Identity Deception Prevention using Common Contribution Network Data

Michail Tsikerdekis

***Abstract*—Identity deception in social media applications has negatively impacted online communities and it is likely to increase as the social media user population grows. The ease of generating new accounts on social media has exacerbated the issue. Many previous studies have been posited that focused on both verbal, non-verbal and network data produced by users in an attempt to detect identity deception. However, although these methods produced a high accuracy, they are mainly reactive to the issue of identity deception. This paper proposes a proactive approach that leverages social network data and it is focused on identity deception prevention for online sub-communities, communities that exist within larger communities (e.g., Facebook groups or Subreddits). The method can be applied to various types of social media applications and produces high accuracy in identifying deceptive accounts at the time of attempted entry to a sub-community. Performance results as well as limitations for the method are presented. A discussion follows on the identification of possible implications of this study for social media applications and future directions on deception prevention are proposed.**

***Index Terms*—identity, deception, prevention, network, data**

I. Introduction

SOCIAL media accounts have substantially increased over the course of the past decade [1], [2] along with an increase of social media platforms flooding both mobile and conventional computer systems. The increase in the number of accounts, which is often associated with an ease of creating online profiles on these media platforms, has attracted industry and scientific attention in relation to aspects of online deception and in particular identity deception. Users can generate fake profiles with objectives that vary from terrorism [3] to scamming and trolling [4]. These malicious accounts have far reaching consequences for the society at large as well as the online communities that they plague.

Solutions to these types of identity deception attacks exist with varying effects in terms of their efficiency, human labor or computational expenses. Typically, social media rely on the sheer numbers of their legitimate community members and administrators to identify malicious accounts. However, this is a labor intensive strategy that has been demonstrated to have a limited effectiveness. Many accounts can remain undetected for months or even years depending on how sophisticated the identity deception attack is [5]. The ability for users to generate new accounts with ease exacerbates the problem. Computational solutions for identity deception detection have also been proposed with varying computational overhead, accuracy as well as complexity of implementation [5], [6], [7]. These solutions focus on verbal and non-verbal approaches to identity deception detection. Non-verbal approaches such as user activity and similar behavioral patterns (rather than text) have demonstrated high success rates in identity deception within both offline [8] and online contexts [5], [9]. Nevertheless, all of the aforementioned examples involve a reactive approach in dealing with identity deception. An alternative approach is focusing on deception prevention instead of deception detection. Although identity deception prevention solutions have not been sufficiently addressed by literature, they have been speculated to be more effective in their applications with online social networks [10]. *The particular problem examined in this paper focuses on preventing access to a sub-community when a user has been active and otherwise potentially appearing legitimate to a larger community or platform. Examples of such sub-communities are Facebook's groups and Reddit's subreddits.*

M. Tsikerdekis is with the Information Communication Technology (ICT) program, University of Kentucky, Lexington, KY, 40506 USA e-mail: tsikerdekis@uky.edu.

The main contributions of this work can be summarized as follows:

- This paper introduces an approach that focuses on the domain of *deception prevention using social network data*. Past methods have used social network data (e.g., friendship network) for deception detection purposes especially in cases of Sybil attacks (e.g., a user creating multiple accounts to elevate their rank in the network) [11], [12], [13]. The method presented in this paper utilizes social network data (in particular common contribution network) in order to establish a community's profile (baseline behavior) and prevent identity deception. It is aimed at uncovering behaviors related to cases of *vandalism and scamming in sub-communities* as opposed to larger network Sybil attacks in past studies. This is achieved by establishing a new user's "fit" with the overall sub-community's profile
- The approach ensures that malicious accounts can be barred from participating in a community before attempting to deceive others. Additionally, the approach presents a more complex obstacle for the deceivers since replicating social fit in a common contribution network involves substantial effort.
- To demonstrate the accuracy and feasibility of the proposed method, this study utilizes publicly available data for a large social media community, Reddit. While Reddit data are used as a proof of concept, the method can be applied to other social media platforms classifications [14] within certain contexts that allow for sub-communities to exist within large communities.

The rest of the paper is organized as follows. In section II, an overview on deception and identity deception is presented

along with a discussion on current identity deception detection, as well as prevention methods. The goal is to frame the work in relation to the larger deception work that has been done in the field. This aims to introduce some of the unique research contributions of utilizing the method proposed in this paper. In section III, the proposed method is described. Section IV presents the performance results for the proposed method. Finally, Section V introduces the implications of the proposed method and proposes future directions for identity deception prevention in the social media domain.

II. Related Works

A. Deception and Identity Deception

Deception is the deliberate transfer of false information to a recipient that is not aware that the information received has been falsified [7], [15]. It is often seen as a way for a deceiver to achieve goal-driven (instrumental), relationship-drive (relational) or identity driven objectives [16]. The intent associated with a deceptive action can be benign or hostile [17]. Factors that can affect the deceiver, and as a consequence deception success, include a deceiver's expectations, goals, motivations, his/her relation to a target as well as a target's degree of suspicion [16]. Others have also suggested that the moral cost for a deceiver can affect the likelihood of engaging in deception [18].

Online, deception can also be affected by the medium [9]. This often refers to software design where one can influence factors such as the perceived level of security that a system provides, or solutions that provide enhanced assurances and trust for online users [19]. Other factors also relate to the Information Communication Technology (ICT) literacy of victims where the knowledge of technology and related security aspects can provide clues to deception [9], [19].

Online deceptive attacks are categorized under three major components: *content manipulation, channel manipulation and identity manipulation* [9]. Most commonly a combination of these attacks can make the overall deception more effective. For example, in online social networks, scamming will often involve a combination of identity and content deception [10].

Identity deception is of a particular interest to the study involved in this paper. A deceiver's goal is to manipulate the sender information with the intent for *identity concealment* (e.g., concealing or altering part of an individual's identity), *identity theft* (e.g., mimicking another individual's identity) or *identity forgery* (e.g., forging a fictional identity) [7], [20]. The latest category is a common attempt of identity deception found in social media [5]. Individuals are often enabled to create an unlimited number of accounts on a social media platform with limited identity verification requirements. This allows for deceivers to achieve identity forgery more easily. However, it should be noted that identity forgery does not necessarily imply malicious intent. There are legitimate reasons for an individual to hide their identity in respect to protecting their privacy. As such, identity forgery may be a desired approach for achieving this objective. Nevertheless, commonly, social media platforms are plagued by individuals that abuse this option in order to achieve malicious objectives. These are the particular cases that are the focus of this study.

B. Deception Prevention

Alternatives to reactive approaches in deception detection have been proposed [9], [10]. These focus on proactively disabling the ability of a deceiver to cause disruption in a social media platform. In particular, identity deception has attracted attention for solutions that can prevent or limit its impact. Further, identity deception can serve as a gateway for content and communication detection. As such, due to a reduction in the risks taken by a deceiver (through a "protected" identity), attacks of content and communication deception appear more feasible.

A proposed solution for limiting identity deception is controlling it at entry points to a social media platform. The most common place is at user registration. Software can place additional verification requirements for new accounts that substantially increase the difficulty for multiple account registration [9]. This paper refers to this option as entry point requirements. However, this strategy may discourage community participation which is a necessary condition for sustaining active social media communities, which rely on user-generated content. Another issue with such an approach is the substantial increase in computational loads that these methods may require [10]. Additionally, many of the additional obstacles can often be bypassed. For example, users required to submit their telephone numbers in order to activate their account can bypass this security measure through disposable phones or Web SMS services.

