How much I can Rely on You: Measuring Trustworthiness of a Twitter User

Rajkumar Das, Gour Karmakar, *Member, IEEE,* and Joarder Kamruzzaman, *Senior Member, IEEE*

Abstract—Trustworthiness in an online environment is essential because individuals and organizations can easily be misled by false and malicious information receiving from untrustworthy users. Though existing methods assess users' trustworthiness by exploiting Twitter account properties, their efficacy is inadequate because of Twitter's restriction on profile and tweet size, the existence of missing or insufficient profiles, and ease to create fake accounts or relationships to pretend as trustworthy. In this paper, we present a holistic approach by exploiting ideas perceived from real-world organizations for trust estimation along with available Twitter information. Users' trustworthiness is determined by considering their credentials, recommendation from referees and the quality of the information in their Twitter accounts and tweets. We establish the feasibility of our approach analytically and further devise a multi-objective cost function for the A* search to find a quasi-optimal path between the trust evaluator and the user whose trustworthiness is being evaluated. We also propose an incentive mechanism to increase user participation in the trust evaluation process, and a threat model and trustworthiness measure of referees to thwart the possibility of providing an untruthful recommendation to inflate one's trustworthiness. The efficacy of our proposed approach is validated through experiments using Twitter data and extensive simulation in various scenarios.

Index Terms—Social Networks, Security, Trustworthiness, Reliable Communication.

1 Introduction

With 310 million active users per month and 500 million tweets per day, Twitter is one of the largest online social network platforms. 69% of the purchases made from small and medium-sized businesses (SMB) are because of the information customers come across in Twitter, and over 80% of their customer service requests are received through this channel [1]. Short but targeted and immediate customer reviews and feedback from large Twitter users enable businesses and organizations to evaluate their products and services within a short time at almost no cost. Consequently, this allows organizations to improve customer experience and meet their demands promptly. Twitter is also a strong vehicle to express and crystallize public opinion on vital socio-political issues as demonstrated in the recent events like Brexit and Greece economic crisis. Therefore, Twitter content analysis is one of the most critical determinants for the success of businesses as well as government and non-government organizations. Another important application of Twitter is that over 63% of its users consider the platform a source of information for events and issues [2]. Accordingly, Twitter emerges as one of the fastest ways to disseminate information. The information is used by users for making decisions on socio-economic issues.

Like other publicly accessible online environments, Twitter is vulnerable to security concerns, which necessitates evaluating the trustworthiness of users and their generated tweets individually. A group of customers or a rival company can generate false reviews to destroy the economic strength and goodwill of an organization. On the other hand, a consumer can be deceived in making an online purchase based on biased Twitter reviews. Similarly, Twitter is also a source of misinformation deliberately propagated through compromised and untrustworthy users [3]. Therefore, it is utmost important to verify the trustworthiness of (i) a Twitter user and (ii) information generated from the user's Twitter account.

The existing research proposed methods mostly based on machine learning techniques to determine trustworthiness of users [7], [8] and their tweets [6] [4] by exploiting several factors derived from Twitter. The efficacy of these learning based techniques is very much limited to the selection and accuracy of their training sets. Most importantly, the main drawback of the existing works is that they are vulnerable to be misled by the users who create fake accounts and relationships to present them as trustworthy. The online predators work in groups by making strongly connected networks among them and can easily manipulate the Twitter statistics to falsely appearing as trustworthy. Anyone can pretend to be a person with good credential by listing false information in profile; there is no way to verify directly. Consequently, we cannot rely on any information from an account which is not verified as trustworthy. We refer to this as the 'trustworthiness of a node' to differentiate it from the 'trustworthiness of data' [32] generated from that account. The problem of verifying a node as trustworthy is further compounded by the fact that around 50% Twitter users have no or limited profile information. These limit the efficacy of the existing methods for measuring trustworthiness. In real-world, a person can be identified as trustworthy only after being verified with

- *R. Das is a Developer, Data Engineer in Information Technology Service Department, Federation University Australia. E-mail: r.das@federation.edu.au; rajkumardash05@yahoo.com*
- *G. Karmakar and J. Kamruzzaman are with School of Science, Engineering and Information Technology, Federation University Australia.*

Manuscript received Jun, 2017

their real identities. This creates a significant research gap between online environments like Twitter and real-world systems in measuring the trustworthiness of a user.

In real-world systems, users are asked to provide their credentials through official identity, based on which they are verified as trustworthy or not. The 100-points identity verification process implemented by the Australian government [12] is a classic example of such system which is adopted by Banks and other financial institutions, real estate agencies and many others. Here, credentials are asked from a number of categories and each category is assigned with some points. A person needs to provide at least 100 points to verify her identity. However, it is not possible to implement such a method in an online environment like Twitter, as people will not provide their credentials online for privacy and security reasons. Therefore, it raises a very potential research question: is it possible to verify a user's identity in the online environment to measure its trustworthiness? If possible, what would be such credentials and how the parties could transfer them in Twitter privately and verify? Any credential which can represent a user's identity in online is referred to as web credential. To address the above research question, for the first time, this paper proposes a method to measure trustworthiness of Twitter users using web credentials. Please note that our mechanism allows only true Twitter users towards trustworthiness evaluation and thus, fake and bot accounts are excluded from all considerations. We shed more lights on this in later sections.

Verifying a user's trustworthiness through web credentials does not provide a complete solution as some users might have no or an insufficient number of credentials despite being trustworthy in their real-world interactions. To address this issue using the most popular and widely adopted approach used in the real-world verification process, we incorporate a further verification process by requesting users to nominate a set of possible referees and users' trustworthiness is verified based on their recommendations. Since online referees may be fake users or generate false recommendations, in this paper, we also introduce the trustworthiness measure of referees from their Twitter accounts and recommendation scores. To the best of our knowledge, evaluating users' trustworthiness using direct recommendation process has not been investigated before for Twitter. Once a user is verified, we evaluate the trustworthiness of tweets of a user in an innovative way.

The success of the proposed trust model depends on (i) the establishment of a communication path in Twitter network between the trust evaluator and a user whose trustworthiness is to be evaluated, (ii) the possibility of the user to provide her web credentials on request and (iii) the availability of trustworthy recommendation from the referees. To address the first issue, considering the characteristics of Twitter users and the underlying network, we analytically ensure the possibility of such a communication path between any two Twitter users within a reasonable delay-bound. Furthermore, we devise a multi-objective cost function for the A* search to find the most reliable path within less amount of delay. For the second issue, we develop an incentive mechanism to encourage users to participate in the trust evaluation process. Finally, we propose a threat model to thwart the trust inflation process by a strongly connected group of users who review each other positively. Especially, the threat model addresses the Sybil attack [26] with respect to our recommendation process and eliminates those referees who pose a higher possibility of being Sybil nodes, attempting to provide fake and inflated recommendations.

In a nutshell, in this paper, we propose a holistic approach to solve the existing research gap by incorporating ideas perceived from physical organizations for trust estimation along with available Twitter information. Our model has two-fold potential benefits. First, verification of a Twitter user by online credentials and referees' recommendation reports will improve the quality of trustworthiness measures. Second, the concept can be used, with minor modification, in other online platforms like Facebook, blogs, and other social networks. Our major contributions are the followings:

- Trustworthiness of a Twitter user is determined by employing the user's web credential and devising an online recommendation system. Credentials and recommendations are sought and exchanged in a way that preserves privacy.
- Trustworthiness of tweets of a user is determined by measuring the quality of information it presents with respect to the information-space on Twitter.
- An incentive mechanism to facilitate communication and a threat model to thwart Sybil attack are proposed.
- Finally, the feasibility and performance of the proposed model are investigated analytically, and an innovative cost function for A* search to find quasi-optimal paths to communicate among nodes within an acceptable time is proposed. The approach is extensively verified using real Twitter and simulated data.

2 Literature Review

Trustworthiness is an important factor to establish a safe, reliable and sustainable social platform [15], where the awareness and guarantee of safety are ensured through developing and strengthening trustworthiness among its users [16]. Consequently, assessing the trustworthiness of an online social network user draws substantial attention from the relevant research community. Despite this, trust evaluation is still poorly understood in an online social network environment and thus, needs extensive investigation in this area [17] [16]. Already being a multifaceted problem, the different characteristics of different online social networks demand platform-specific solutions by considering their distinguished features and corresponding vulnerabilities. In this section, we mainly focus on measuring the trustworthiness of users on Twitter. The existing approaches can be categorized into two different but closely related streams, namely (i) assessing user trustworthiness and (ii) detecting a spam user account.

To address the first issue, the existing research proposed methods to determine the trustworthiness of users [7], [8] and their tweets [6] [4] by exploiting several factors derived from Twitter. It includes (i) the graph based factors that exploit the structural and contextual properties of users such as the number of followers, friends and list-memberships and (ii) factors extracted from the characteristics of the user

generated tweets including the number of times a tweet has been re-tweeted or liked by others and the quality of the links included in a tweet.

Yu et al. built a hybrid graph with users, tweets and topics as nodes to represent the trust relationship among the nodes, and conduct a semi-supervised learning algorithm inspired by the Label Propagation Algorithm to detect untrusted users [18]. Their main drawback is that they need a seed set of known trusted users, which may be challenging to get all the time, especially a user may change her behaviour over time or depending on a topic. Similarly, in [7], Zou et al. proposed a graphical probabilistic model based on Pairwise Markov Random Field (PMRF) considering both user features and social relationships, and employ the Belief Propagation algorithm for trust inference. However, trustworthiness is highly subjective [17] that might not converge to a single value for a user across the social network.

Salih et al. in [8] formulate a domain-specific trustworthiness metric incorporating several attributes extracted from content and user analysis and considering temporal factors. However, the main concern is that, instead of comprehending the user's trustworthiness, the attributes mainly relate to the influence of a user on Twitter-sphere on a particular topic. Rather than relying on a training set, in [6], Twitter is treated as a news channel, and the credibility of user tweets is verified against the information from contextually similar trustworthy news sources. The authors further devised trust propagation rules by combining contextual and social networking information to estimate users' trustworthiness. The major drawback is that, for any past event, the trustworthiness computation heavily relies on the availability of event-related posts generated by credible news outlets. Any error generated in this step will be propagated in users' trustworthiness estimation.

On the other hand, a series of trust propagation and inference models is proposed where the connected peers of the user assign trust scores through the networks [5] [11] [9]. Although propagative nature of trust can provide a rough estimation of an unknown user, the subjectivity in trust necessitates an individual evaluation of each user's trustworthiness, especially in the information-centric network like Twitter.

Trustworthiness is also related with whether the corresponding Twitter account is a spammer and disseminates spam. Several existing models deal with the spam/spammer detection on Twitter. In [19], Benevenuto et al. recognized a number of characteristics relating to spammers' social behaviour and their tweet content and applied machine learning techniques using these attributes to classify users as either spammers or non-spammers. The characteristics of tweets include (i) fraction of tweets containing URLs, (ii) fraction of tweets that contain spam words, and the average number of words that are hash-tags on the tweet. Twenty-three user attributes are considered, including account age, the existence of spam words on the user's screen-name, the minimum, maximum, average, and median of the time between tweets and the number of tweets posted per day. Similarly, in [21], Wang proposed a spammer detection strategy using both graph- and content-based features, such as the number of followers and friends, user reputation, duplicate tweets, number of HTTP links, and number of replies and mentions.

Instead of using account features, Song et al. proposed a spam tweet filtering method in [20] by using relation features, namely the distance and connectivity between the sender and receiver of a tweet. Amleshwaram et al. [23] introduced bait-oriented features that capture the techniques used by spammers to grab victims' attention to lure them into clicking malicious links. Examples include the number of unique mentions, unsolicited mentions, hijacking trends and others. They also clustered the spammers into prevalent spam campaign groups on Twitter.

Lin et al. [24] analyzed the effectiveness of the common features used in prior studies for detecting long surviving Twitter spammer account, and their study yields two very simple yet, effective features, namely (i) the URL rate and (ii) the interaction rate. Likewise, very recently, Herzallah et al. [22] conducted a comprehensive analysis of all the features that were investigated for spam detection in the literature to determine the most effective ones. The most effective features in detecting spam Twitter account are reported, and the prediction performance of well-known classifiers (e.g., Naive Bayes, support vector machines, neural networks, Decision Trees, k-Nearest Neighbour) using these features is evaluated. They obtained a very high prediction accuracy of 99% in detecting spam accounts, though they experimented with small dataset having balanced spammer and non-spammer class size. In reality, class sizes are unequal and whether similar high accuracy is achievable with large scale real dataset is yet to be tested.

