Disambiguation of Identities using User Ranking in Social Networks

Mohnish Naidu
Department of Computer Science
Maulana Azad National Institute of Technology, Bhopal

Sri Khetwat Sarita
Department of Computer Science
Maulana Azad National Institute of Technology, Bhopal

ABSTRACT
An approach to disambiguate the identity information between a Sybil and a Non-Sybil by using the social circles and weighted graphs is being presented in this paper. The disambiguation is done in two phases the Static phase and the Dynamic phase. In Static phase, the social circles are generated through the extraction and pruning of social networks. In Dynamic phase, the initial parameters are further refined by the use of polling from the other users. The approach provides a methodology to monitor existing identity information, applicable to addressing issues like identity theft, online fraud and lateral surveillance.

General Terms
Sybil Defense, User Ranking, Social Networks, Web Security, Ontology, Semantic Web.

Keywords
Sybil Defense; Pagerank; Semantic Web; Social network; User Rank.

1. INTRODUCTION
Online social networking sites are gaining popularity day by day. Online social networks represent a new kind of information network that differs significantly from existing networks like the Web. For example, in the Web, hyperlinks between content form a graph that is used to organize, navigate, and rank information. The properties of the Web graph have been studied extensively, and have lead to useful algorithms such as PageRank. In contrast, few links exist between content in online social networks and instead, the links exist between content and users, and between users themselves. However, little is known in the research community about the properties of online social network graphs at scale, the factors that shape their structure, or the ways they can be leveraged in information systems.

Users join a network, publish their own content, and create links to other users in the network called "friends". This basic user-to-user link structure facilitates online interaction by providing a mechanism for organizing both real-world and virtual contacts, for finding other users with similar interests, and for locating content and knowledge that has been contributed or endorsed by "friends".

A disambiguation of identities using user ranking in social networks is presented in this paper. The approach provides a methodology to monitor existing identity information, applicable to addressing issues like identity theft, online fraud and lateral surveillance.

2. RELATED WORK
Traditional defense against Sybil attack are based on central certification authority which checks and certifies each node as in [9, 10]. The cost of monitoring each node individually is high, hence the use of various decentralized approaches are given in [1, 2, 3, 4, 11]. Most of these approaches are based on the fact that the Sybil Nodes are loosely connected in the network with the Non-Sybil Nodes and the Non-Sybil nodes are tightly bounded.

An approach in [5], provides a method of node ranking for the Sybil defense. The nodes that have better connectivity to the trusted node are ranked higher and are seems to be a more trustworthy. This defense scheme runs on ranking node similarity. The main approaches have been provided in [11][12] to defend against Sybil attacks.

A method using random walks is used to defend against Sybil attack is given in SybilGuard and SybilLimit [3, 4]. In SybilLimit, an honest node can certify other nodes as "probably honest", accepting at most O(log n) Sybil identities per attack edge. SybilLimit also uses a voting procedure to find the likely hood of a node being honest. An attempt for community detection by finding clusters of nodes directly could be used for Sybil defense [6]. An approach is given in [4], of detecting the Sybil by the use of random walks from the trusted nodes, but in this we have to assume a group of Non-Sybil nodes.

3. PROPOSED ARCHITECTURE
The architecture lays on the concept of ranking system as referenced in [9] which could be used to disambiguate the Sybil and Non-Sybil. Each user's are identified by their individual ranks and the privileges are given to them accordingly. There are two phases for providing ranking of the users. The first phase, static phase, gives the ranking based on the initial parameters provided by the users. The information is verified by the MSG (Matcher and the Score Generator), against the information collected from the web by the Web Mining Engine. In the second phase, dynamic phase, the value changes according to the poll results from the users of the social circles. For example if the user wants to expand his social circle, user sends a friend request. The polling depends on the receiver whether he accepts the request or not. The poll result helps to change the users' rank value.