An alternative solution to entry criteria is the gradual addition of privileges to a new user as they transition through levels of progression in a community [10]. For simplicity, this paper refers to this option as progressive privileges granting. This method can be seen as a golden medium between applying strict entry criteria to newcomers and having no entry criteria at all. Nevertheless, the drawbacks of this method are often that the levels of progression can be easily identified by users and as a consequence, deceivers. As such, the cost for identity deception often increases but only by a factor of time, not necessarily difficulty.

The two aforementioned approaches relate to the method proposed by this study, which attempts to combine them within a particular social media context, which is explained in later sections.

C. Identity Deception Detection and Prevention

The method proposed by this paper is aimed at identity deception prevention. However, there is an absence of studies on deception prevention that utilize a machine learning approach similar to this study. Two deception detection methods are used for comparison with the proposed deception prevention method. In practical terms, the method proposed by this study could be developed to act as a deception detection method depending on the different context of application. As such, references to its computational overhead and detection accuracy are directly comparable to the methods described below.

Human identity deception detection is the most commonly used method by social media users today. Human judgment

has also been proposed as deception prevention mechanism [10]. However, the efficiency of deception detection can vary substantially depending on cues and time available to an individual [19], [21]. Statistical chance for distinguishing a fake from a non-fake account with no prior belief on the matter, would produce an equal probability between the two options (50 percent chance of proper identification). Human deception detection can vary between 55 to 60 percent [22] and results have been known to fall to as low as 34 percent [23].

Algorithmic approaches to identity deception detection have been found to have substantially higher accuracy rates but at a cost of computational expense [24]. For large social media communities the impact of computational overheads can render the implementation of an algorithm infeasible. Nevertheless, they offer the ability to decrease labor for administrators and are therefore important solutions.

A study by Solorio et al. attempted to identify malicious accounts on Wikipedia. These accounts are often referred to as sockpuppets, and are owned by previously blocked users [6]. Since entry criteria for registering new accounts are not so restrictive, users that are blocked for violating community rules can generate new accounts. In effect, they forge new identities that are meant to be difficult to link to their previous accounts. Nevertheless, when natural language processing techniques were utilized in order to compare new accounts with known blocked sockpuppet accounts, similarities were revealed in the written language. Punctuation count, quotation count, variations between capital and lowercase "I" among other textual features linked the owner of a new account to his or her already blocked sockpuppet accounts. By tracking the revision history for a sample of 77 cases of legitimate and sockpuppet accounts, the similarity-based method was able to identify cases correctly with a 68.83 percent accuracy using a Support Vector Machines (SVM) model. The drawback of the method is its complexity due to the need of comparing a user to a complete dataset of known blocked sockpuppets. An additional drawback is the need for the account to generate written content. This adds a requirement for a certain amount of time to pass or content added before the method can be utilized. Assuming that this requirement is met, testing for a new user to the existing blocked user dataset will result in a time complexity of $\mathcal{O}(N * R)$, where N is the number of blocked users as well as user of interest and R is the number of revisions made by all the users in the user set.

A more recent method that examined the multiple account identity deception problem attempted to increase accuracy while reducing the time complexity [5]. Looking at non-verbal behavioral indicators, the authors of the study were able to identify malicious accounts in sample of approximately 12,000 legitimate and malicious accounts with a 71.3 percent accuracy using an Adaptive Boosting Algorithm (ADA) model. Similar to the previous method, the approach requires a certain amount of time to pass before detection is feasible. Within 1 day after user registration the method's accuracy was 67 percent. After 30 days since a user's registration the same accuracy jumped to 71.3 percent. The time complexity of the algorithm was reported to be $\mathcal{O}(R)$ requiring for the method to read only the number of revisions (or actions) made by the single user of interest. The main computational overhead for the method is the requirement for building a proper training set in order to establish a legitimate user baseline (as anticipated by EVT). The method also coincides with known deception theories that highlight the inability for deceivers to control non-verbal behavior to the degree that they do with verbal behavior. In fact, many users may not be aware that their accounts are being monitored. Literature suggests that a deceiver will maximize his/her effort in ensuring that his/her verbal behavior does not expose the deception being carried out [5], [25], [26].

Studies utilizing social network features have also been developed in order to identify Sybil attacks [11], [12], [13]. These demonstrated the effectiveness of detecting fake accounts on online social networks by leveraging on network structure features. A study that investigated malicious accounts on Twitter has utilized user profile as well as ego (user-centric) network metrics [27]. For example, one measure established that fake accounts tend to have less triangles than legitimate user accounts. Such network statistics were demonstrated to be good classifiers for detecting malicious Twitter accounts, however, they have not been applied for preventing access to sub-communities and are not directly applicable to such a problem. For one, many accounts may demonstrate legitimate behavior in a larger community (Facebook) with the sole purpose of gaining access to a particular sub-community (Facebook group). This is also supported by literature regarding more complex deceptive attacks [9]. Further, the aforementioned method implicitly requires data on the complete graph, which renders it computationally expensive for large networks. Finally such studies also make an implicit assumption on the uniformity of behavior among users of large social networks, which for diverse online networks it may not hold. This may be especially true if the networks derive their edges using user interest, messages exchanged, or topics liked and commented as opposed to friendship or follower relationships.

A more relevant study to this paper's problem (preventing access to a sub-community) utilized common network features in order to determine the legitimacy of a relationship between two users on Facebook [28]. The purpose was to identify if a user is a likely to be "foe or friend" based on common characteristics (e.g., the two users have liked the same items or belong to the same groups). The method yielded high accuracy results in the testing dataset producing an Area Under Curve (AUC) of 0.778 and an F-measure of 0.696.

Some of the aforementioned methods demonstrate relatively high detection rates at the expense of computational cost. On one hand, verbal cues are easily conceptualized and applied for deception metrics. On the other hand, non-verbal cues have been known to be more effective in exposing deceivers [29], [30]. However, such studies do not identify the potential for incorporating social constructs, which are at the core of social media platforms. Studies that utilized social network features have demonstrated high detection results of malicious accounts for online social networks. However, such studies focus largely on detecting malicious accounts when an attack occurs on the larger platform (e.g., a user creates an account with the purpose of creating a Sybil attack or posting malicious links). They are

not applicable for cases when a malicious user attempting to gain access to a group is otherwise appearing to be using a platform in a legitimate manner. Further, such methods have been explicitly applied to platforms where some sort of "follower" relationship exists (e.g., Facebook's friendships or Twitter's followers). An alternative approach is determining if a user is a likely friend based on the set of common interests [28]. However, the concept has yet to be applied for assessing the legitimacy of an access request by a user against the profile of a whole sub-community. Further, some of the aforementioned methods (with the exception of [27]) provide predictable patterns where if a deceiver follows; he or she could then masquerade as a legitimate user and avoid detection. Features that establish a "social fit" to a larger sub-community are much more difficult to fake. Common contribution network data, utilized in this study, can uniquely answer such questions and it is likely more rigorous in determining the legitimacy of an account comparing directly a user's profile with that of the whole sub-community based on a network of interests.