The main problem of these machine learning based techniques is that its detection performance mostly depends on the appropriateness of the training set used in the learning phase. On Twitter, users can easily manipulate the features used for learning, and the current machine learning based approaches in literature do not embed a mechanism to verify and filter out this false/manufactured information provided by untrustworthy Twitter users. For the first time, the proposed model addresses these issues by introducing the concepts of real-world identity verification process along with referees' recommendation to evaluate user trustworthiness in the context of Twitter-like public online environment. Since our method does not require a high number of web-credentials and recommendations, the data is either very small in size or low dimensional indicating a machine learning based approach is not suitable here.

Evaluating trust is also essential in other systems, such as a system for recommending web services to users [38] [39]. In [38], Wang et al. propose a system for web service recommendation to users. In recommending a service to a user, it takes into consideration the trust relationship between a user and the recommenders who have earlier used those services. It employs the well-reputed beta trust model to evaluate both direct and indirect trust between a user and recommenders before taking the recommendations for web-services. On the other hand, the model Personalized Service Recommendation Based on Trust Relationship (PSRTR) [39] proposed by Tian et al. compute user trust based on their interest background, evaluation tendency and recommenda-

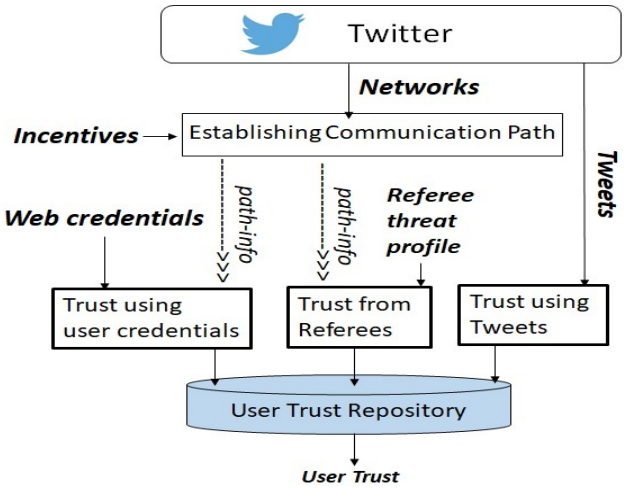


Fig. 1. Conceptual framework of a Twitter user trustworthiness measurement.

tion effect. Our model that applies to Twitter is completely different from the web-service recommendation platform. The innovation comes from the process of establishing a communication path among users, collecting web credentials from them and attaining recommendation reports from trustworthy referees. It is a new concept first introduced for Twitter Platform, and the methods for web service recommendation are not directly applicable to Twitter.

3 Conceptual Framework

The conceptual framework depicted in Fig. 1 represents the core components of our proposed framework. Here, the three most fundamental modules of the framework are: (i) communication path establishment, (ii) verifying trustworthiness using web credentials and referee recommendations and (iii) measuring trustworthiness from user tweets.

First, a user (trust evaluator) establishes a communication path with another user (destination) by exploiting the Twitter network's follower-followee relationship. Incentives encourage users to participate in the path establishment process. Using this path, web credentials are obtained from the destination user and her trustworthiness is evaluated.

Second, a list of referees is also sought after using this communication channel. Further communication paths between the trust evaluator and the referees are established to obtain the referees' report and using them user's trustworthiness is further evaluated. Referee threat profile is built to thwart the Sybil attacks in the recommendation process and to pick the most reliable referees.

Third, a user's tweets are used to further evaluate that user's trustworthiness. User Trust Repository is the trust evaluator's personal repository to store the three trust scores that can be referred to when interacting with any user on Twitter.

In the following section, we briefly describe the three fundamental components of our framework.

3.1 Communication Path Establishment

Twitter does not support private communication path between two arbitrary users which is necessary for them to exchange web credentials in privacy-preserving way. This is because (i) traditional tweets are public and visible to all and (ii) sending a private message is only limited to one-hop communication from a followee to its follower. To work around this restriction, as depicted in Fig. 2(a), a communication path between any two Twitter users (A, Z) needs to be established by a chain of one-hop followee-follower interactions (A, B, X and Z). Please note that followers connect to their followee, whereas the message is sent in the opposite direction. Even to reach a followee (F), a follower (A) has to establish such a path using one of the followers as an intermediate node (C).

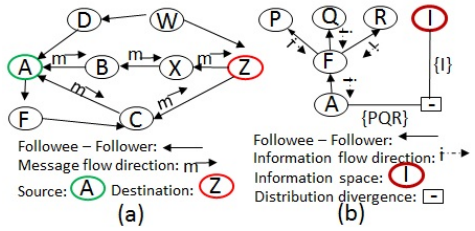


Fig. 2. Illustrations of (a) communication paths and (b) information quality measurement. Here, the *source* is the trust evaluator, and the destination is the one whose trustworthiness is being measured. The source initiates communication regarding credentials and referee reports, and the destination responds.

Using the path $A \leftarrow B \leftarrow X \leftarrow Z$, A can send messages to Z, but will be visible to the intermediate nodes, thus violating privacy of communication. Moreover, Z cannot send any message using this path due to its unidirectional flow of information, and need to establish another path from Z to A if any. Thanks to Twitter feature "group conversation", a bidirectional private communication between A and Z can be established. With respect to Fig. 2(a), on Twitter, node A can start a group conversation by including any combination of followers B, C, D in a private message. Being a part of the conversation, any node can add their followers to the list. Following this chain of followee-to-follower interactions, the destination node Z will eventually be added to that conversation by any of her followees (X or C) previously added to the group. After Z joins the conversation, the message originator A can remove others, thus having a communication group of only two users, A and Z, effectively communicating directly to Z.

The participation of intermediate users in the communication path establishment and of destination users to exchange information need incentives. An incentive mechanism to facilitate participation is discussed in Section 9.

Please note that, on Twitter, there always exists a communication path between two users A, Z consisting of a few intermediate nodes. Twitter network has very good connectivity and a user can reach 96% of all Twitter users by visiting the friends of friends at six levels (https://sysomos.com/inside-twitter/twitter-friendship-data/), and 99.99% of all users at eight levels.

3.2 Verifying Identity - Web Credentials and Referees' Recommendation Report

A person is adjudged trustworthy by financial organisations (e.g., bank, government organisation) globally through verifying their identity documents. Similarly, the online presence of a person, such as a Twitter account can also be verified as trustworthy using her online identity - her

public manifestations on the Internet. Examples of such web presences include, but are not limited to, a person's presence in various social media (LinkedIn, Facebook and social question-answer communities), personal and institutional websites, Youtube channels and shared videos, personal blogs, membership to scientific communities like Google-Scholar, authorship of news items published in a newspaper, and so on. We refer them as "Web Credentials" in this paper, and use them to evaluate a Twitter user's trustworthiness.

Referees recommendation is also a well-practiced method for verifying facts and identity in real life, and can be adopted in our model to verify a user trustworthiness. A set of referees (Twitter accounts) can evaluate a user on a set of online identity criteria, and the evaluation results comprise the recommendation report.

3.3 User Tweets - Information Quality Measurement

On twitter, a user's (refer to A in Fig. 2(b)) news-feed is populated by tweets from the Twitter accounts that the user follows (here F). Similarly, F obtains information from P, Q and R. Thus, the followee F acts as an information gateway to A. The quality of this information needs to be measured to determine its trustworthiness. For this, we need to extract the information from a user's tweets, and assess its quality according to its similarity to the majority-supported information in the Twitter-space.

Content analysis tools can be used to extract information from a user's tweets on a particular topic ($\{PQR\}$ in Fig. 2(b)). A set of keywords and hash-tags can be generated from a user's tweets and be used to extract tweets from the Twitter-space on that topic. This represents the Twitter-space information (I in in Fig. 2(b)). The difference between the Twitter-space information and that of the user can be used to assess the quality of her tweets.

4 Model Description

In our model, we represent the Twitter relationship structure using a directed graph $G = (V, E)$. For any two Twitter users $u, v \in V$, the directed edge $(u, v) \in E$ denotes that u (follower) is following v (followee). $\mathcal{M}_u$ and $\mathfrak{F}_u$ are the sets of followers of u and the followees that u follows, respectively. $\mathcal{W}_u$ denotes the set of web credentials comprising the online identity of u. Now, Θ_u represents the set of tweets originating from or retweeted by u on a particular topic, while Θ symbolises the tweets generated by all the users in the Twitter-space on that topic. Table 1 lists and briefly explains the symbols used in this section for a quick reference.

TABLE 1
Important symbols used in the paper and their meaning.

Symbols	Description
A, Z	Twitter users - trust evaluator (communication source) and user whose trust is evaluated (destination), respectively.
$\mathcal{M}_u$, $\mathfrak{F}_u$	The set of followers and followees for any Twitter user, respectively.
$\mathcal{W}_u$, Θ_u	The set of web credentials of a Twitter user and the Tweets generated from that account, respectively.
$\Psi_Z^{\mathcal{W}}$, $\Psi_Z^{\mathcal{R}}$, Ψ_Z^{Θ}	Trust of user Z computed by user A based on web credentials, referees' recommendation and generated tweets, respectively.
$\nu_A(\mathcal{W}_Z(k))$	User A's evaluation on the k^{th} web credential of user Z.
$\mathcal{C}$, Ω	The number of categories web credentials can be divided into and total number of points a user needs to be verified as trustworthy, respectively.
$\mathcal{R}_Z$, Φ_Z	The set of referee for user Z and the evaluation criteria which the referees will recommend Z on, respectively
$\psi_{Z,R\Phi}^{A}$, τ_r	The recommendation report - recommendation received from all referees $\mathcal{R}_Z$ and on all identity criteria Φ_Z, and the trustworthiness of any referee $r \in \mathcal{R}_Z$, respectively.

4.1 Evaluating Trustworthiness from Web Credentials

To measure the trustworthiness of a user by her web credentials, we adopt the personal identification system (known as the "100-point-check") adopted by the Australian Government. In that system, every identification document is assigned a value (points) which can be further divided into primary, secondary and tertiary categories. For a user's identity to be considered verified by the system, she must provide documents totalling a certain number of points (e.g., 100 points). However, the number of points assigned to a particular document, the number of document categories, and the total number of points required are highly adaptable based on the specific needs of the implementing organisation. The web presences of a person can likewise be divided into multiple categories based on how informative they are and their usefulness as credentials. For example, a person's LinkedIn profile carries more reliable information than her Facebook profile. Likewise, information in web profiles in domains such as .gov and .edu, and in corporate employee profiles, are far more reliable than those presented in personal website or blog. Consequently, they can be grouped in different categories. Consider that we can divide user credentials into $\mathcal{C}$ number of categories.

Now, consider that a Twitter user A is evaluating the trustworthiness ($\Psi_Z^{\mathcal{W}}$) of another user Z based on the latter user's web credentials $\mathcal{W}_Z$. If Ω represents the total credential value required for a user to be verified, her trustworthiness $\Psi_Z^{\mathcal{W}}$ can be computed using Eq. (1).

$$\Psi_Z^{\mathcal{W}} = \sum_{c \in \mathcal{C}} \max_{\mathcal{W}_Z(k) \in \mathcal{W}_{Z,A}^c} \frac{\nu_A(\mathcal{W}_Z(k))}{\Omega} \quad (1)$$

Here, $\mathcal{W}_{Z,A}^c \subseteq \mathcal{W}_Z$ represents the subset of credentials of Z in category c that is available to A when computing the trustworthiness. Depending on the value of Ω, the number of credential categories $\mathcal{C}$ and the value of any credential $\mathcal{W}_Z(k)$ of Z, $\nu_A(\mathcal{W}_Z(k))$ is assigned by A. Unlike physical identity documents, web credentials can be perceived differently by different individuals. The perceived value of a web credential can depend on the quality of the corresponding web source. Page rank is a good indication of such quality. In addition, the value of a credential also reduces any confusion in unambiguously identifying a user from her web credentials. Consequently, $\nu_A(\mathcal{W}_Z(k)) = p(\mathcal{W}_Z(k)) \times \pi(\mathcal{W}_Z(k))$ is an estimation of the value of a credential, where $\pi(\mathcal{W}_Z(k))$ represents the page rank of the website corresponding to the credential $\mathcal{W}_Z(k)$, and $p(\mathcal{W}_Z(k)) \in [0, 1]$ is the probability that it is indeed a web presence of Z.

4.1.1 Dealing with Number of Credentials

It can be argued that a user with more web credentials is not necessarily more trusted than someone with fewer credentials. Dividing credentials into multiple categories

and assigning different scores to them address this issue. Having fewer credentials in high valued categories is sufficient for a user to be considered trustworthy. Depending on the requirements and situation, users who do not have highly-reputed credentials can use credentials from low-valued categories. However, limiting the value of $\mathcal{C}$ helps the proposed model to negotiate the situation where users may generate many fake accounts on public web services, while still having low or insufficient scores as per Eq. (1). Moreover, it is not possible for a user to create fake credentials in the most-reputed category, such as profiles in government or educational websites.

