACM COSN'15*.

[43] 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 *USENIX Security'14*.

[44] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, and B. Y. Zhao. Social turing tests: Crowdsourcing sybil detection. In *NDSS'13*.

[45] G. Wang, T. Wang, H. Zheng, and B. Y. Zhao. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *USENIX Security'14*.

[46] G. Wu, D. Greene, B. Smyth, and P. Cunningham. Distortion as a validation criterion in the identification of suspicious reviews. In *SOMA'10*.

[47] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. SybilLimit: A Near-optimal Social Network Defense Against Sybil Attacks. In *IEEE S&P'08*.

[48] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard: Defending Against Sybil Attacks via Social Networks. In *ACM SIGCOMM'06*.

[49] Z. Zhao, J. Cheng, F. Wei, M. Zhou, W. Ng, and Y. Wu. Socialtransfer: Transferring social knowledge for cold-start crowdsourcing. In *ACM CIKM'14*.

[50] C. Zhong, M. Salehi, S. Shah, M. Cobzarenco, N. Sastry, and M. Cha. Social bootstrapping: how pinterest and last. fm social communities benefit by borrowing links from facebook. In *WWW'14*.

[51] E. Zhong, W. Fan, J. Wang, L. Xiao, and Y. Li. Comsoc: Adaptive transfer of user behaviors over composite social network. In *ACM KDD'12*.