III. Proposed Method for Preventing Online Identity Deception

A. Problem Domain

This study aims to develop a method that can automatically prevent identity deception in particular social media contexts. These contexts specify social media platforms that allow users to join and participate in sub-communities. An example of this is Facebook's closed groups. For instance, a school may decide to utilize a closed group so that parents and children can have access to discussions about events and other associated news within a school or group. In the event that the school does not have access to emails of parents, parents may have to initiate registration themselves. Determining which requests are legitimate is of the utmost importance in such a scenario. A deceiver can potentially engage in identity theft and aim to gain access to this community. Once access is gained, the deceiver has access to all the information protected under the closed group and can further expand his or her goals using content deception. Trading or meet up groups (e.g., minority groups) are other examples of such closed groups that could benefit from this approach.

This study utilized publicly available data from the social bookmarking site Reddit as a representative example for such closed groups. While the study focuses on a particular example of social media platform, the method is applicable to many social media platforms that are similar to Reddit in terms of the ability to contain "locked" sub-communities and social network data. The next section elaborates more on the nature of social network data as well as sub-communities on Reddit.

B. The Reddit Environment

Reddit is a social bookmarking platform that anyone can join by registering an account. At the time of writing, the only necessary criteria for creating a new account is an email address. After registering an account in the community, a user can post and comment on messages to *subreddits*. These can be seen as tags associated with posts that revolve around a particular topic of interest. A related term (although not exact) for a subreddit, are the hashtags used by users on Twitter. Some subreddits can apply restriction criteria for posting content or commenting. Others also maintain lists of legitimate users as well as scammers (another term for deceiver).

The data used in this study came from the subreddit called *giftcardexchange* [1]. The subreddit serves as a sub-community that allows for users to exchange their gift cards. Since the trade occurs across geographical distances, identifying legitimate users with whom one can trade is important. The community employs several tools in order to decrease the likelihood of scammers. It maintains other subreddits that assist them with maintaining active reputation profiles for existing members [2] as well as a wiki within Reddit that helps maintain a scammer's list (also called a banlist) [3]. Gift card exchanges are assisted by a bot that helps identify reputation profiles for members. The bot also assists with banning individuals that do not qualify for a reputation profile. The subreddit uses entry point requirements based on two rules: an account needs to be two weeks old and active. Active, meaning a balanced amount of posts and comments on other subreddits with a rough rule of one activity page (approximately 30 posts or comments) per one month of account age.

C. Social Network Data and Variables

Reddit users generate social network data based on their verbal and non-verbal activity. While verbal and non-verbal behaviors of previous detection methods determine user activity, social network data focuses on the social profile of a user and his or her "social fit" with the giftcardexchange sub-community. This is in stark contrast to the existing method used by the sub-community of interest, where users are compared to what is considered to be an "active" reddit user rather than how well one fits in with the existing users of a sub-community. This is also, arguably, an ever-changing attribute as the sub-community expands on its members.

For the purposes of this study, common contribution networks for users were utilized as the primary social network; that is, the number of pages a user has commonly contributed with existing members in the giftcardexchange community. For example, if a user has made a post in the pics subreddit and several existing giftcardexchange members have also made posts there, they form a connection (edge in social network terms) in the social network with that user. Subsequent connections that exist between all existing members of the giftcardexchange community are also taken into account. As such, the overall fit of a new member with the existing community's interests can be examined with a social network analysis approach. Such common contribution networks have been examined in various studies on Wikipedia and were important in uncovering behavioral patterns for members that would have been otherwise inaccessible [31]. Fig. 1 depicts the formation of a social network based on common contributions

[1] https://www.reddit.com/r/giftcardexchange/
[2] https://www.reddit.com/r/GCXRep
[3] https://www.reddit.com/r/UniversalScammerList/wiki/banlist

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2016.2607697, IEEE Transactions on Information Forensics and Security

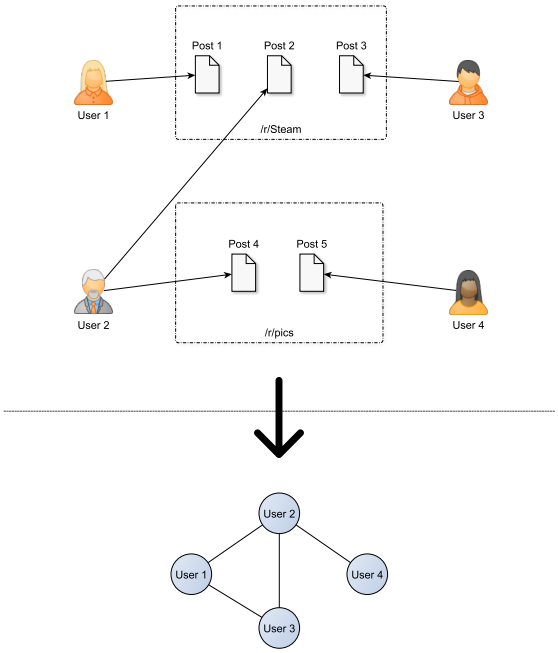

Fig. 1. Formation of a common contribution network based on Reddit data. Subreddits in this example include /r/pics and /r/Steam. All users are assumed to be part of the sub-community of interest (e.g., giftcardexchange).

made by members. Users 1, 2 and 3 form an edge between one another through their participation in the Steam subreddit. Similarly, Users 2 and 4 form an edge through participating in the pics subreddit. The network structure reflects shared common interests between users. For example, user 2 has common interests with all users in the network whereas user 4 is disconnected from the others.

Algorithm 1 is used to create a view of a sub-community's network which is part of the proposed deception prevention method. The algorithm has been built so that it can be easily generalized. Also, by maintaining only a list of edges as opposed to nodes, nodes that do not have common contributions with other nodes in the network are not included. Simply put, the network does not include isolate nodes, which would otherwise affect node-specific metrics. Also, since this is a common contribution network, edges are considered to be undirected. Finally, a date restriction is included in the algorithm for generating snapshots of the network at a particular date.

Once the network is built, social network analysis can be applied in order to determine a node's "fit" within the network. While social network analysis often involves examining the structural properties of a network as a whole, in the method proposed by this study, node-level metrics are more important, since the aim is to determine whether the account attempting to join the sub-community is malicious.

The core idea behind the approach posits that individuals that are a good and legitimate "fit" with other community members will be more central in the network structure. This is due to the fact that they are bound to share contributions to subreddits that existing members also contribute. As such, centrality metrics that quantify the position of a user in the

Algorithm 1 Build an edge list of a sub-community's network based on user common contributions.

procedure COMMONCONTRIBUTIONNETWORK($Date$ = Today's Date, $Uid = 0$)
  $U \leftarrow$ [all legitimate users of the sub-community and Uid (if not 0)]
  for i in U **do**
    $P \leftarrow$ [all posts (except posts in sub-community) made by i prior to $Date$]
    for j in P **do**
      $UP[j]$.append(i)
    end for
  end for
  $E \leftarrow List()$
  for i in UP **do**
    $US \leftarrow List()$
    for j in i **do**
      US.append(j)
    end for
    CALCULATE unique pairs (x, y) from US set, store result in $Pairs = [[x1, y1], [x1, y2], , [xn, yn - 1], [xn, yn]]$
    E.append($Pairs$)
  end for
  E = Unique(E) ▷ Clears duplicates but this information can also be kept for weighted networks
  return E
end procedure

overall network's structure can describe the "fit" of a user. Further, manipulation of centrality metrics by a deceiver is more complex since prior knowledge of the network structure is necessary. The method involves several node specific social network analysis metrics which are further described below.