4.1.2 Dealing with Fake and Bot Accounts

Another problem the proposed model needs to deal with is the identification of fake web credentials provided by users whose trustworthiness is being evaluated. Usually, a real Twitter user creates legitimate accounts on other online public media, whereas a fake Twitter user remains fake in all her public online appearances. Usually, 10% of Twitter users are fake [37]. To deal with these fake Twitter accounts, a user is tested as real or fake before her trustworthiness is evaluated. There are several existing methods [28], [29] to determine fake account on Twitter with good accuracy, and our model leverages their outcomes. It is unnecessary to evaluate the trustworthiness of a user who is deemed highly fake by those methods. Therefore, our model only consists of those users that pass the fake user detection process.

In some occasions, a real Twitter user might provide fake credentials to the evaluator. In that case, the credentials must be evaluated as real or not. Likewise on Twitter, a fake user can be identified in other online platforms [29], such as LinkedIn and Facebook. However, to minimise the overhead, we can employ this step only when the information provided by the destination and the referee differs, or when there is no high quality referee available for that destination. Moreover, between 9% and 15% of active Twitter users are bots [36] whose activities are controlled by computer programs. Since our framework relies on answering the messages that are sent to the intermediaries or the destinations, it can detect these bot accounts. This is possible because bots usually do not reply to messages.

4.2 Evaluating Trustworthiness from Referee Recommendations

In addition to the web credentials, A can request Z to provide a list of referees $\mathcal{R}_Z$, and use their recommendations to better determine the trustworthiness of Z. Like the identity verification system, the referees are asked to provide reports on Z using multiple identity criteria, which are then converted into a combined score representing trustworthiness. User Z might choose her referee from her followers and followees, thus $\mathcal{R}_Z \in \mathfrak{F}_Z \cup \mathcal{M}_Z$. As they are just a one-hop distance from her, this increases the availability of recommendations. However, since the referees are not limited to relationships within Twitter, a user can select referees who might not have a Twitter account. For this, Z needs to provide A with a way to communicate with those referees (e.g., email addresses). Nevertheless, followers of a Twitter account would be the best choice as referee, because they have the best knowledge on that user.

If Φ_Z^A represents a set of identity criteria on whose basis a referee is evaluating Z, the recommendation report received by trust evaluator A from referees $\mathcal{R}_Z^A$ on the target user Z can be represented as $\psi_{Z,R\Phi}^A = [\psi_{Z,r\phi}^A]$, where $r \in \mathcal{R}_Z^A$ and $\phi \in \Phi_Z^A$. Thus, $\psi_{Z,r\phi}^A$ denotes the recommendation on Z received by A from referee r on an identity criteria ϕ. The trustworthiness $\Psi_Z^{\mathcal{R}}$ of Z using this recommendation report is estimated as:

$$\Psi_Z^{\mathcal{R}} = \mathrm{Mo}(\psi_{Z,R\Phi}^A) = \underset{\forall r \in \mathcal{R}_Z^A, \forall \phi \in \Phi_Z^A}{\mathrm{Mo}}([\psi_{Z,r\phi}^A]) \tag{2}$$

Here, Mo takes the statistical mode across the full recommendation report, i.e., the mode of all recommendations received by trust evaluator from the referees on the destination node. We choose Mo as it provides the most common evaluation of the destination received from the referees.

4.2.1 Trustworthiness of the Referee

To rely on the user trustworthiness values computed from recommendations, it is important to consider the trustworthiness τ_r of a referee $r \in \mathcal{R}$, where $\mathcal{R}$ represents the set of referees a trust evaluator considers. For this, we can leverage the existing methods for computing trustworthiness, which are based on the characteristics of a Twitter account as proposed in the literature. Instead of using every possible feature and training a large dataset using a machine learning algorithm, the approach proposed in [25] suits our purpose well, which is described in the following paragraph.

We choose the two most important features reported in [22], namely (i) reputation and (ii) age, of a Twitter account, to compute the trustworthiness of a referee. Here, the reputation ($\mathbf{R}$) of a Twitter account is defined in [21] as:

$$\mathbf{R}_r = \frac{|\mathcal{M}_r|}{|\mathcal{M}_r| + |\mathfrak{F}_r|} \tag{3}$$

where $\mathcal{M}_r$ and $\mathfrak{F}_r$ are the set of followers and followees of the Twitter account of referee r, respectively. The reasoning behind such a definition is that, on Twitter, a reputable account usually has more followers than followees.

Using these factors, the trustworthiness of a referee is defined using Eq. (4).

$$\tau_r = \frac{\mathcal{A}_r}{\max\limits_{\forall \beta \in \mathcal{R}}(\mathcal{A}_\beta)} \times \frac{\mathbf{R}_r}{\max\limits_{\forall \beta \in \mathcal{R}}(\mathbf{R}_\beta)} \times \frac{\mathcal{M}_r}{\max\limits_{\forall \beta \in \mathcal{R}}(\mathcal{M}_\beta)} \tag{4}$$

Here, $\mathcal{A}$ is the age of the account that represents its length of exposure to Twitter platform. This is an important parameter generally used in calculating the trustworthiness of a referee [25]. While selecting the referee from the set of followers of a user, those with high values of τ_r are chosen and asked for their recommendation values. After obtaining their recommendations, the trustworthiness of the referees can be further scaled as per the following equation.

$$\tau_r = \frac{\tau_r}{\underset{\forall \beta \in \mathcal{R}}{max}(\tau_\beta)} \times (1 - |\widetilde{\psi}_{Z,\mathcal{R}\Phi}^A - \widetilde{\psi}_{Z,r\Phi}^A|) \tag{5}$$

Here, $\widetilde{\psi}_{Z,\mathcal{R}\Phi}^A$ and $\widetilde{\psi}_{Z,r\Phi}^A$ represents the median of the recommendation values obtained from the all referees $\mathcal{R}_Z^A$ and from the referee r in consideration, respectively. The median is used to remove the effects of any outliers on the recommendation system.

Since any follower of a Twitter user can be selected as a referee, the most reliable ones should be chosen to give their recommendations. For this, we need to filter out those who are members of a strongly-connected network (potentially

a Sybil [26] user) and would be likely to manipulate the recommendation system by providing untrue recommendations. To address this issue, we propose a threat model by defining the quality of a follower using social graph-based properties. The threat model is discussed separately in Section 5.

4.3 Evaluating Trustworthiness from Tweets

So far, we have calculated the trustworthiness of a Twitter user based on her web credentials or recommendations from referees. However, to evaluate user trustworthiness from her tweets on a particular topic, an application system needs to know how much it can rely on information disseminated by tweets generated from the account on that topic. This emphasises the importance of assessing the trustworthiness of Twitter data. There are already many methods proposed in the literature for mining trust information from tweet content. Existing methods are primarily machine learning-based and accordingly, their performance is heavily training data-dependent. We adopt a different approach, where each Twitter account is considered a source of information, and the quality of its information is used to determine the trustworthiness of the data generated from that Twitter account.

Consider that $I(\Theta_Z)$ represents information extracted from the tweets Θ_Z of user Z on a particular topic, whereas $I(\Theta)$ is the information extracted from a set of all tweets Θ from the Twitter-space on the same topic. We adopt the Euclidean distance between the two data sets as a measure of their similarity. This defines the trustworthiness of Z's tweet data on the topic, as approximated by A (Eq. (6)).

$$\Psi_Z^{\Theta} = 1 - \mathsf{D}(I(\Theta_Z), I(\Theta)) \tag{6}$$

Here, $\mathsf{D}(I(\Theta_Z), I(\Theta))$ represents the difference between the information extracted from the two sets of tweets. For a specific topic, to represent the information that a particular tweet conveys, we extract the opinion from the tweet text. Content analysis tools can extract opinions from the text of a user's tweets and assign it a numerical value. However, to measure the Euclidean distance between two unequal tweet datasets, we first convert the extracted opinions into a probability space by computing the probability using a histogram. For this, we consider z equal-length bins $B = \{b_1, b_2, ..., b_z\}$, and compute the probability of the bins for each dataset. Finally, the Euclidean distance is measured.

$$\mathsf{D}(I(\Theta_Z), I(\Theta)) = \sqrt{\sum_{b \in B} (P(\Theta(b)) - P(\Theta_Z(b)))^2} \tag{7}$$

Here, the probability $P(\Theta(b))$ is defined as $\frac{|O_{\Theta(b)}|}{|\Theta|}$, where $O_{\Theta(b)}$ is the set of extracted tweet opinions that belong to the opinion block b. In a similar way, $P(\Theta_Z(b))$ is calculated using the tweets Θ_Z of user Z.

4.3.1 Generating a Reference Sample

In Section 4.3, it was assumed that, in measuring the trustworthiness of a Twitter user based on her tweets, all the tweets on a certain topic are collected. In practice, this is not possible due to the well-known practical limitations of the Twitter APIs, i.e., the Twitter Search API only returns tweets published in the past seven days. To overcome the limitations of data availability, a sample set of tweets is extracted from Twitter using the combination of Search and Streaming APIs. Here, the Streaming API is configured to collect all tweets generated on a particular topic from Twitter's global stream of tweet data. The sample should be a quality representation of the information that exists in the Twitter-sphere. We consider two properties of the sample to ensure its quality, namely (i) expertise and (ii) diversity. These are discussed in detail in the next paragraph.

Topic Expertise of the Sample Set: Usually, experts (and leaders) have greater influence on general people. Hence a quality sample should contain tweets from experts. Let, $\mathcal{S}$ and $\mathcal{U}_S$ represent our sample and the set of users in the sample, respectively. For any user $u \in \mathcal{U}_\mathcal{S}$, $\mathcal{E}_u$ denotes their expertise level, which can be measured using the algorithm proposed in [27]. Now, Eq. (8) defines the ratio $\mathcal{E}_\mathcal{S}$ of experts to non-experts in a sample.

$$\mathcal{E}_\mathcal{S} = \frac{|\{u \in \mathcal{U}_\mathcal{S} : \mathcal{E}_u >= \mathcal{E}_{th}\}|}{|\mathcal{U}_\mathcal{S}|} \tag{8}$$

Here, $\mathcal{E}_{th}$ is the level of expertise that a user must have to be considered an expert on a topic. If expertise values range between 0 to 1, $\mathcal{E}_{th}$ can be assigned as 0.5. As per the definition, any sample with $\mathcal{E}_\mathcal{S} > 0.5$ is a good quality sample, since it contains more expert users than lay people.

Diversity of the Sample Set: As discussed earlier, a Twitter account works as an information source for its followers. For example, if u follows v, tweets from v populate the news-feed of u. Here, $v \in \mathfrak{F}_u$, where $\mathfrak{F}_u$ is the set of followees of u on Twitter. A sample is considered *diverse* when it is open to various information sources, whereas a *closed* sample is restricted to information from its members. Consequently, the *diversity* $\mathcal{O}_\mathcal{S}$ of the sample can be defined as follows.

$$\mathcal{O}_\mathcal{S} = \frac{|\cup_{u \in \mathcal{U}_\mathcal{S}} (\mathfrak{F}_u) - \mathcal{U}_\mathcal{S}|}{|\mathcal{U}_\mathcal{S}|} \tag{9}$$

Here, $\cup$ represents set union operation, and thus $\cup_{u \in \mathcal{U}_\mathcal{S}}(\mathfrak{F}_u)$ denotes the set of information sources for the sample. The numerator denotes the number of information sources that are new with respect to user set $\mathcal{U}_\mathcal{S}$ in the sample set. A sample with $\mathcal{O}_\mathcal{S} \approx 0$ is a very closed sample with very few new information sources. This might happen when all the tweets are published by a close-knit group of users and, hence do not comprise a representative sample. The higher the value of $\mathcal{O}_\mathcal{S}$, the more acceptable the sample is to be considered in measuring user trustworthiness using tweets.

Using Eqs. (8) and (9), a good quality reference sample set of tweets can be extracted from Twitter. The sample can be used to calculate $I(\Theta)$, which is then applied in Eq. (6) to estimate the trustworthiness of tweets generated by user Z.

4.4 Decision on a User

Before creating any association with a user, her trustworthiness needs to be evaluated using web credentials and referee recommendations. This is also important for building a trusted community. On the other hand, when using the user as an information source, trustworthiness based on generated tweets is more important. A user confirming with the information of a sample is as trustworthy as the sample for obtaining information. However, a user cannot be considered untrusted if her opinions differ significantly from the "common sense" of the Twitter network or from that of the representative sample. Under these circumstances, if the user is already verified by credentials, she can be regarded as a source of alternate information, or be ignored while making decisions based on Twitter information.