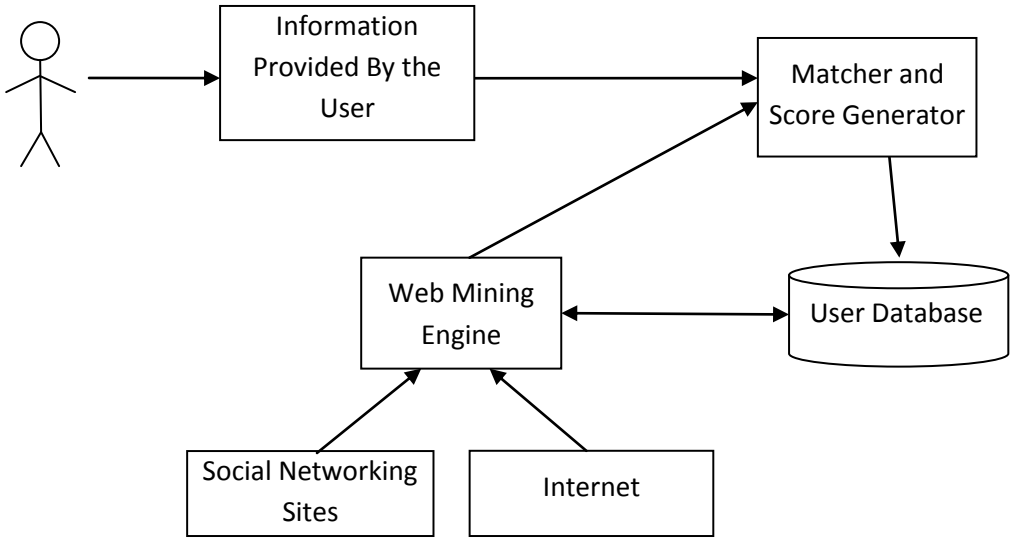

Figure 1 User Ranking Architecture

The change will be proportional to the rank value of the voter i.e. if the rank value of the voter is high than the change is measurable else negligible for the low ranked voter. Architecture is being shown in figure 1.

The Web Mining Engine (WME) is used to prune the data available from the various social networking sites, web and communities etc. for a particular user. It acts as a web crawler which extracts the information available on the web and the network. This information is filtered and essential and relevant data is pruned out.

The MSG (Matcher and the Score Generator) is used to match the information provided by the user and the WME. The relevant data can be obtained by the relative match of data obtained from the web and some from basic heuristics like qualification according to the age, friends in the circle sharing same communities like school, places, graduation school, interests etc. The match results are then used to generate a score value, which is the initial value associated with the user.

3.1 Static Ranking

The social web can be used to extract the information about a person's identity from social networking sites, web, communities etc [7]. As shown in figure 1, the user provides the information like academics, basic information about him and other social interests. Depending upon this, more information regarding the user is mined by Web Mining Engine later the information from these two sources are matched at the MSG to get an initial parameter which are stored in user's database along with the rank value calculated.

3.2 Dynamic Ranking

Dynamic ranking is used to further refine the initial rank value provided by the static phase. When the user send a friend request, the requested person has an option of accept or deny of friendship in addition to this he will have to poll for the question that "Whether the person is genuine or not?". The poll given by the person is used to increment or decrement the rank value of the user.

The change in rank value of the user is directly proportional to the rank of the voter i.e. if the rank of the voter is high then change in the ranking will be more and if rank of the voter is low then the change in the rank will be low or sometimes negligible.

3.3 Rank Value Calculation

The calculation of the rank is based on the acceptance or rejection of the friend request as well as the positive or negative poll results. The possible cases which are possible are discussed below.

3.3.1 Friendship Accepted and Positive Poll

In this case, the requested person has accepted the friend request and verified the user as genuine, and then the friend count is incremented to one.

$$F_R = F_R + 1$$

A s the friend circle count is increased then the new increased rank value calculated as

$$R_R = \left\{ \frac{R_R * (F_{RRC} + F_R - 1) + R_V + 3.5}{F_R + FRRC} \right\}$$

Where,

FR = Total friends count in the circle of the user
RR = Rank of the user/requester
RV = Rank of the voter/requested node.
FRRC = Friend request rejection count

3.3.2 Friendship Accepted and Negative Poll

This case is same as the previous case except that the requested node has given the negative poll (or verified the user as the fake). So, instead of increasing the rank value it is decremented. The friend count is incremented as the new connection is added in the social circle and the new rank value is calculated.

$$F_R = F_R + 1$$

$$R_R = \left\{ \frac{R_R * (F_R + F_{RRC} - 1) + R_V - 3.5}{F_R + FRRC} \right\}$$

3.3.3 Friendship Rejected and Positive Poll

This is the case in which the friend request is being rejected by the voter but, he has verified the requester as a genuine. So in this case no changes are done to the user rank value as the same node could be used to increase the rank value of a user by continuously rejecting the request and providing positive poll.

3.3.4 Friendship Rejected and Negative Poll

This is the case in which the requested person has rejected the friend request but, given a positive poll. In this case we decrement the rank value and increment the count of FRRC which is used to count all those nodes which are not in the circle of the user but there poll has been counted to calculate the rank value of the user.

$$FRRC = FRRC + 1$$

$$R_R = \left\{ \frac{R_R * (F_R + F_{RRC}) + R_V - 3.5}{F_R + FRRC + 1} \right\}$$

The poll count is considered in this case but not in the previous case, as it would decrement the rank value of a user. If a user continuously sending the friend request to the same node which is rejecting the friend request along with the negative poll then their Rank value of the requesting user will fall at each rejection from that node. The low rank value will provides less privileges' to the user hence sending friend request to these nodes will be avoided.

3.3.5 Friend Removed from Friend List

If any friend is removed from the friend list then we have to decrement the friend count and the rank of the user whose friend being removed will also have to suffer by a proportional decrement in rank according to the friend being removed

$$F_R = F_R - 1$$

$$FRRC = FRRC + 1$$

$$R_R = \left\{ \frac{R_R * (F_R + F_{RRC}) + R_V - 3.5}{F_R + FRRC + 1} \right\}$$

4. EXPERIMENT ANALYSIS AND RESULT DISCUSSION

4.1 Static Phase :

Initially data is retrieved from the WME by using various available API's of social networking sites like LinkedIn, Facebook, Twitter, Google [19, 20, 21, 22]. These API's provides us the information and list of the used identities related to a particular user like first-name, last-name, and location, industry, current-status, network, connections, specialties, honors, interests, positions, skills, certification, education, courses etc in XML format which could be stored temporarily till the MSG calculates the initial value. Before initiating the request to retrieve the information each API needs an authentication.