The least sophisticated metric that determines a node's position in a network compared to other nodes in the network is a node's degree ($deg(v)$) or more simply put, the number of edges to which a node is connected. This is also often referred to as degree centrality ($C_D(v)$) for a particular node. However, the metric is blind to the overall network's structure especially when used for node-specific calculation. Nevertheless, it is important because it shows common "interests" between a node and other nodes in the sub-community.

Eccentricity ($e(v)$) is another rudimentary metric of a node's position, however, the structure of the overall network is taken into account. It is the maximum network distance (also known as geodesic distance) for a node v in a network G between v and any other node i in G. In turn, a geodesic distance for nodes (v, i) is the shortest path between them in G (usually obtained through walks in a graph).

A more refined metric that takes into account the total structure of the network in determining how central a node is, also utilizes geodesic distances between nodes. It is referred

to as closeness centrality and is defined as follows:

$$C_C(v) = \frac{\sum_{i:i\neq v} \frac{1}{d(v,i)}}{|V(G)| - 1} \quad (1)$$

where $d(v,i)$ is the geodesic distance between the node of interest v and another node i in a network G [32].

While closeness centrality aims to identify how central a node is in the overall network's structure, betweenness centrality aims to identify how critical a node is as a "bridge" [33]. It is defined as:

$$C_B(v) = \sum_{i,j:i\neq j,i\neq v,j\neq v} \frac{g_{ivj}}{g_{ij}} \quad (2)$$

where g_{ij} are the number of geodesic distances from i to j, while g_{ivj} are the geodesic distances from i to j that pass through v.

Alternative centrality metrics also exist that utilize walk-based eigenvalues. Eigenvector centrality $C_E(v)$ focuses on the idea that a central node should also have connections to other "powerful" nodes. The algorithm assigns an attribute to each node in a graph $x_i = 1$. Scores are then updated based on the following formula:

$$C_E(v) = \sum_{j,N} a_{vj} * x_j \quad (3)$$

where N is the number of all adjacent nodes to node v, a_{vj} is an adjacency matrix where $a_{vj} = 1$ if an edge exists between v and j and $a_{vj} = 0$ otherwise. Values in $C_E(v)$ are then normalized by dividing based on the largest value. The updating and normalization process keeps repeating until values in $C_E(v)$ stop changing.

Burt's constraint was also used as a metric in order to identify how critical is the position of a node. The metric identifies a node's value through its access to non-redundant contacts. An individual that bridges two separate parts of the network is more likely to encounter unique information coming from separate groups and as such they are in an advantageous position compare to other nodes. In network representations other than friendship networks (e.g., common contribution networks), an individual with a high score in this metric can be seen as a more diverse individual. Constraint is formally defined as:

$$C(v) = \sum_{j\in V_v\setminus\{v\}} (p_{vj} + \sum_{q\in V_v\setminus\{v,j\}} p_{vq}p_{qj})^2 \quad (4)$$

where V_v is the ego network for a node v. An ego network consists of a focal node (ego) and nodes (called alters) that the focal node is directly connected to along with all ties that may exist among the alters. p_{vj}, p_{vq}, p_{qj} are derived by the following formula:

$$p_{xy} = \frac{a_{xy} + a_{yx}}{\sum_{k\in V_x\setminus\{x\}}(a_{xk} + a_{kx})} \quad (5)$$

where x and y are the indexes for p and a_{xy} are elements in graph's adjacency matrix A.

The aforementioned metrics can generate a profile for the position of each node in a network and describe how critical a node is. They were implemented as part of the deception prevention mode in order to monitor how the "networking" behavior of a user's account fits with the behaviors exhibited by previous members.

The final step in the method is generating training data that would formulate a baseline behavior. This is achieved by utilizing supervised machine learning methods where once the models are computed based on a training set they are ready to be utilized for new account cases that attempt to join a sub-community.

D. Data Retrieval and Model Testing

Data were collected based on the banlist provided by the giftcardexchange subreddit. The banned user logs span a period from September 2014 until December 2015. Each ban includes a comment describing the reason a user has been banned. The banned log included 2,719 accounts (relating to the subreddit giftcardexchange). Nevertheless, many of these were banned for violating the existing entry point requirement of having an active account of at least 14 days old. In fact, only an approximate 15 percent of the accounts that were banned had posted a topic on other subreddits before the time they were banned (not accounting for comments to other subreddits). Simply put, 85 percent of banned accounts were inactive. Users that had no previous posts and were banned using the aforementioned account activity rule, were eliminated from the sample. The accounts left in the sample were 419 users that managed to evade entry point requirements and were reactively detected as deceivers by an administrator at a later time. Given that the banlist is relatively young (15 months during the time of observation) these accounts result in an estimated average of 23 identity deception cases per month. These cases evade the entry rules of the subreddit (not accounting for improvements that may have been made to the bot over time).

Many of these users had a substantial history of posts in the Reddit community and arguably went to a great effort in attempting to masquerade as legitimate users. For example, one banned user, submitted the posts displayed on table I up to the point when they were banned by the sub-community. The particular user had evaded administrators for approximately 2 months until finally, the user was banned on November 29, 2014 at 21:05. On average for the 419 users that evaded the initial entry point requirements, it takes approximately 266 days (median is 84 days) for one of these types of accounts to be banned.

The sample was utilized in order to test the proposed deception prevention method. The sample constitutes the most difficult set of identity deception cases that the subreddit community encountered. The accounts in the set managed to evade entry point requirements. In order to build a baseline for the machine learning models, these 419 accounts were paired with another 419 legitimate accounts which were already part of the giftcardexchange subreddit. The total number of legitimite accounts that were not banned and posted on giftcardexchange were 8026.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2016.2607697, IEEE Transactions on Information Forensics and Security

TABLE I
EXAMPLE OF POSTS MADE BY USER UP TO THE TIME THEY WERE BANNED ON GIFTCARDEXCHANGE

Date (YYYY-MM-DD HH:MM:SS)	Subreddit
2014-08-03 22:26:10	/r/giftcardexchange
2014-08-08 18:26:54	/r/pics
2014-08-08 20:47:51	/r/pics
2014-08-08 21:21:53	/r/giftcardexchange
several lines omitted due to space limitations	[...]
2014-09-28 01:02:46	/r/REDDITEXCHANGE
2014-10-19 01:58:13	/r/giftcardexchange
2014-10-21 17:15:46	/r/giftcardexchange
2014-10-23 21:20:33	/r/giftcardexchange
2014-10-23 21:36:22	/r/giftcardexchange
2014-10-23 23:44:02	/r/giftcardexchange
2014-11-25 00:00:11	/r/hardwareswap
2014-11-26 07:15:02	/r/hardwareswap
2014-11-26 20:13:21	/r/DSLR

The method proposed in this study evaluates a particular user at the time they attempt to join a sub-community, which is perceived to be the current date. For testing purposes, all dates were adjusted in order to simulate the time when accounts attempted to join the sub-community of interest. Simply put, for each one of the 838 accounts in the sample, Algorithm 1 was used to generate a network at a time set by the daterestriction optional parameter. This procedure is an especially important part of building a proper training set. Once the training set is built, the time utilized for every new user that attempts to join a community is expected to be the current date.