5 THREAT MODEL

The trust component measured according to recommendation is prone to attack when malicious users work in groups to form strongly-connected networks that positively review each other. To be resilient to such attacks, the proposed model needs to distinguish between referees that are trustworthy real users, and the fake ones that are generated by untrustworthy users (possibly by the user whose trustworthiness is to be evaluated). To address this issue, we propose an innovative way to measure the quality of potential referees and select those deemed to be suitable.

Let $\mathcal{R}_Z$ represent the set of potential referees for destination user Z, where most are fake Twitter accounts created by Z. They are termed as *Sybil* in the literature [26]. A network formed by Sybils (i) exhibits *fast mixing* property [26], that is, it is a strongly-connected graph and (ii) is connected by a limited number of edges to the honest part of the network– referred to as *attack edges*. These two properties ensure that most Sybil users are only connected to themselves, with a few having connections to the honest network through the attack edges. This observation leads us to formulate the quality measure of a referee as discussed below.

For any user $r \in \mathcal{R}_Z$, $\mathcal{Q}_r^1$ denotes the immediate quality of r, which is defined as:

$$\mathcal{Q}_r^1 = \frac{|\mathcal{M}_r - \mathcal{M}_Z|}{\hat{\mathcal{M}}_r^1} \tag{10}$$

where, $\mathcal{M}_r$ and $\mathcal{M}_Z$ represent the set of followers of referee r and user Z, respectively. The numerator signifies the number of new users that can be encountered through this connection of Z and r. The denominator $\hat{\mathcal{M}}_r^1 = |(\mathcal{M}_Z \cup \mathcal{M}_r)| + |(\mathcal{M}_Z \cap \mathcal{M}_r)|$ is a normalising factor, where $(\mathcal{M}_Z \cup \mathcal{M}_r)$ represents the users covered by Z and r together, and $(\mathcal{M}_Z \cap \mathcal{M}_r)$ denotes the users they share between them. If the number of new users encountered through r is smaller, or the number of users r and Z share is larger, the quality $\mathcal{Q}_r^1$ is lower. By exploring one hop further, we can define the 2-hop quality of r as per Eq. (11).

$$\mathcal{Q}_r^2 = \frac{1}{|\mathcal{M}_r|} \times \sum_{s \in \mathcal{M}_r} \frac{|\mathcal{M}_s - (\mathcal{M}_Z \cup \mathcal{M}_r)|}{\hat{\mathcal{M}}_r^2} \tag{11}$$

where $\hat{\mathcal{M}}_r^2 = |(\mathcal{M}_Z \cup \mathcal{M}_r \cup \mathcal{M}_s)| + |(\mathcal{M}_Z \cap \mathcal{M}_r)| + |((\mathcal{M}_Z \cup \mathcal{M}_r) \cap \mathcal{M}_s)|$. The 3-hop quality of referee r is similarly defined as:

$$\mathcal{Q}_r^3 = \frac{1}{|\mathcal{M}_r|} \times \sum_{s \in \mathcal{M}_r} \left(\frac{1}{|\mathcal{M}_s|} \times \sum_{k \in \mathcal{M}_s} \frac{|\mathcal{M}_k - (\mathcal{M}_Z \cup \mathcal{M}_r \cup \mathcal{M}_s)|}{\hat{\mathcal{M}}_r^3} \right) \tag{12}$$

where $\hat{\mathcal{M}}_r^3 = |(\mathcal{M}_Z \cup \mathcal{M}_r \cup \mathcal{M}_s \cup \mathcal{M}_k)| + |(\mathcal{M}_Z \cap \mathcal{M}_r)| + |((\mathcal{M}_Z \cup \mathcal{M}_r) \cap \mathcal{M}_s)| + |((\mathcal{M}_Z \cup \mathcal{M}_r \cup \mathcal{M}_s) \cap \mathcal{M}_t)|$. Similarly, the quality of r in further hops can be defined.

The quality of any referee of a user Z residing in the Sybil region yields smaller values for $\mathcal{Q}_r$, since its connections are limited in that region, which reduces the possibility of exploring new users as the number of hops increases. Consequently, the quality values drop quickly with increasing hops to a very small value near zero. However, for any user in the honest region, this will not occur, and their qualities have higher values even after considering a substantial number of hops. Therefore, if we restrict the referees from those with high quality, the proposed model can restrict malicious users to be selected as referee.

5.1 Finding a Lower Bound for $\mathcal{Q}_r$

A lower bound for the values of $\mathcal{Q}_r$ can be computed so that any user r having a quality value less than this lower bound can be discarded as a referee. Consider, ρ represents the fraction of users that a particular user Z is connected to within a Sybil region. Through a user r having connection only in that region, Z can explore, at most, $(1-\rho)$ new users. Following Eq. (10), $\mathcal{Q}_r^1$ can be bounded by $\mathcal{Q}_r^1 <= \frac{1-\rho}{1+(2\times\rho-1)} = \frac{1-\rho}{2\times\rho}$. Similarly, from Eqs. (11)-(12), $\mathcal{Q}_r^2 <= \frac{1-\rho}{1+\rho+(2\times\rho-1)} = \frac{1-\rho}{3\times\rho}$ and $\mathcal{Q}_r^3 <= \frac{1-\rho}{1+\rho+\rho+(2\times\rho-1)} = \frac{1-\rho}{4\times\rho}$. In general, for n-hop quality, $\mathcal{Q}_r^n <= \frac{1-\rho}{(n+1)\times\rho}$. This lower bound is applicable for any Sybil region with $\rho > 0.5$. However, the lower bound for any region with $\rho < 0.5$ is set empirically as $\frac{1}{n+2}$ for n-hop quality $\mathcal{Q}_r^n$.

6 FEASIBILITY ANALYSIS

As alluded to in Section 4, the computation of the trustworthiness of a user and her data is formulated statistically. However, the success of our model depends on the cooperation of intermediate users through which the trust evaluator can communicate with the destination user. Let α_u represent the probability that a user u will respond to any Twitter message forwarded to her. For the sake of analytical simplicity, we assume an equal response probability α for all users. For any path having length $\mathcal{L}$ from any source A to destination Z, the probability of such a path being established successfully depends on the response of all the intermediate users along with the destination. Equation (13) captures that probability.

$$P(A \leftarrow Z) = (\alpha)^{\mathcal{L}} \tag{13}$$

Here, we consider the path establishment a success only if all users along the path respond to A so that it can receive the credentials and list of referees from Z. Now, exploring through each of the followers $\mathcal{M}_A$, A can independently try to establish a successful path to Z with a success probability of $p = (\alpha)^{\mathcal{L}_m}$, where $\mathcal{L}_m$ is the path length from A to Z through the m^{th} follower of A. This process of establishing a successful communication path by exploring each of the followers is referred to as *forward exploration*, and is depicted in Fig. 3(a). Consequently, the expected number of paths established successfully from the independent exploration of $\mathcal{M}$ paths is given by the following equation.

$$\mathbb{E}(A \leftarrow Z) = \sum_{m=1}^{\mathcal{M}} (\alpha)^{\mathcal{L}_m} \tag{14}$$

Now, according to the "six degrees of separation" theory [30], any two users are connected through six or fewer steps on Twitter. Therefore, we can assume that there is, at best, a six-step path from A to Z with a successful communication probability of α^6. Moreover, each of the other $\mathcal{M}-1$ followers of A will provide at least one seven-step path, with each having a success probability of α^7. Therefore, the estimated expected success in establishing a path $\hat{\mathbb{E}}(A \leftarrow Z)$ is:

$$\begin{aligned} \mathbb{E}(A \leftarrow Z) \geq \hat{\mathbb{E}}(A \leftarrow Z) &= \alpha^6 (1 + \sum_{m=1}^{\mathcal{M}-1} \alpha) \\ &= \alpha^6 \left[(\mathcal{M}-1)\alpha + 1 \right] \end{aligned} \tag{15}$$

In Eq. (15), we consider the worst case scenario of having a six-step path between any two users; however, they can be connected by fewer than six steps. As suggested in [31], the average path length between any two users on Twitter

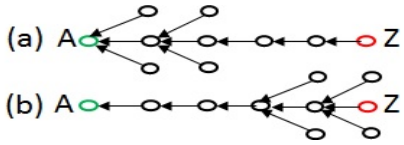


Fig. 3. Establishing a communication path: (a) forward exploration from the source (A); (b) backward exploration from the destination (Z).

is 4.2. Therefore, the actual number of successes will be greater than or equal to our estimated value $\hat{\mathbb{E}}(A \leftarrow Z)$, i.e., $\mathbb{E}(A \leftarrow Z) \geq \hat{\mathbb{E}}(A \leftarrow Z)$. Establishment of one successful path from A to Z is all that is needed in our model for the trust evaluator (A) to obtain the credentials and list of referees from Z. Therefore, equating $\hat{\mathbb{E}}(A \leftarrow Z) = 1$ with typical response probability $\alpha = 1/3$ [14], Eq. (15) yields $\mathcal{M} \geq 244$, whereas with $\alpha = 1/2$, it yields $\mathcal{M} \geq 21$. Finally, with $\alpha = 1$, $\mathcal{M} = 1$.

Exploring more than one path through each of the followers of A further ensures better success with smaller values of α and $\mathcal{M}$. For example, with each follower of A, if we explore the paths through all of their successive followers, $\hat{\mathbb{E}}(A \leftarrow Z)$ is evaluated using Eq. (16).

$$\hat{\mathbb{E}}(A \leftarrow Z) = \sum_{m=1}^{\mathcal{M}} \alpha(\sum_{n=1}^{\mathcal{M}}(\alpha)^{\mathcal{L}_n}) = \sum_{m=1}^{\mathcal{M}} \left[\alpha^6(\alpha^2(\mathcal{M}-1)+1)\right] \tag{16}$$

In simplifying Eq. (16), we again apply the six degrees of separation theory to establish the fact that there is at least one path of length seven through each of the followers of A's followers. Again, equating a successful path establishment with $\alpha = 1/3$, we have a reasonable estimate of $\mathcal{M} \geq 138$ for the average number of connections in an OSN. These estimates of $\mathcal{M}$ match the distribution of the average number of followers on Twitter, as found in [31]. This indicates high possibility of being successful in establishing a path between any two users A and Z on Twitter.

7 Searching for an Optimal Path

The previous section's analysis demonstrated that, in the worst case, our model is capable of establishing at least one successful path with high probability from any trust evaluator to the target user. However, exploring each path sequentially will not provide an optimal result in terms of success probability and exploration delay. Consequently, we need a guided search to find the best possible path from source A to destination Z. Here, we adopt the A* search to find such a heuristically-defined optimal path. For this, we need to define the heuristics. For a path to be optimal, it needs to have the following properties.

- **Shortest path**: The optimal path should be the shortest of all paths from A to Z; thus, the optimal path length is $\mathcal{L}(OP) = \underset{\forall \mathcal{P}(A,Z)}{min}(\mathcal{L}(\mathcal{P}(A,Z)))$, where $\mathcal{P}(A,Z)$ is a path between A and Z with length $\mathcal{L}(\mathcal{P}(A,Z))$.
- **Reliable path**: We need the optimal path to be reliable. For reliability, we consider the following two factors.
  1) **Trustworthy path**: A reliable path should be the one which includes the more trustworthy users. Traditionally, the trustworthiness of a path is defined as the lowest trustworthiness value of all users constituting a path. We use Ψ_X as the trustworthiness value for user X on a path, which can be computed hop-by-hop using her publicly-available tweets as per Eq. (6). The best possible path is the one that has the highest value for the lowest trustworthiness of its constituting users. Therefore, the trustworthiness Υ of the optimal path is defined as $\Upsilon(OP) = \underset{\forall \mathcal{P}(A,Z)}{max}\Psi(\mathcal{P}(A,Z)) = \underset{\forall \mathcal{P}(A,Z)}{max}(\underset{u \in \mathcal{P}(A,Z)}{min}(\Psi_u^{\Theta}))$.
  2) **Responsive path**: For reliability, we need a path where every user responds to messages quickly and positively. For this, we need to maximise the cumulative response probability of all users in a path. Here, the response probability of a path $\alpha(\mathcal{P}(A,Z)$ is defined as $\Pi_{u \in \mathcal{P}(A,Z)}\alpha_u$, which can be simplified using Eq. (13).
- **Expected success**: We need to consider the expected success of a path $\mathbb{E}(\mathcal{P}(A,Z))$ to ensure it is optimal. Since A* evaluates all successors of a user to select the best one in each iteration, it needs to consider the expected success of each of the users in the path to ensure the highest value of $\mathbb{E}(\mathcal{P}(A,Z))$. Therefore, $\mathbb{E}(\mathcal{P}(A,Z)) = \sum_{X \in \mathcal{P}(A,Z)}\mathbb{E}(X \leftarrow Z)$.

7.1 Cost Function and Proof of Obtaining a Quasi-Optimal Path

Based on the criteria described above, we define the following cost functions (heuristic) for any intermediate user X while it is explored by the A* search.

$$g(X) = \frac{1}{\alpha(\mathcal{P}(A,X)} \times \frac{1}{\Psi(\mathcal{P}(A,X))} \times \frac{\mathcal{L}(\mathcal{P}(A,X))}{\mathbb{E}(\mathcal{P}(A,X))} \tag{17}$$

$$h(X) = \frac{1}{\hat{\alpha}(\mathcal{P}(X,Z)} \times \frac{1}{\hat{\Psi}(\mathcal{P}(X,Z))} \times \frac{\hat{\mathcal{L}}(\mathcal{P}(X,Z))}{\hat{\mathbb{E}}(\mathcal{P}(X,Z))} \tag{18}$$

Here, $g(X)$ is the cost used to reach X from A, and $h(X)$ is the estimated heuristic (cost) from X to Z. In Eq. (18), $\hat{(\bullet)}$ represents the estimate of the corresponding parameters. For the A* search to guarantee a quasi-optimal path using a graph search from A to Z, we need to prove our heuristic to be *admissible* and consistent.

- **Admissible**: For admissibility, we have to prove that, for all intermediate users X, $h(X) \leq h^*(X)$ holds true, where $h^*(X)$ denotes the true cost of reaching the destination Z from X. Thus, the heuristic never overestimates the cost. While estimating a successor, we assume that the destination is one step away ($\hat{\mathcal{L}}(\mathcal{P}(X,Z)) = 1, \hat{\mathbb{E}}(\mathcal{P}(X,Z)) = \mathcal{M}_X$) from the successor. Moreover, the successor will evaluate the destination as trustworthy ($\hat{\Psi}(\mathcal{P}(X,Z)) = 1$) and responsive ($\hat{\alpha}(\mathcal{P}(X,Z) = 1$) with the highest possible values. Thus, $h(X) = \frac{1}{1} \times \frac{1}{1} \times \frac{1}{\mathcal{M}_X}$ will be always smaller than $h^*(X)$ with destinations more than one step from the successor. Here, $\mathcal{M}_X$ is the number of successors for X.
- **Consistent**: For consistent heuristic, we have to prove that for any edge (X, X'), $h(X) \leq h(X') + g(X, X')$ holds true, where $g(X, X')$ is the cost to reach X' from X. From the discussion above, $h(X) = \frac{1}{\mathcal{M}_X}, h(X') = \frac{1}{\mathcal{M}_{X'}}$, and $g(X, X') = \frac{1}{\Psi_{X'}} \times \frac{1}{\alpha_{X'}} \times \frac{1}{\alpha_{X'}}$. The minimum possible value of $g(X, X')$ is 1. Therefore, for the heuristic to be consistent, the inequality $\frac{1}{\mathcal{M}_X} \leq \frac{1}{\mathcal{M}_{X'}} + 1$ must hold for all possible edges (X, X') on Twitter.

Now, for any number of followers $\mathcal{M}_X, \mathcal{M}_{X'} >= 0$, the inequality holds, which proves that our heuristic is consistent.

Therefore, using the cost functions defined in Eqs. (17)-(18), A* search guarantees to return the best path to explore first, which reduces the delay and ensures the highest possibility of a successful transfer of credentials from Z to A.

8 Expected Delay

Besides finding an optimal path from A to Z, we need to estimate the average time required to establish the first successful path, or the time it might take before terminating the process without any successful communication with Z. Consider that a user takes x minutes to reply to any message forwarded to her, and $\mathcal{X}$ is the time-out period used to decide that there will be no reply from her. With a response probability α, the expected delay for a one hop communication is:

$$\mathbb{D}_1 = \alpha x + (1-\alpha)\mathcal{X} \tag{19}$$

For a path with length $\mathcal{L}$, the expected end-to-end communication delay is:

$$\mathbb{D}_{\mathcal{L}} = \sum_{l=1}^{\mathcal{L}}(\mathbb{D}_1) = \mathcal{L}\mathbb{D}_1 = \mathcal{L}(\alpha x + (1-\alpha)\mathcal{X}) \tag{20}$$

Because of the real-time nature of Twitter, for the majority of the questions, the earliest response arrives within five minutes [13]. Therefore, $x = 5$ is a rational assumption for reply time. Now, if we assume $\mathcal{X} = 30$ minutes for the time-out period and $\alpha = 0.33$ for the response probability, the expected delay for a path of length six would be $\mathbb{D}_{\mathcal{L}} \approx 132$ minutes, which is a reasonable waiting time.

As established in Section 6, to obtain credential information, our model can guarantee a minimum number of successful communication between a source and destination by exploring a path through each of the source's followers. This depends on (i) the number of followers ($\mathcal{M}$), (ii) the response probability (α) of a user, and (iii) the average path length ($\mathcal{L}$) from source to destination. Here, we define the average delay ($\overline{\mathbb{D}}$) in receiving the first successful response from the destination through such an exploration. Equation (21) captures such a delay.

$$\overline{\mathbb{D}} = \mathbb{D}_{\mathcal{L}} + \overline{\mathcal{M}} \times (\overline{\mathcal{L}} \times \mathbb{D}_1 + \mathcal{X}) \tag{21}$$

Here, $\mathbb{D}_1$ and $\mathbb{D}_{\mathcal{L}}$ are defined in Eqs. (19) and (20), respectively. In Eq. (21), $\overline{\mathcal{M}}$ represents the average number of paths that need to be explored before getting the first success, whereas $\overline{\mathcal{L}}$ denotes the average number of users on a path of length $\mathcal{L}$ that is traversed before being unsuccessful on that path. Considering that each user on a path makes her decision individually, $\overline{\mathcal{M}}$ and $\overline{\mathcal{L}}$ can be computed using the binomial distribution as per Eqs. (22) and (23).

$$\overline{\mathcal{L}} = (1-\alpha) \times \sum_{i=0}^{\mathcal{L}-1}(i) \times \alpha^{i-1} \tag{22}$$

$$\overline{\mathcal{M}} = (\alpha)^{\mathcal{L}} \times \sum_{i=0}^{\mathcal{M}-1}(i) \times (1-\alpha^{\mathcal{L}})^{(i-1)} \tag{23}$$

The average delay in establishing the first successful path is further investigated experimentally in Section 10.3.

The amount of delay can be reduced significantly if we use the A* search, instead of the above-mentioned exhaustive search. We analyse the delay incurred in the A* search as follows. Although the A* search provides a quasi-optimal path from source to destination, it does not guarantee that the path will be established successfully. If any of the users on the optimal path does not respond, an A* search from the last responding user needs to be resumed after excluding the non-responsive link from consideration. This method of searching sub-optimal paths from the last responding user on the heuristically optimal path is termed *backward exploration*, and is depicted in Fig. 3(b). However, we can limit the search by $\mathbb{P}$ number of such sub-optimal paths from the last responding user, so that other potential paths through the responsive users on the initial heuristically optimal path can be explored. The worst possible delay ($\mathbb{D}^*$), encountered when not getting a response from any of the sub-optimal paths from $\mathbb{P}$, is defined by:

$$\mathbb{D}^* = \sum_{q=1}^{\mathbb{P}}(\mathcal{X} + \frac{(\mathcal{L}_q - 1)}{2}\mathbb{D}_1) \tag{24}$$

where $\mathcal{L}_q$ is the length of the qth path. With the optimal path of length $\mathcal{L}$, the worst-case delay for a path establishment process without success is defined using Eq. (25).

$$\overline{\mathbb{D}}^* \leq (\mathcal{L}-1) \times (\mathcal{X} + \sum_{q=1}^{\mathbb{P}}\mathcal{L}_q\mathcal{X}) \tag{25}$$

The delay in establishing the first successful communication between source and destination using the A* search with backward exploration is further analysed experimentally in Section 10.4.

9 Incentive Mechanism

For the successful implementation of the proposed framework, (i) the destination user needs to agree to provide the credentials asked for by the trust evaluator and (ii) the intermediate users have to respond by forwarding the message to the next hop. In this section, we discuss an incentive mechanism which can convince users to participate in the above tasks. Following this incentive mechanism, a Twitter user will receive help to separate fake messages from the good ones depending on how much the user help other in similar cases. To be specific, collaboration through our approach helps real Twitter users to receive untainted/reliable information only and this is the biggest incentive for the Twitter community as a whole.

Referring to Fig. 2, let trust evaluator A ask destination Z for credentials. For this, A first introduces herself to Z by providing A's credentials. Moreover, the intermediate users evaluate their relationships with the previous user in the communication chain, e.g., B evaluates A and X evaluates B. These evaluation scores and A's credentials together assure Z that A is a real and trustworthy. This encourages Z to participate in the trust evaluation process.

When destination Z provides her credentials to trust evaluator A, the users along the path (A, B and X) inform their respective neighbours of this behaviour. Moreover, when an intermediate user agrees to forward the message, this information is also communicated to the neighbours of the users along the path. The neighbours then record the information for future reference. Using this shared information, an incentive mechanism is proposed as follows.

9.1 Incentive for the destination

After being assured that the trust evaluator A is a real user with a good reputation, the destination (Z) examines

A's earlier participation behaviour, i.e., whether A provided credentials previously to other users upon request. Any evidence of such behaviour will certainly increase the probability of Z responding to A's request. For this, Z will communicate with her neighbours to obtain the participation record of A. The adjustment of Z's response probability (α_Z) is captured as:

$$\alpha_Z(\mathcal{I}_Z) = \alpha_Z + \frac{1-\alpha_Z}{1+e^{-\mathsf{k}\mathcal{I}_Z}} \tag{26}$$

where, $\alpha_Z(\mathcal{I}_Z)$ is the modified response probability of Z after considering the incentive, whereas $\mathcal{I}_Z$ is the corresponding incentive metric which is computed as $\mathcal{I}_Z = \mathcal{N}_Z(A)/\mathcal{N}_Z$. $\mathcal{N}_Z$ and $\mathcal{N}_Z(A)$ represent the number of neighbours of Z and the number of neighbours having a record that A provided credentials on request, respectively. In Eq. (26), k controls the steepness of the logistic curve, which signifies the personal trait of the user in changing her decision upon the availability of information.

9.2 Incentive for Intermediate users

The incentive to forward a message by an intermediate user X can be modelled using similar reasoning to that described above. Any history of forwarding of X's message by the trust evaluator A certainly acts as a greater incentive for X to respond to A's current request. Considering these incentives, X's increased response probability can be modelled using Eq. (27)

$$\alpha_X(\mathcal{I}_X) = \begin{cases} 1 & \text{A forwarded X's request} \\ \alpha_X + (1-\alpha_X)\mathcal{I}_X & \text{Otherwise} \end{cases} \tag{27}$$

Likewise Eq. (26), $\mathcal{I}_X = \mathcal{N}_X(A)/\mathcal{N}_X$, which represents the fraction of neighbours of X who have the information that A has participated in the trust computation process by forwarding a message. Please note that incentives for destinations and intermediate users are defined differently (exponential vs. linear, as in Eqs. (26) and (27), respectively) to provide more incentives to the destinations, since getting credentials from a destination user is the ultimate goal of this communication process.

10 Simulation Results

We assessed the potential efficacy of the proposed framework for determining user trustworthiness through simulations using synthetic data, along with experiments using real Twitter data. For simulation, we implemented the Watts-Strogatz model [35] to generate a graph having 1000 nodes (users) with directed edges among them, which emulates the real connectivity structure of Twitter. It ensures the small-world properties of Twitter [34], which has short average path lengths and high clustering coefficients [33]. Each user is equipped with a number of web credentials which other users can evaluate to compute her trust. Web credentials are divided into four categories. A user can ask for any number of credentials from another user to evaluate the latter's trust.

Similarly, a user seeks recommendation from a number of referees to evaluate another user's trust, and a referee is asked to provide a recommendation report by evaluating a number of criteria (web credentials). The choices for all simulation parameters are specified in the respective sections. Moreover, the publicly accessible Streaming and Rest APIs were used to collect Twitter data. The description of the data set is presented in Section 10.5.