The dates that could be set for the training dataset varied for the legitimate accounts and banned users. Since conceptually, deception prevention attempts to evaluate and render a decision for a user at the time of attempted entry in a community, the time utilized for the training set is expected to be the entry point. However, realistically there were two potential options for the banned accounts; the time right before when they initiated their first post on the sub-community (before they were banned) and the time right before they were banned. The latter option helps assess the efficiency of the method compared to human deception detection but also renders the method as identity deception detection. For comparative purposes both options were tested. Henceforth, the two time estimates will be referred to as T_{entry} and T_{banned} respectively. The optional parameter relating to date restriction on Algorithm 1 for legitimate users was set at the time right before the first post to the sub-community.

After executing the aforementioned procedure, each social network metric mentioned in the previous section was calculated for each user. For example, user A is a banned user and made their first post on giftcardexchange at T_{entry} time. Algorithm 1 is called using the following parameters: CommonContributionNetwork(A, T_{entry}). The result generates a snapshot of the sub-community's network with the banned user included in that snapshot. For user A (now a node in the network) the following are calculated:

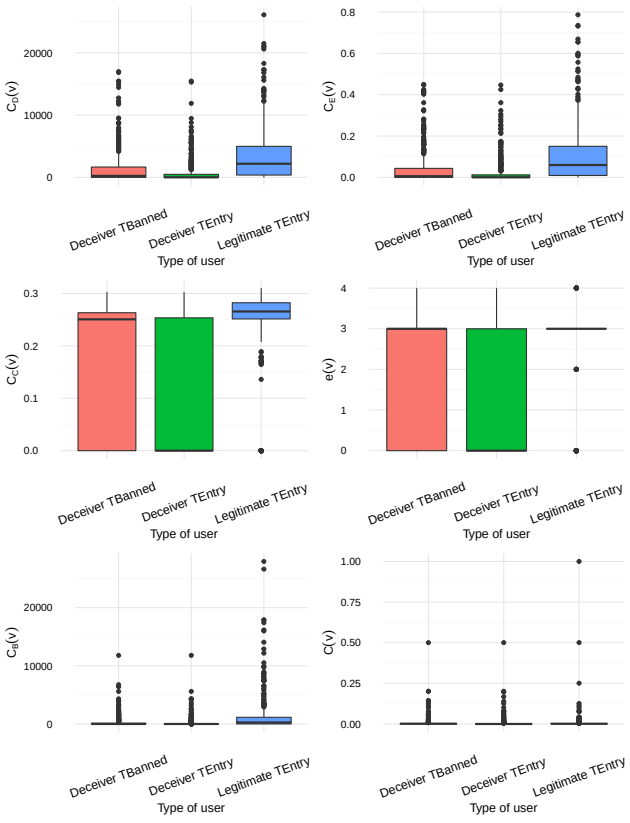

Fig. 2. Boxplots depicting the differences between banned accounts (deceivers) and legitimate users at T_{entry} (time of entry in the sub-community) and T_{banned} (time of being banned by the sub-community). Social network metrics include degree (C_D), closeness (C_C), betweeness (C_B), eigenvector centrality (C_E), eccentricity (e) and constraint (C).

$C_D(A), C_C(A), C_B(A), C_E(A)$, $e(A)$ and $C(A)$. Once the same procedure is applied for all users, the final sample contained legitimate and banned users along with their social network metrics in the particular network snapshot in time (T_{entry} or T_{banned}).

Fig. 2 depicts the differences between these metrics and respective groups of users. These were calculated for banned accounts (henceforth deceivers) at T_{entry} and T_{banned} as well as legitimate accounts at T_{entry}. It is evident that there is a visual difference between the group of banned users and the legitimate users. Deceivers potentially have a more difficult time constructing a common contribution network that relates well to existing members of the sub-community. This may be due to a lack of focus since banned accounts may have not originally intended to be part of the particular sub-community. Another reason is that deceivers may not have allowed enough time for connections to be generated. Further discussion on these results is offered at a later section of this paper.

Deceivers appear to have substantial differences in social network metrics compared to legitimate users shown on fig. 2. Metrics indicate that deceivers are overall further apart in terms of how they fit in the community, compared to legitimate users. The effect seems to be higher when measurements are made at the time of entry (during the first post in the sub-community) compared to the time a user was banned. This

is promising for utilizing the method for identity deception prevention.

Based on the above differences, different models were specified with the intent to identify further how differences in time affect performance of the deception prevention method. Mainly, the testing was separated into two versions of the deception prevention method: calculating metrics for banned users at the time they were banned and at the time right before they made the first post in the sub-community. The two versions were further split into two categories, where users of interest that were isolates were included and excluded in the training and testing of models. The different models that were used for evaluation are the following:

- MB Includes legitimate users at time of entry in the sub-community and banned users at the time they were banned. Users of interest with degree 0 are excluded from the sample.
- $MBwI$ Includes legitimate users at time of entry in the sub-community and banned users at the time they were banned. Users of interest with degree 0 are included in the sample.
- ME Includes legitimate users at time of entry in the sub-community and banned users at the time of entry in the sub-community. Users of interest with degree 0 are excluded from the sample.
- $MEwI$ Includes legitimate users at time of entry in the sub-community and banned users at the time of entry in the sub-community. Users of interest with degree 0 are included in the sample.

All models involved the same social network metric variables described in this paper. Nevertheless, sample sizes varied. For models MB and ME, the numbers of legitimate and banned users for training and testing the models varied due to some users being isolates at the time of entry in the sub-community's common contribution network. The final samples that were used for training and testing of models were 560 (377 legitimate and 183 deceivers) for ME and 650 (377 legitimate and 273 deceivers) for MB.

IV. Performance Evaluations

Models were tested using a set of supervised machine learning algorithms, which include: Support Vector Machine (SVM), Random Forest (RF) and Adaptive Boosting (ADA) and are considered ideal for models that involve binary outcomes [34]. Support Vector Machines build a representation of values in a set as points in space (the Cartesian coordinates plain is often used as an example) and then a clear gap is identified between distinct groups of points. Random Forests tackle the problem of binary classification using an ensemble of decision trees (e.g., if user has a degree < 10 then he or she is a deceiver). Different decision trees will result in different prediction efficiency. The ensemble of decisions made by a series of decision trees creates a more accurate and stable prediction model called Random Forest. Finally, Adaptive Boosting Algorithm is an algorithm (also referred to as a meta-heuristic) similar to the RF algorithm. It uses an ensemble of decision trees but it assigns weight on each case in the training data based on classification accuracy. Problematic cases that are mis-classified are boosted (having their weights increase). This way, the overall ADA model compensates for these difficult cases. More details about these machine learning algorithms can be found in literature (See [34], [35]).

TABLE II
Classification Matrix Used to Evaluate the Efficiency of a Model-Algorithm Pair

	Verified Identity Deception		**Verified Legitimate User**	
Predicted Identity Deception	True (TP)	Positive	False (FP)	Positive
Predicted Legitimate User	False (FN)	Negative	True (TN)	Negative

A. Performance Metrics

Model performance for the proposed deception prevention method was measured using a classification matrix as shown on Table II. Classification matrices are commonly used in order to develop a set of metrics that measure recall (the fraction of valid deceiver cases that are identified), precision (the fraction of deceiver cases that are identified and are valid), F-measure (a model's accuracy bounded between 0 and 1 and is derived from recall and precision), accuracy (the sum of true positive and true negatives that are identified out of the total cases), false positive rate (FPR) (fraction of falsely identified deceiver cases) and Matthews Correlation Coefficient (MCC) (a machine learning performance metric bounded between 0 and 1 that is often considered more rigorous in cases in which sample size varies). Formal definitions for these metrics are provided below [36]:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

$$Accurasy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

B. Testing Procedure

Model testing was conducted using a repeated ten times ten-fold cross-validation procedure based on the mean values for all aforementioned performance metrics. Algorithm 2 includes the procedure used for model testing (also utilized in a

TABLE III
PERFORMANCE RESULTS FOR SVM MODELS.

Model	Precision	Recall	F-meas.	Accuracy	FPR	MCC
ME	0.58	0.06	0.13	0.67	0.03	0.10
$MEwI$	0.82	0.59	0.68	0.73	0.13	0.48
MB	0.55	0.55	0.55	0.62	0.33	0.23
$MBwI$	0.65	0.71	0.67	0.66	0.39	0.32

TABLE IV
PERFORMANCE RESULTS FOR RF MODELS.

Model	Precision	Recall	F-meas.	Accuracy	FPR	MCC
ME	0.48	0.35	0.40	0.66	0.19	0.18
$MEwI$	0.72	0.72	0.72	0.72	0.28	0.44
MB	0.57	0.52	0.54	0.63	0.29	0.24
$MBwI$	0.66	0.69	0.67	0.66	0.34	0.32

TABLE V
PERFORMANCE RESULTS FOR ADA MODELS.

Model	Precision	Recall	F-meas.	Accuracy	FPR	MCC
ME	0.50	0.35	0.41	0.67	0.17	0.20
$MEwI$	0.74	0.71	0.73	0.73	0.25	0.46
MB	0.56	0.49	0.51	0.62	0.28	0.21
$MBwI$	0.66	0.66	0.65	0.66	0.34	0.32

previous study [5]) using a random forest algorithm as an example. The method splits the dataset into ten equal segments. Then, nine segments are used for building the model while the last segment is used for prediction testing. The separate segment is considered previously unseen data for the machine learning algorithm and as such results obtained are considered to be more reliable. The procedure is repeated ten times. Additionally, the dataset is split ten times using a different seed number. The total mean approximates the performance of the models and as a consequence the performance of the deception prevention method in unseen data. As such, reliability of the results is expected to be high [34].

Algorithm 2 A repeated ten times ten-fold cross-validation algorithm for testing a model using random forest.

procedure TENTIMESTENFOLDCROSSVALIDATION
▷ Algorithm builds a single model (e.g., RF) and produces final results
  $w \leftarrow$ predefined number
  for n in $T = [1, 2, ..., 9, 10]$ **do**
    $S \leftarrow w * n * 10$ ▷ Set random seed
    Create fold sample list FLi by randomly assigning fold numbers to the full length of dataset
    for f in $TT = [1, 2, ..., 9, 10]$ **do**
      Build Random Forest model RF based on training data (FLi not equal to f) and S
      Calculate predictions Pi using RF for testing data (FLi equal to f)
      Set Oi as observed values (is or is not a deceiver) based on testing data
      Build classification matrix using observed Oi and predicted Pi values
      Calculate Recall RE_f, Precision PR_f, and F-measure FM_f
    end for
    Calculate $RE_n = \sum_{f=1}^{10} \frac{RE_f}{10}$, $PR_n = \sum_{f=1}^{10} \frac{PR_f}{10}$, and $FM_n = \sum_{f=1}^{10} \frac{FM_f}{10}$
  end for
  Calculate $RE = \sum_{n=1}^{10} \frac{RE_n}{10}$, $PR = \sum_{n=1}^{10} \frac{PR_n}{10}$, and $FM = \sum_{n=1}^{10} \frac{FM_n}{10}$
end procedure

Test results are presented in tables III, IV and V for all models and machine learning algorithms that were used, rounded to three decimal digits.

A comparison between the proposed deception prevention technique and other techniques mentioned in the literature review is presented in Table VI. High-level efficiency is estimated based on the approaches taken by each method. These summaries demonstrate the superiority of the technique not just because of its proactive nature (being a prevention technique) but also due to its accuracy. In terms of the conceptual aspects of the approach, the "Friend or Foe" method can be seen as the logical extension of attempting to verify a user's relationship with a whole sub-community as opposed to just another user. The largest downside of the approach compared to the non-verbal approach presented for comparison is the computational overhead, which can increase substantially, as the number of users grow. However, the method is still more effective than other network methods (e.g., [27]) that may require a complete or near complete snapshot of the network rather than a sub-community's network which is substantially smaller.

V. DISCUSSION OF RESULTS

Both the deception prevention $MEwI$ and detection $MBwI$ models have yielded high accuracy results. However, recall and precision varied substantially between different models. Overall, accuracy rates were similar between the model when the time was set at the instance of entry in a sub-community ($MEwI$) and the time a deceiver was banned ($MBwI$). However, the same cannot be said for the models where isolates were not taken into account in the training process of the models. In fact, when isolates (users of interest with degree 0) are not considered, then precision and recall for models was substantially lower for MB and ME. Further, considering precision, recall and f-measure, the prevention model ($MEwI$) appears to be stronger than the detection model ($MBwI$). The opposite effect exists if isolates are removed (MB and ME). This is due to the likelihood that as people are allowed to enter the community and operate, they will likely also post in subreddits that other users post. As such, deceivers seize to be isolates MB performs better than ME.

A possible explanation for these differences is due to the behavior of many deceivers. As previous studies on online deception have indicated, time is an important factor for uncovering deception cases [5], [9]. The longer an individual

TABLE VI
COMPARISON OF IDENTITY DECEPTION PREVENTION TECHNIQUE WITH DECEPTION DETECTION TECHNIQUES.

	Non-Verbal Expectancy Violations Detection [5]	**Natural Language Processing Similarity Searching [6]**	**Friend or Foe Detection Method using Network Features [28]**	**Deception Prevention using Social Network Analysis (Proposed Method)**
Accuracy	71.3%	68.8%	69.6%	73%
Indicators used	Non-verbal	Verbal	Verbal	Verbal for building common contribution network
Limitations	Needs data on user activity	Needs user text	Applies only for two users	Needs design that allows for sub-communities
Efficiency	$\mathcal{O}(1 * R')$, R' limited amount of revisions ($R' < R$) made during observation window	$\mathcal{O}(N * R)$, N total number of users and R all revisions made by each user	$\mathcal{O}(P)$, P is the volume of activity made by two users	$\mathcal{O}(N * P)$, where N is number of users in sub-community and P all posts made up to date
Time of application	After a set time (e.g., 12 hours) window	As soon as a user posts a text somewhere	As soon as a friend request is active	At the time of attempted entry in the sub-community

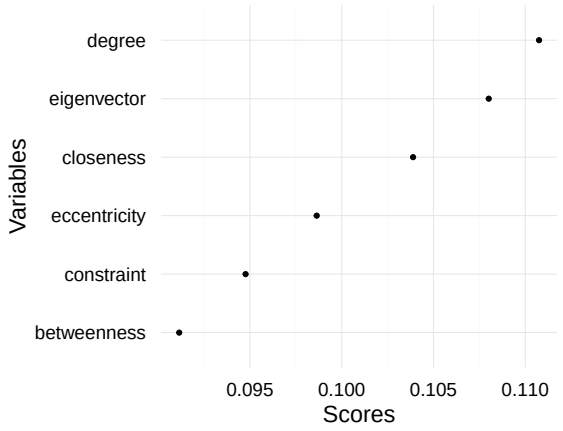

Fig. 3. Variable importance plot for adaptive boosting model MEwI illustrating the importance of social network metrics. Abbreviations are as follows: degree C_D, eigenvector_centrality C_E, closeness C_C, constraint C, betweenness C_B and eccentricity e.