10.1 User Trustworthiness

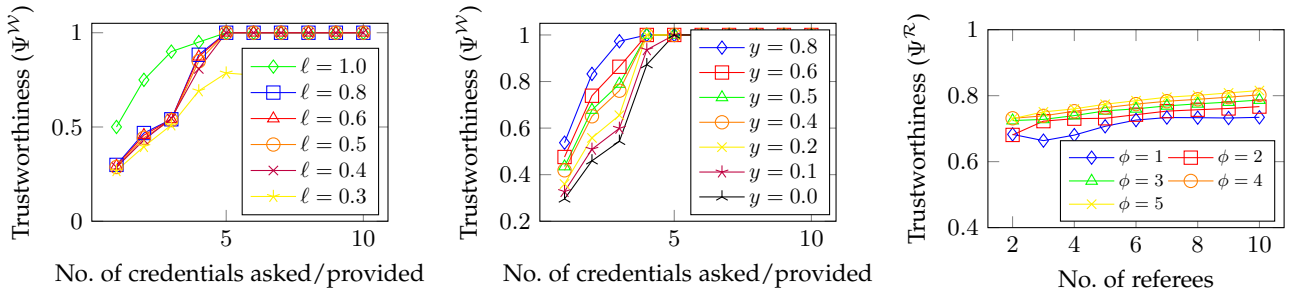
In this section, to emulate people's natural way of developing trust of others through an identity verification process, we present experiments that investigated the impacts of a number of factors on users' evaluated trustworthiness; namely the number of credentials, the number of referees, the number of identity criteria presented in a recommendation report, and the recommendation values available to the trust evaluator.

In the first experiment, we considered that a user could have a maximum of 15 possible web credentials. Here, to represent various scenarios, we varied the average number of credentials, as represented by ℓ in Fig. 4(a), that a user actually had. $\ell = 1.0$ denotes every user had all 15 credentials, whereas $\ell = 0.5$ means, on average, a user had only half of this. Similar to the real-world identity verification system established by the Australian Government [12], where different credential categories (e.g., primary photo IDs like a driver's licence or passport constitute 60 points) score different amounts of points, the credentials were classified into four categories in the simulations, worth $\omega = 100, 50, 30, 10$ points. We assigned 100 points to the first category, which includes credentials that are very hard to fake; hence a user with such a credential is deemed trustworthy.

When a user provides a credential of a specific value ω, the trust evaluators have their own perceptions or evaluations of the merit of that credential, which is within a range $[0, \omega]$. In the experiment illustrated in Fig. 4(a), we varied the number of credentials. For each set of credentials provided, the trustworthiness of the users was computed. For this experiment, divergence in the perceived value of credentials by the trust evaluators was incorporated by varying individual evaluation within $[0, \omega]$. Here, for statistical validity, trustworthiness was taken as an average across the network. For averaging, trustworthiness was evaluated for each pair of users in the network for each respective set of credentials requested or provided.

From the figure, it is evident that a user's trustworthiness does not depend on the value of ℓ, i.e., the number of credentials that the user has. Rather, trustworthiness depends on the number of credentials requested and provided by the users. Nevertheless, for a user to be considered trustworthy, at least five credentials are required in all the scenarios, except when ℓ is too low ($\ell \leq 0.3$).

In Fig. 4(b), a similar set of results as described above is presented, where users' average trustworthiness is depicted against the number of credentials available to the trust evaluator, but with a different experimental condition. For this, we considered $1 - y$ as the divergence of users' perceptions on the value of a credential ω, where this value can be perceived in the range $[y \times \omega, \omega]$. For a group of trust evaluators with high level of knowledge, and with the authenticity of the credentials provided being very high, y would be higher, and consequently, the diversity of the evaluated values (perceptions) would be less. On the other hand, in the presence of possible fake and untrue credentials, y would be lower, and the resulting diversity in human



(a) Trustworthiness with various average numbers of user credentials

(b) Trustworthiness with various users perceptions of other users' credentials

(c) Trustworthiness with varying numbers of referees and evaluated criteria

Fig. 4. (a)-(b) Trust vs. number of credentials requested/provided; (c) trust vs. number of referees. Here, ℓ represents the average fraction of credentials users possess, $1 - y$ captures the variations in perceptions of the credential values (ω) by individuals where the perceived values are distributed in the range $[y \times \omega, \omega]$, and represents the number of identity criteria based on which a referee recommends.

perception would be high. In this experiment, ℓ was chosen as 0.6, representing a society of people having 60% of the possible credentials.

From the figure we see that user trustworthiness can be measured accurately with a lower number of credentials when the corresponding value of y is higher. However, when the value of y is lower, more credentials are required to verify a user as trustworthy. These observations are realistic in nature. When people have reasonably good understanding of the credentials' true values and the credentials are information-rich, they can infer the trustworthiness of others with a limited number of credentials. On the other hand, when people have a different understanding of the value of web credentials, or when ambiguous and fake web presences are very common in the society, we need more credentials to cross-match with to verify a user as trustworthy. For both the cases described in Figs. 4(a)-(b), the number of credentials needed to assess the trustworthiness of a user reasonably is five; however, with richly-perceived credentials, even four is good enough (Fig. 4(b)).

Referee recommendation is used, in parallel with web credentials, to evaluate user, which in turn cross-validates these two measures of trustworthiness. In the third experiment, we captured such a process of assessing user trustworthiness with recommendation reports received from nominated referees. For this experiment, we considered a maximum of 10 referees, and trust is computed using Eq. (2). Each referee provides a report by evaluating the user's credentials from different perspectives, such as the nature of the existing relationships between the user and the referee, and its duration, the user's profession, her acceptance in the society, and other factors that may be pertinent. We called these *identity criteria* as represented by ϕ in Fig. 4(c), and described in Section 4.

From these results, we observed that trustworthiness from recommendation increases with the number of identity criteria and with the number of referees. More identity criteria mean better perception and observation on a user. Similarly, with the presence of more referees, there is less confusion on a user's trustworthiness. Interestingly, trustworthiness computed using recommendations from three referees with 3–5 identity criteria closely matches with trustworthiness evaluation based on three web credentials.

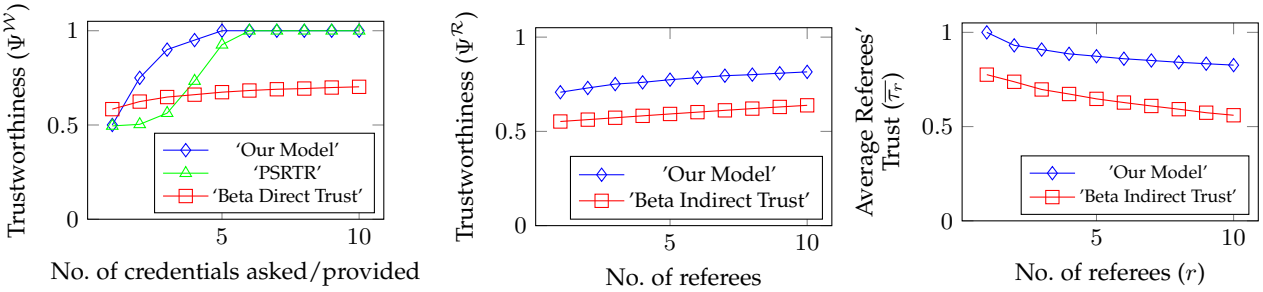
10.2 Comparison

This section presents the performance of our model in evaluating user trust. First, user trust using web credentials ($\Psi^{\mathcal{W}}$) is analysed against two relevant and contemporary models ('Beta Trust' [38] and PSRTR [39]), where trust is evaluated in the context of web-services recommendation system. The first one is called PSRTR [39] that computes trust based on interest background, evaluation tendency and recommendation effect. The other one is the well-reputed beta trust model that is utilised in [38] by a user to evaluate direct trust of another user before taking web-service recommendation from. Although these two models mainly recommend web services to a user based on the experience of other users, the technique of evaluating the trust of a recommending user aligns with that of our model.

Trust values are computed based on the simulating settings presented in Fig. 4(a) with $\ell = 1.0$. The number of credentials used to evaluate a user trust varied from 1 to 10, and the value of a credential can be perceived by the trust evaluator in $[0 - \omega]$ where ω is the highest possible value of a credential. The parameters for the beta model, as referred to by success and failure, are evaluated based on the difference between the perception of the values of web credentials by the trust evaluator and the user (whose trust is being evaluated).

Fig. 5(a) shows the trust values of our and PSRTR models, and the direct trust of Beta trust model for verifying a user identity using the available web credentials. Fig. 5(a) evidences our model always outperforms Beta trust model for any number of web credentials. It also performs better than PSRTR model when verifying users with a smaller number (≤ 5) of credentials. The requirement of five credentials to obtain a trust value of 1.0 is sufficient in web platform as in the real-world identity verification process between two to four identities are usually used [12]. The categorisation of web credentials into multiple groups and accumulating the credential values for user verification are the key to the better performance of our model.

Second, user trust using referees' recommendations ($\Psi^{\mathcal{R}}$) is evaluated comparing with the indirect trust of a user as computed in [38] using the weighted average of the direct trust values between the recommended user and a set of users targeted by the trust evaluator. The number of



(a) Trustworthiness computed using Our Model, PSRTR and Beta Direct Trust

(b) Trustworthiness ($\Psi^{\mathcal{R}}$) from referees' recommendations

(c) Trustworthiness with varying numbers of referees and evaluated criteria

Fig. 5. (a) Trust vs. number of credentials requested/provided; (b) trust vs. no of referees; (c) average trust of the referees ($\overline{\tau}_r$) vs. no of referees.

referees is varied from 1 to 10, and 5 identity criteria are used for trust evaluation. The results are presented in Fig. 5(b). The simulation settings are equivalent to that used in Fig. 4(c) with $\phi = 5$.

The figure shows that the use of statistical mode operation to compute trust from recommendation reports in our method outperforms the Beta indirect trust computed in [38]. Mode can capture the commonalities among the referees evaluations, which eventually helps the trust evaluator to better verify user trustworthiness in our method. On the other hand, varying referee number does have a lesser impact on the trust computed for the user network, because of the nature of the average function used for trust computation.

One strength of the proposed model is the selection of referees (recommenders) ($r \in \mathcal{R}$) based on their trust (τ_r) scores. The model always selects the highly trusted recommenders. In the absence of any mention of referee selection method in [38], a random choice of recommenders yielded the trust of the selected referees using the Beta indirect trust model is depicted in Fig. 5(c). It is observed that our model always picks referees that are highly trustworthy. On the other hand, random selection can result in recommenders of low average trust score.

10.3 Average Success and Delay

In this section, the simulation results for success (number of successfully established paths between the trust evaluator and the destination) and delay (in establishing the first successful path) of our proposed framework are presented. A social network graph with small world property with 5000 users was constructed with an average path length of around 3.2. We present the results for users with follower counts from 88 to 113 to provide an indication that our model will work in the Twitter network. The response probability α was set as 0.33.

First, we present the simulation results for the average percentage of successes obtained by exploring a single path to the destination through each follower of a trust evaluator. Therefore, for a user with $\mathcal{M}$ followers, $\mathcal{M}$ paths are explored and the number of successfully established paths is recorded. This is compared with the theoretical lower bound as defined by Eq. (15) in Section 6. In addition, the success ratio after applying the proposed incentive mechanism is also recorded. Figure. 6(a) illustrates these findings.

From the results, it is evident that the success ratio we obtained experimentally is close to those computed analytically, which validates the analytical computation. Moreover, the results imply that the number of experimentally obtained successes and those guaranteed analytically, are greater than 1, which ensures that our model will be successful in contacting and receiving credentials from destination users, even if the response probabilities of those users are as low as 33%. With the use of our easily implemented incentive mechanism, the success rate is three to four times greater than that obtained analytically and without an incentive mechanism. This validates the efficacy of our proposed incentive mechanism.

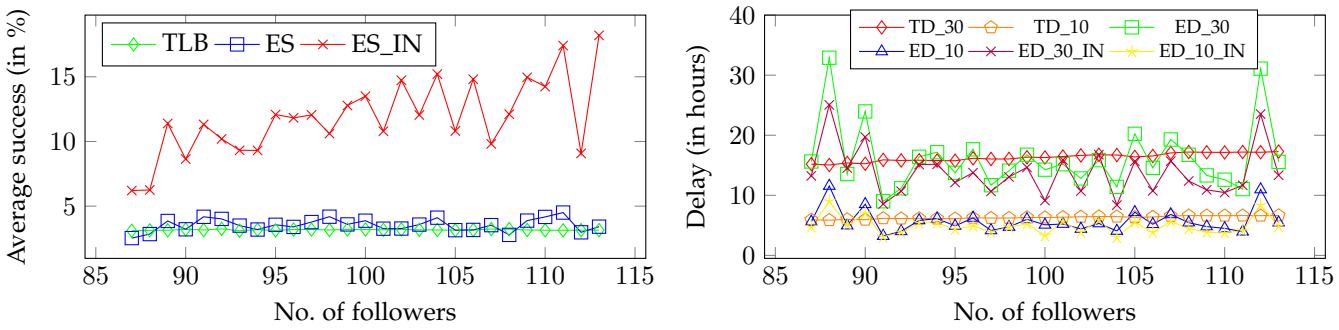
Second, we elaborate the results for the delays (in hours) in getting the first successfully established path to contact the destination. Both the theoretical and experimental average delays for the first success are presented, along with the delays experienced when the incentive mechanism is employed. The delays were computed separately by considering 30-min and 10-min time-out periods. Figure 6(b) demonstrates these results. The average expected delay for the first success with a 30-min time-out is around 16 hours, and that with a 10-min time-out is 5.7 hours.