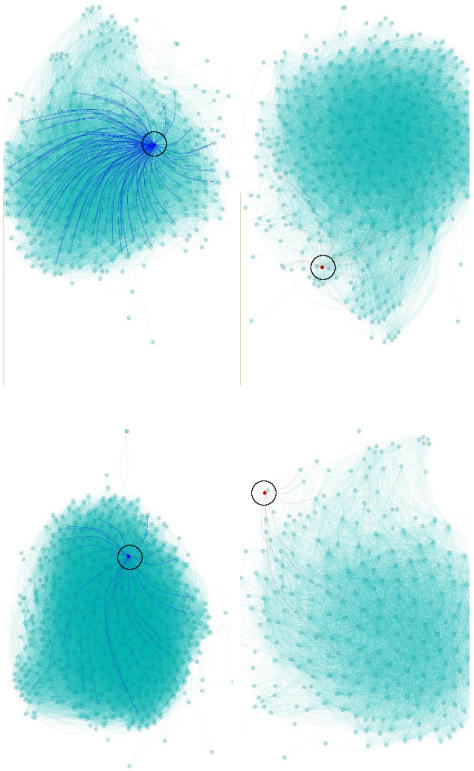

Fig. 4. Examples of common contribution networks for the sub-community at different times for legitimate (top-left, bottom-left) and banned (top-right, bottom-right) users. Users are identified with a black circle.

remains undetected, the more apparent his or her deviation from legitimate user behavior will appear. In other words, shared interests by the sub-community are not reflected by deceivers.

In an effort to understand how social network factors can influence how deceivers are detected, variable importance derived from ADA model for $MEwI$ (the model that yielded the best results) is provided at fig. 3. The graph has been generated by the function provided by the R package ada that utilizes the adaptive boosting algorithm [37]. Metric importance varies between the rest of the models. For example, MB tends to place the highest importance on the closeness metric C_C. The difference can be explained in the presence or absence of isolates, which can affect model statistics and for many users having a degree of 0 can be telling sign of deceptive accounts.

Further visual examination of networks can also explain what is demonstrated by the variable importance graph. Fig. 4 depicts two examples of social networks of banned accounts at the time of entry and two legitimate accounts at the time of entry. Legitimate users are represented in blue and have noticeably more connections with the rest of the sub-community at the time of attempted entry. They are also more central in the overall network's structure. Deceivers (depicted in red) have fewer connections and reside at the fringes of the common contribution network at the time of attempted entry.

Overall, legitimate users appear to have more connections to other users (higher C_D) compared to deceivers at the time of entry T_{entry}. In other words, users that are legitimate are expected to have more common interests with the rest of the sub-community's members. Legitimate users also appear to be more central to the network. Additionally, their position also distinguishes them as hubs of distribution of information based on the information centrality metric. Given that the network is not a communication network but rather a common "interest"

network, it is important to not oversimplify this information by determining that these individuals act as bridges of communication. Instead, this finding could be interpreted in the looser sense that these individuals are in the center of "things" in the larger Reddit community. As such, their bridge-like properties could be a reference to a diverse set of interests that make the sub-community wealthier. Further, their diversity could also be used as a metric of the novelty that they offer as a new users that joins a sub-community. Deceivers on the other hand not only have less in common with the rest of the sub-community at the time they join the group, but they also tend to have less diversity in the topics (subreddits) affiliated with their posts. This could be due to inactivity or due to narrow objectives (e.g., forging a new identity with the explicit interest to attack a sub-community).

The technique demonstrates that identity deception prevention using common contribution network data is feasible and reliable for communities (e.g., Reddit) that allow for the existence of sub-communities (e.g., giftcardexchange within Reddit). In theory, as long as formation of common contribution networks is possible, the technique can be applied to a number of social media platforms. This study has also demonstrated that such a deception prevention method is computationally feasible for networks that are not large. However, certain social network metrics are computationally intensive for large cases. For example, betweenness centrality metrics (C_B) often incur computational overheads at $\mathcal{O}(v)^3$, where v is the number of nodes in G.

From a theoretical perspective, the deception prevention method presented in this study can be thought to be possible on the basis of IDT, LT and EVT. Identity deception success is dependent on a deceiver's ability to "read" on the interactions with victims and attempt to adjust his or her behavior in order to hide the attempt for deception. In social networking terms this means establishing connections in order to appear legitimate. Nevertheless, since the network tested here is a common contribution network, it is unclear on whether a deceiver would cognitively attempt to fit well into the network. One can speculate that most deceivers attempt to look like legitimate users by generating posts in various subreddits. Generating a proper "fit" within an existing sub-community's social network is challenging for any user even if visualization of such a network is possible. The difficulty of achieving such a deceptive feat is also due to the nature of communities which are often found to have users with many secondary ties. Often this is translated as common interests between members of a community [38]. As such, a random order of posts in various subreddits would not result in an appearance that a user "fits" well with the rest of the community. The lack of common interest between a deceiver and normal users become apparent through the position of the deceiver in the common contribution network. In some ways, this can be thought of as a leakage cue.

The results of the method also demonstrate that deception prevention is possible by utilizing previously generated account data, as opposed to having stricter registration methods (e.g., asking for a telephone number). This is in line with a previous study [5] and demonstrates that not only EVT can be used as base theory for identity deception detection, but also identity deception prevention.

A. Limitations of Proposed Detection Method

The method proposed by this paper has several limitations that are dependent on the context of application. The main requirement is the need for sub-communities to exist within a larger community and users in principle to be allowed to participate to these sub-communities. Participation can be restrictive or non-restrictive. That is, users can contribute to sub-communities that are either moderated or non-moderated. The sub-community that wants to utilize the deception prevention method will have to be able to access the data on a user that attempts to join the sub-community. There is also a need for the sub-community to build a baseline model which would require a substantial number of legitimate as well as banned user cases. The method appears to be most effective as a deception prevention mechanism, but it can also be used as a detection method. However, computational efficiency is a serious limitation especially for large networks. Consequently, larger networks are probably the ones that would need the method the most. A way that the approach can perhaps be balanced is by trading off some of the more computationally intensive social network metrics with less computationally intensive. Even so, that will influence the accuracy of the prevention method. Nevertheless, the method is still superior in terms of its accuracy to previous identity deception detection methods [5], [6].

B. Future Work

Future work will need to examine the application of this deception prevention method in different social media contexts. Additionally, even though this study utilized common contribution networks in order to illustrate the deception prevention method, other networks may also prove to be effective. The degree of effectiveness may prove to be higher with other types of networks. A combination of various types of networks could also prove to be useful. Further, aside from standard social network metrics, other social network features should be tested as well as weighted and longitudinal networks. Finally, one of the important aspects relating to the feasibility of this method is identifying computationally efficient methods where networks and metrics can be more easily generated for the deception prevention method.