From these results, it is evident that the average delays predicted analytically match those obtained experimentally, which validates our analytical computation of expected delays. Considering the real-time nature of Twitter [13], 10 minute is a more realistic assumption for a time-out period, and the corresponding delay of 5.7 hours is a reasonable period to wait to receive credentials from a user. In addition, when the incentive mechanism is employed, the delay can be further reduced by three hours for 30-min and by one hour for 10-min time-out periods.

10.4 Average Delay using A* Search

To evaluate the effectiveness of the devised A* search in obtaining the best path from the source to the destination user, we conducted the same experiments as described above, but with a response probability α distributed within (0.25, 1.0). The results are presented in Fig. 7.

In Fig. 7(a), the delays in obtaining the first success are presented, considering a 30-min and 10-min time-out periods. For both cases, the delays incurred by A* search were smaller than those obtained by exploring a single path through each follower. For a 30-min time-out, on average,



(a) Success in path establishment

(b) Delay in establishing a path

Fig. 6. (a) Average success (average rate of successfully established paths), (b) delay for the very first success vs. the number of followers. In (a) TLB: theoretical lower bound of success (Eq. (14)), ES and ES_IN: success obtained without and with the incentive mechanism, respectively. In (b), TD_30, TD_10: average delay for the first successful communication with the destination user determined analytically considering 30-min and 10-min time-out periods, respectively; ED_30, ED_10: the corresponding delays obtained experimentally, and ED_30_IN, ED_10_IN : the experimental delays with the incentive mechanism.

the delay was reduced by around 2.5 hours, whereas it was decreased by 15 minutes for a 10-min time-out.

Figure 7(b) shows the number of users that has to be explored before obtaining the first success, for both the exploration methods. The improvement by A* search is clearly evident. On average, A* search needed to explore around three users fewer to obtain the first success, which increases the possibility of establishing the communication path within a realistic time and distance. It also validates the definition of the cost functions (Eqs. (18)-(19)).

10.5 Trustworthiness of User Tweets

We evaluated the trustworthiness of user data by measuring the quality of their tweets, as proposed in Section 4. For this, we streamed Twitter using its Streaming API for a period of 30 consecutive days between November 04, 2015 and December 03, 2015 on the topic of the "Vaccination Debate", which is globally important because of conflicting expert opinions. A substantial total of 113270 tweets were collected with 55930 unique users.

The expertise of the users was evaluated using the method proposed in [27]. To generate a reference sample $\mathcal{S}$, a total of 515 users were selected with a sample expertise $\mathcal{E}_{\mathcal{S}} = 1.0$ and a sample openness $\mathcal{O}_{\mathcal{S}} = 25.9$, as computed using Eqs. (8) and (9), respectively. The high value of $\mathcal{O}_{\mathcal{S}}$ indicates that the sample included information from a large number of sources, as many as approximately $13,350$ in this case. Our reference sample included a total of $3,285$ tweets from the 515 selected users.

To extract information from a user tweet, we used sentiment analysis tools, namely the commercially-available tool found in http://text-processing.com/. Current sentiment analysis tools can achieve more than 80% accuracy [40], which might vary depending on the underlying approach of using machine learning or lexicon/dictionary based algorithms. Note that the information presented in a tweet refers to the opinion of a user on the topic, and is represented in the range $[0, 1]$. The information extracted from the $3,285$ tweets constituted the sample of the Twitter-sphere's information on that topic. To compute a user's trustworthiness, the information extracted from her tweets was evaluated using the sample information according to Eqs. (6) and (7).

TABLE 2
The performance of statistical and machine learning models

	Accuracy	Sensitivity	Precision	F1 Score
Our Model	0.73	0.89	0.74	0.80
Machine Leanring	0.6	0.90	0.64	0.74

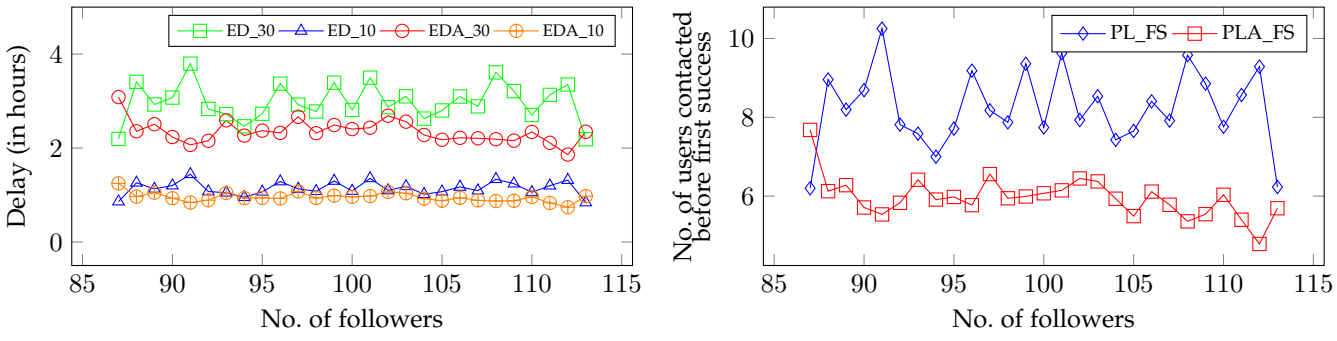
We evaluated the trustworthiness of 244 users who had more than one tweet in our dataset. This constitutes our evaluated sample.

To evaluate the performance of our trustworthiness measure based on user tweets, we considered the fact that Twitter routinely suspends accounts which mainly engaging in spamming or abusive behaviours.[1] Consequently, they are definitely untrustworthy. In our evaluation, 55 accounts among the 244 we considered for assessment were deemed untrustworthy (trust scores ≤ 0.5). We tried to extract their profiles from Twitter, and found that 34 of them had already been suspended by Twitter. Therefore, more than 60% of the users we evaluated as untrustworthy were also deemed untrustworthy by Twitter.

In our evaluated sample, there were 59 suspended accounts with an average trustworthiness score of 0.51. On the other hand, the average trustworthiness of this sample was 0.65. If we discard the suspended accounts from consideration, the sample's average trustworthiness score rises to 0.70. This indicates that our method rightfully assigns lower trustworthiness (in the vicinity of 0.5) to accounts that are eventually suspended by Twitter and, hence, the method is effective in assessing trust and finding untrustworthy users.

To evaluate the performance of our statistical model, a machine learning approach is employed to measure the trustworthiness of user Tweets. Table 2 compares the outcomes. The training data set was prepared using the reference sample of 515 users, where two persons independently evaluated the trustworthiness of user tweets subjectively, and the data was labeled using the average of these evaluations. Deep learning neural network model was implemented using Tensorflow (https://www.tensorflow.org/) to train a binary classifier, and then used to classify the 244

1. https://support.twitter.com/articles/15790



(a) Delay for the first success

(b) users visited before first success

Fig. 7. (a) Average delay (in hours) for the first success; (b) average number of users to be contacted to before reaching the destination vs. number of followers. In (a) ED_30, ED_10: the average delay for the first successful communication with the destination user obtained through simulation (exploring one path through each follower) considering 30-min and 10-min time-out periods, respectively, and EDA_30, EDA_10: the corresponding delay obtained by exploring paths determined by A* search and the backward exploration method. In (b): PL_FS and PLA_FS represent the number of users to be explored before obtaining the first success, using one path per follower, and A* with backward exploration, respectively.

users into trustworthy and non-trustworthy classes. Please note that a user was classified as trustworthy when her trustworthiness score was greater than 0.5.

From the results, it is evident that our statistical model performs better in predicting users as trustworthy compared with the deep learning-based model. Our model achieves better accuracy, precision and F1 score. The sensitivity (recall) of both models are nearly equal. Being a statistical model, the proposed model performs better than machine learning to predict the trustworthiness of user tweets, as with fewer attributes (one in our case – sentiment value), the statistical model performs better than machine learning

10.6 Performance of the Threat Model

In this section, we present the performance of the proposed threat model in detecting Sybil users. For this, a Sybil region is added with the social network created earlier for the other simulations. In this context, the original social network is referred to as the *honest region*. Any user is detected as Sybil if its n-hop quality $\mathcal{Q}^n$ is below the threshold, as defined in Section 5. The results are presented in Fig. 8.

First, we applied our Sybil detection mechanism for both the honest and Sybil regions using a 1- to 4- hop (n) quality index and a threshold of $\frac{1}{(n+2)}$. For this experiment, a standard Sybil region was created with 100 users that had a strongly connected community structure. The results are presented in Fig. 8(a). From the results, it is evident that the performance of our threat model in detecting Sybil users is very good for all hop quality scores except for $n = 1$, where it is very poor for honest regions using 1-hop quality. For $n = 1$, around 60% honest users are classed as Sybil.

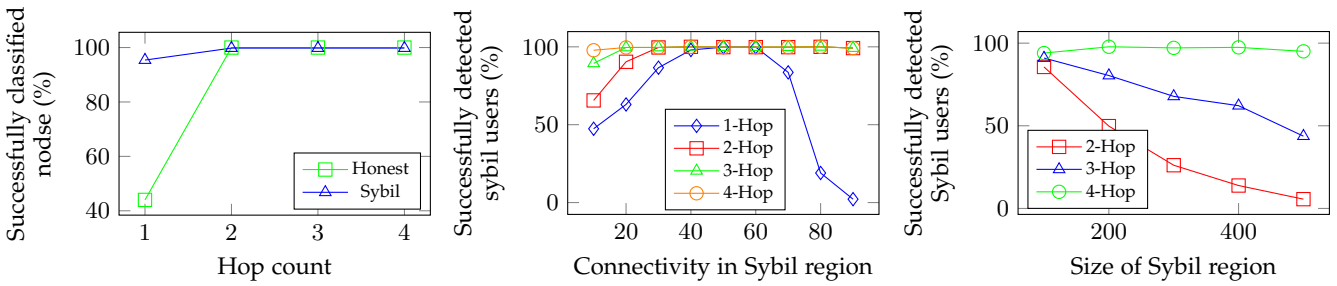
Second, the connectivity (ρ) of the Sybil region was varied from 10% ($\rho = 0.1$) to 90% ($\rho = 0.9$), and the performance of the model was evaluated for 1-4 hop quality. Here, the Sybil region had 100 users, and as discussed in Section 5.1, the threshold was chosen as $\frac{1-\rho}{(1+n)*\rho}$ for $\rho > 0.5$, and $\frac{1}{(n+2)}$ otherwise. The results are depicted in Fig. (8)(b). The performance using 3 and 4 hop quality was very high, with almost every user in the Sybil region being detected. Performance with 2-hop quality was quite acceptable, especially when there was enough connectivity in the Sybil region. Though not presented here, all the users in the honest region were detected successfully under this scenario. Based on these two experiments, it can be concluded that the n-hop quality index is suitable for differentiating between Sybil and honest users for higher values of n.

In the third set of experiments, the number of Sybil users was increased from 100 to 500 to create a larger Sybil region. The Sybil users were also connected following the small world property, to emulate a scenario where attackers try to create a network with the properties of a real social network. Consequently, the threshold was selected as $\frac{1}{(n+2)}$. From the results of this experimental scenario (Fig. 8(c)), it is observed that Sybil detection performance decreased with the increase in Sybil region size, for both 2- and 3-hop quality scores, although the latter performed better than the former. In contrast, using 4-hop quality index, more than 95% of users were successfully detected as Sybil users even when its region size was large.

Based on the findings discussed above, we can conclude that for better performance, we need to consider the highest possible quality indices in terms of hop count, although this might be computationally expensive for larger networks. A 4-hop quality strikes a balance between performance and computational complexity. In addition, it is possible to fine-tune the thresholds rather than using values computed from the aforementioned formulae. In that case, it is possible to obtain acceptable performance with a lower hop quality index. This is investigated in the following section.

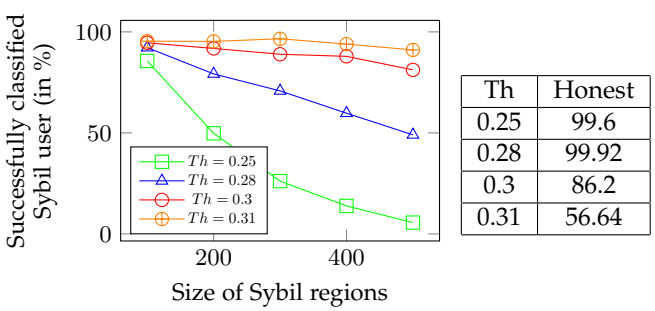
10.6.1 *Choosing threshold for 2- and 3-hop quality*

As discussed above, the performance of 2- and 3-hop quality indices in detecting Sybil users degrades as the Sybil region size increases. Especially, if the connectivity in Sybil regions follows that of the honest region, the theoretical threshold chosen as $\frac{1}{n+2}$ (see Section 5) cannot differentiate successfully between Sybil and honest regions. However, it is possible to improve this situation by fine-tuning the thresholds empirically around the theoretically-derived value in such a way that our method can yield acceptable results using the less computationally-intensive 2- and 3-hop quality values. The empirical analysis is presented in Fig. 9. For this, the experiment represented by Fig. 8(c) was

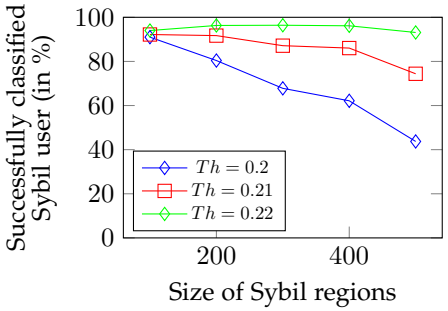


(a) Performance vs. hop count (b) Performance vs. connectivity in Sybil region (c) Performance vs. Sybil region size

Fig. 8. Performance of the proposed threat model. (a) Percentage of users correctly classified in the respective categories. Honest: users belong to the honest region of the network, Sybil: users belong to the Sybil region of the network, (b) percentage of successfully detected Sybil users vs the connectivity of the Sybil region in terms of ρ $(\times 100)$ and (c) percentage of detected Sybil users vs the size of the Sybil region.



Th	Honest
0.25	99.6
0.28	99.92
0.3	86.2
0.31	56.64



Th	Honest
0.2	100
0.21	99.96
0.22	95.24

(a) 2-hop quality: Sybil region (b) 2-hop quality: honest region (c) 3-hop quality: Sybil region (d) 3-hop quality: honest region

Fig. 9. The effects of the threshold on the performance of the proposed threat model. (a) Percentage of detected Sybil users vs Sybil region size for the 2-hop quality measure, with a threshold in the range $[0.25 - 0.31]$; (b) average percentage of successfully detected honest users for the same thresholds with 2-hop quality; (c) percentage of detected Sybil users vs the size of the Sybil region for the 3-hop quality measure, with a threshold in the range $[0.2 - 0.23]$; (d) average percentage of successfully detected honest users for the same thresholds with 3-hop quality.

repeated for 2- and 3-hop quality values using thresholds of $[0.25 - 0.31]$ and $[0.2 - 0.23]$, respectively, mainly in the vicinity of the corresponding theoretical value of $\frac{1}{n+2}$.

For Sybil detection, Fig. 9(a) shows that our model attains very good performance for thresholds ≥ 0.3. For those values, the performance for the honest region starts decreasing, especially when the threshold reaches 0.31 (Fig. 9(b)). We can set the value at 0.3, as this gives reasonably better results for both regions compared with other values. Note that the performance in the honest region is less important as long as the method can select the required number of referees for a user.

Similar set of results are demonstrated for 3-hop quality in Figs. 9(c) and (d). The threshold value for 3-hop quality can be set as 0.22, which is a trade-off in detection performance for the Sybil and honest regions.

11 Conclusion

In this paper, we propose a holistic approach to determine the trustworthiness of a Twitter user and her tweets. We employed the concept of identity verification in real-world system to compute trustworthiness using a number of her information-centric sources, namely web credentials and recommended referee reports. Referee recommendations are used for those users who have no or limited number of web credentials. On the other hand, the trustworthiness of tweet data generated by a user is computed by determining their quality of information with respect to that of a quality reference sample extracted from Twitter-space on a particular topic. Simulation results evidence that, with information-rich web presences and recommendations, 3 to 5 web credentials or the same number of referee recommendations are required to verify a user as trustworthy. This is consistent with the number of identity documents and referees needed to verify a user's identity trust in a physical system.

We analysed the feasibility of our proposed model analytically and through simulation. We observed that it is possible to successfully establish a communication path between the trust evaluator and any Twitter user within a reasonable time period, and web credentials and referee reports can be obtained through this established path. Furthermore, the heuristic cost functions that we defined for the A* search is found to be effective in producing better success and delay. Therefore, for time critical applications, instead of the exhaustive search, one can implement our heuristic based A* search to establish a communication path quickly. In addition, an easy to implement incentive mechanism has been devised that attains better performance in terms of both success and delay. Our trustworthiness computation method when applied to real-world user tweets revealed some suspected accounts which were later founded suspended by Twitter itself, substantiating its applicability in

evaluating users' data trustworthiness. Finally, the performance of the proposed threat model ensures that Sybil nodes cannot inflate the trustworthiness computation by providing untrue and high recommendation.

REFERENCES

[1] *Twitter for Business: Understand and Unlock the Power of Twitter for Your Business*, [online]. Available: https://business.twitter.com/en.html, retrieved on 26th July 2016.

[2] *The Evolving Role of News on Twitter and Facebook*, Pew Research Center, [online]. Available: http://www.journalism.org/files/2015/07/Twitter-and-News-Survey-Report-FINAL2.pdf, retrieved on 26th July 2016.

[3] R. Thomson and N. Ito and H. Suda and F. Lin and Y. Liu and R. Hayasaka and R. Isochi and Z Wang, *Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter*, in ISCRAM, 2012.

[4] A. Anderson and D. Huttenlocher and J. Kleinberg and J. Leskovec, *Effect of User Similarity in Social Media*, in WSDM, 2012.

[5] Y. A. Kim, *An Enhanced Trust Propagation Approach with Expertise and Homophily-Based Trust Networks*, Knowledge-Based Systems, vol. 82, pp. 20-28, 2015.

[6] L. Zhao and T. Hua and C.-T. Lu and I.-R. Chen, *A Topic-Focused Trust Model for Twitter*, Computer Communications, vol. 76, pp. 1-11, 2016.

[7] J. Zou and F. Fekri, *Leveraging Online Social Relationships for Predicting User Trustworthiness*, in GLOBECOM, 2015.

[8] B. A. Salih and P. Wongthongtham and S.-M.-R. Beheshti and D. Zhu, *A Preliminary Approach to Domain-Based Evaluation of Users' Trustworthiness in Online Social Networks*, in IEEE International Congress on Big Data, 2015.

[9] Y. Yao and H. Tong and X. Yan and J. Lu, *Multi-Aspect + Transitivity + Bias: An Integral Trust Inference Model*, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 7, pp. 1706-1719, 2014.

[10] B. Soeder and K. S. Barber, *A Model for Calculating User-Identity Trustworthiness in Online Transactions*, in Privacy, Security and Trust, 2015.

[11] S. Hamdi and A.L. Gancarski and A. Bouzeghoub, *TISoN: Trust Inference in Trust-Oriented Social Networks*. ACM Transactions on Information Systems, vol. 34, no. 3, 2016.

[12] 100 Point Check, [online]. https://en.wikipedia.org/wiki/100_point_check, retrieved on 26th July 2016.

[13] S. A. Paul and L. Hong and H. Chi, *Is Twitter a Good Place for Asking Questions? A Characterization Stud*, in ICWSM, 2011.

[14] J. Mahmud and M. X. Zhou and N. Megiddo and J. Nichols and C. Drews, *Optimizing the Selection of Strangers to Answer Questions in Social Media*, CoRR, 2014, abs/1404.2013.

[15] Z. Zhang, *Security, Trust and Risk in Multimedia Social Networks*, The Computer Journal, vol. 58, no. 4, 2015.

[16] Z. Zhang and B. Gupta, *Social Media Security and Trustworthiness: Overview and New Direction*, Future Generation Computer Systems, 2016.

[17] W. Sherchan and S. Nepal and C. Paris, *A Survey of Trust in Social Networks*, ACM Computing Surveys, vol. 45, no. 4, 2013.

[18] Z. Yu and H. Yu, *Untrusted User Detection in Microblogs*, in Trust, Security and Privacy in Computing and Communications (TrustCom), 2014.

[19] F. Benevenuto and G. Magno and T. Rodrigues and V. Almeida, *Detecting Spammers on Twitter*, in CEAS 2011

[20] J. Song and S. Lee and J. Kim, *Spam Filtering in Twitter Using Sender-Receiver Relationship*, in RAID 2011.

[21] A. H. Wang, *Don't follow me: Spam detection in Twitter*, in SECRYPT, 2010.

[22] W. Herzallah and H. Faris and O. Adwan, *Feature engineering for detecting spammers on Twitter: Modelling and analysis*, Journal of Information Science, 2017.

[23] A. A. Amleshwaram and N. Reddy and S. Yadav and G. Gu and C. Yang, *CATS: Characterizing Automation of Twitter Spammers*, in COMSNETS 2013.

[24] P. Lin and P. Huang, *A study of effective features for detecting long-surviving Twitter spam accounts*, in ICACT, 2013.

[25] R. Hassan and G. Karmakar and J. Karmuzzaman, *Reputation and User Requirement Based Price Modelling for Dynamic Spectrum Access*, IEEE transactions on Mobile Computing, vol. 13, no. 9, 2014.

[26] M. Al-Qurishi and M. Al-Rakhami and A. Alamri and M. Alrubaian and S. M. M. Rahman and M. S. Hossain, *Sybil Defense Techniques in Online Social Networks: A Survey*, IEEE Access, vol. 5, 2017.

[27] R. Das and J. Kamruzzaman and G. Karmakar, *Modelling Majority and Expert Influences on Opinion Formation in Online Social Networks*, WWW, 2017.

[28] S. Cresci and R. Pietro and M. Petrocchi and A. Spognardi and M. Tesconi, *Fame for sale: Efficient detection of fake Twitter followers*, Decision Support Systems, vol. 80, 2015, pages 56-71.

[29] Q. Cao and M. Sirivianos and X. Yang and T. Pregueiro, *Aiding the detection of fake accounts in large scale social online services*, in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12), 2012.

[30] S. Milgram *The small world problem* Psychology today, vol. 2, no. 1, 1967, pages 60–67.

[31] H. Kwak and C. Lee and H. Park and S. Moon, *What is Twitter, a social network or a news media?*, in proceedings of the 19th international conference on World wide web (WWW '10), 2010.

[32] R. Thomson, N. Ito, H. Suda, F. Lin, Y. Liu, R. Hayasaka, R. Isochi and Z. Wang, *Trusting tweets: The Fukushima disaster and information source credibility on Twitter*, in proceedings of the 9th International ISCRAM Conference, 2012.

[33] S. Aparicio, J. Villazón-Terrazas, and G. Álvarez, *A Model for Scale-Free Networks: Application to Twitter*, Entropy, vol. 17, no. 8, pages 5848-5867, 2015.

[34] E. Ch'ng, *Local Interactions and the Emergence of a Twitter Small-World Network* Social Networking, vol. 4, pages 33-40, 2015.

[35] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, vol. 393, pages 440-442, 1998.

[36] O. Varol, E. Ferrara, C. A. Davis, F. Menczer and A. Flammini1, *Online Human-Bot Interactions: Detection, Estimation, and Characterization*, in Eleventh International AAAI Conference on Web and Social Media, 2017.

[37] B. Wang, L. Zhang and N.Z. Gong, *SybilBlind: Detecting Fake Users in Online Social Networks Without Manual Labels*, in Research in Attacks, Intrusions, and Defenses (RAID), 2018.

[38] H. Y. Wang, W. B. Yang and S. C. Wang, *A Service Recommendation Method Based on Trustworthy Community[J]*, Chinese Journal of Computers, 2014.

[39] H. Tian and P. Liang, *Personalized Service Recommendation Based on Trust Relationship*, Scientific Programming, 2017.

[40] V. A. Kharde and S. Sonawane, *Sentiment analysis of twitter data: A survey of techniques*, International Journal of Computer Applications, vol. 139, no. 11, pages 5–15 2016.

Rajkumar Das obtained his PhD degree in Computer Science from Monash University, Australia, and his B.Sc. and M.Sc. degree in computer science and engineering (CSE) from Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2008 and 2011, respectively. He worked as an assistant professor in CSE, BUET. His research interests include social dynamics, user trust, Machine Learning.

Gour Karmakar received the B.Sc. degree from CSE, BUET in 1993 and Masters and Ph.D. degrees in Information Technology from the Faculty of Information Technology, Monash University, in 1999 and 2003, respectively. He is currently a senior lecturer at Federation University Australia. His research interest includes multimedia signal processing, wireless sensor and social networks.

Joarder Kamruzzaman received the B.Sc. and M.Sc. degrees in Electrical and Electronic engineering from Bangladesh University of Engineering and Technology, Bangladesh, and Ph.D. degree in Information Systems Engineering from Muroran Institute of Technology, Japan. Currently, he is a Professor in Federation University Australia. His research interest includes sensor networks and computational intelligence.