VI. Conclusion

The growth of social media applications appears to be occurring at an unpresented rate, which has resulted in an increased interest in identifying novel ways to battle identity deception and in particular identity forgery. However, despite the efforts and prevention techniques identified by designers, developers, and researchers, deception prevention has not been given enough attention. This study offers a novel attempt to demonstrate a computational approach to identity deception prevention by utilizing social network data and specifically a common contribution network. Past studies that utilized social

network data focused largely on detection and Sybil attacks [11], [12], [13] as opposed to cases of scamming that were present in the community that was examined in this paper. The approach of deception prevention yields high accuracy rates and has the potential to curve deceptive accounts for communities that exist on social media platforms. Given the increased number of incidents that arises with deceivers being able to generate an unlimited number of accounts, a proactive approach to identifying these accounts is essential for the long-term sustainability of "healthy" communities.

REFERENCES

[1] J. Brenner and A. Smith, "72 % of Online Adults are Social Networking Site Users groups continue to increase their engagement," PewResearchCenter, Tech. Rep., 2013. [Online]. Available: http://pewinternet.org/Reports/2013/social-networking-sites.aspx

[2] X. Shen, "Security and privacy in mobile social network [Editor's Note]," *IEEE Network*, vol. 27, no. 5, pp. 2–3, sep 2013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6616107

[3] J. P. Farwell, "The Media Strategy of ISIS." *Survival (00396338)*, vol. 56, no. 6, pp. 49–55, nov 2014.

[4] K. Bergstrom, ""Don't feed the troll": Shutting down debate about community expectations on Reddit.com," *First Monday*, vol. 16, no. 8, 2011.

[5] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 8, pp. 1311–1321, 2014.

[6] T. Solorio, R. Hasan, and M. Mizan, "A Case Study of Sockpuppet Detection in Wikipedia," in *Proceedings of the Workshop on Language Analysis in Social Media*, A. Farzindar, M. Gamon, M. Nagarajan, D. Inkpen, and C. Danescu-Niculescu-Mizil, Eds., no. Lasm. Stroudsburg, PA, USA: The Association for Computational Linguistics, 2013, pp. 59–68. [Online]. Available: http://www.aclweb.org/anthology/W13-1107

[7] G. Wang, H. Chen, J. Xu, and H. Atabakhsh, "Automatically detecting criminal identity deception: an adaptive detection algorithm," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 36, no. 5, pp. 988–999, 2006.

[8] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, J. F. Nunamaker, D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 36–43, 2005.

[9] M. Tsikerdekis and S. Zeadally, "Online Deception in Social Media," *Communications of the ACM*, vol. 57, no. 9, pp. 72–80, 2014.

[10] ——, "Detecting and Preventing Online Identity Deception in Social Networking Services," pp. 41–49, 2015.

[11] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *NSDI'12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, ser. NSDI'12. Berkeley, CA, USA: USENIX Association, 2012, p. 15.

[12] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An analysis of social network-based Sybil defenses," in *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 4. ACM, 2010, p. 363.

[13] N. Tran, J. Li, L. Subramanian, and S. S. M. Chow, "Optimal Sybil-resilient node admission control," pp. 3218–3226, 2011.

[14] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[15] P. Ekman, "Deception, lying, and demeanor," in *States of mind: American and post-Soviet perspectives on contemporary issues in psychology*, D. F. Halpern and A. E. Voiskounsky, Eds. Oxford University Press, 1997, pp. 93–105.

[16] D. B. Buller and J. K. Burgoon, "Interpersonal Deception Theory," *Communication Theory*, vol. 6, no. 3, pp. 203–242, aug 1996.

[17] J. Burgoon, M. Adkins, J. Kruse, M. Jensen, T. Meservy, D. Twitchell, A. Deokar, J. Nunamaker, Shan Lu, G. Tsechpenakis, D. Metaxas, and R. Younger, "An Approach for Intent Identification by Building on Deception Detection," *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 21a–21a, 2005. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1385270

[18] A. C. Squicciarini and C. Griffin, "An Informed Model of Personal Information Release in Social Networking Sites," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, sep 2012, pp. 636–645.

[19] S. Grazioli and S. L. Jarvenpaa, "Perils of Internet fraud: An empirical investigation of deception and trust with experienced Internet consumers," pp. 395–410, 2000.

[20] T. Madhusudan, "On a text-processing approach to facilitating autonomous deception detection," pp. 43–52, 2003. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1173789

[21] H. S. Park, T. Levine, S. McCornack, K. Morrison, and M. Ferrara, "How people really detect lies," *Communication Monographs*, vol. 69, no. 2, pp. 144–157, jun 2002.

[22] T. H. Feeley and M. J. Young, "Humans as lie detectors: Some more second thoughts," *Communication Quarterly*, vol. 46, no. 2, pp. 109–126, mar 1998.

[23] T. R. Levine, R. K. Kim, H. Sun Park, and M. Hughes, "Deception Detection Accuracy is a Predictable Linear Function of Message Veracity Base-Rate: A Formal Test of Park and Levine's Probability Model," *Communication Monographs*, vol. 73, no. 3, pp. 243–260, sep 2006.

[24] M. P. Ebenazer and P. Sumathi, "An Overview of Identity Deception Approaches and Its Effects," *International Journal of Computer Trends and Technology*, vol. 25, no. 3, pp. 123–126, 2015. [Online]. Available: http://www.ijcttjournal.org/2015/Volume25/number-3/IJCTT-V25P124.pdf

[25] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraudlent financial statements using linguistic credibility analysis," *Decision Support Systems*, vol. 50, no. 3, pp. 585–594, feb 2011.

[26] J. S. Donath, "Identity and deception in the virtual community," in *Communities in Cyberspace*, M. A. Smith and P. Kollock, Eds. Routledge, 1999, ch. 2.

[27] C. Yang, R. Harkreader, and G. Gu, "Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013.

[28] M. Fire, D. Kagan, A. Elyashar, and Y. Elovici, "Friend or foe? Fake profile identification in online social networks," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–23, 2014. [Online]. Available: http://dx.doi.org/10.1007/s13278-014-0194-4

[29] S. L. S. Lu, G. Tsechpenakis, D. Metaxas, M. Jensen, and J. Kruse, "Blob Analysis of the Head and Hands: A Method for Deception Detection," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, vol. 00, no. C. IEEE, 2005, pp. 1–10. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1385269

[30] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, and L. Michaelis, "Distinguishing deceptive from non-deceptive speech," in *Interspeech 2005*. Proceedings of Eurospeech'05, 2005, pp. 1833–1836.

[31] G. C. Kane, "It's a Network, Not an Encyclopedia: A Social Network Perspective on Wikipedia Collaboration." *Academy of Management Proceedings*, vol. 2009, no. 1, pp. 1–6, aug 2009.

[32] J. Gil and S. Schmidt, "The Origin of the Mexican Network of Power," in *Proceedings of the International Social Network Conference*, Charleston, 1996, pp. 22–25.

[33] S. P. Borgatti and M. G. Everett, "A Graph-theoretic perspective on centrality," *Social Networks*, vol. 28, no. 4, pp. 466–484, oct 2006.

[34] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*, ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.

[35] G. Williams, *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*, ser. Use R! Springer, 2011.

[36] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[37] M. Culp, K. Johnson, and G. Michailides, "ada: An r package for stochastic boosting," *Journal of Statistical Software*, vol. 17, no. 2, 2006.

[38] T. W. Howard, *Design to Thrive: Creating Social Networks and Online Communities That Last*. Burlington, MA, USA: Morgan Kaufmann, 2010.