The MSG access the XML information available in the temporary files generated by the WME for each user identity. It also contains the information given by user at the time of registration, which is cross verified against the crawled data. The score is generated after the comparison and the amount of information provided by the user. This score is set as the initial rank value of the user in the database.

4.2 Dynamic Phase :

This phase recalculates the rank value of the given user each time when it gets the response for the friend request based on the various cases discussed in section III. A table named USER_RANK is used to store the parameter values of dynamic phase for each user and is updated every time when the rank value is recalculated. The threshold value is so chosen in accordance with the static phase (here threshold value is taken as 3.5).

Table 1. Structure of USER_RANK table

Column Name	Type
USER_ID	INT
RANK_VALUE	FLOAT
FRIEND_COUNT	INT
FRIEND_REQUEST_REJECT_COUNT	INT

The experiments were performed on 46 nodes (23 Sybil Nodes and 23 Non-Sybil Nodes). The figure 2 shows the number of nodes detected as Sybil after each interval of time. After the initial ranking, 12 amongst 23 nodes are detected as Sybil whereas due to lack of information provided by the Non-Sybil nodes, 5 nodes from the Non-Sybil nodes are also detected as the Sybil nodes. The data is collected after every 5 days which shows the role of dynamic phase to improve the detection mechanism.

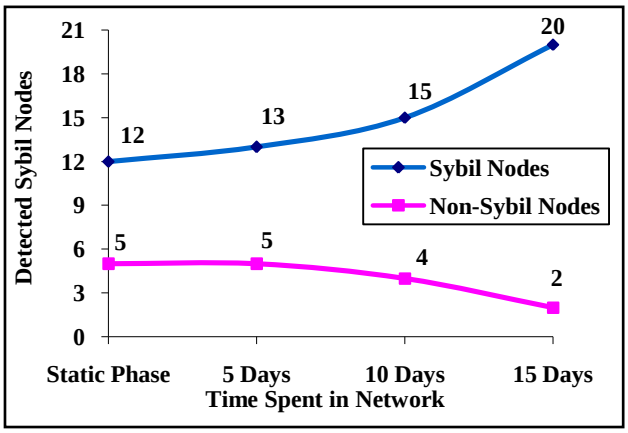

Figure 2 Sybil Identities detected over time interval

At the end of 15 days the precision is improved over 0.62 to 0.87. Figure 3 depicts the precision, recall, accuracy and F-factor of the system over the time intervals.

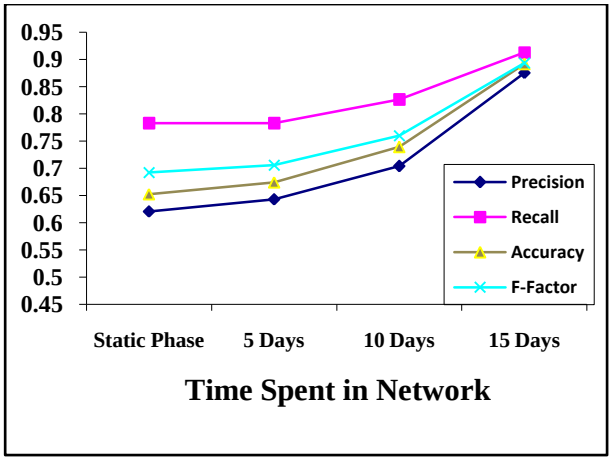

Figure 3 Performance Measurement

5. CONCLUSION

The approach provides us a way to measure the value of trust in the Social Networks. This value could be used to disambiguate between the Sybil and Non-Sybil Identities. The value also determines the reliability of any identity in the network. The privileges are granted or revoked according to their rank values. The user ranking system also solves the problem of Identity theft in the social network and providing only a single identity to each user. And hence the problems of online fraud by the identity theft could be solved.

6. REFERENCES

[1] G. Danezis and P. Mittal. SybilInfer: Detecting Sybil Nodes using Social Networks. In Proc. NDSS'09, San Diego, CA, Feb 2009.

[2] N. Tran, B. Min, J. Li, and L. Subramanian. Sybil-Resilient Online Content Voting. In Proc. NSDI'09, Boston, MA, Apr 2009.

[3] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks. In Proc. IEEE S&P, Oakland, CA, May 2008.

[4] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard: Defending Against Sybil Attacks via Social Networks. In Proc. SIGCOMM'06, Pisa, Italy, Sep 2006.

[5] B. Vishwanath, K.P. Gummadi, A. Post, Alan Mislove : An Analysis of Social Network-Based Sybil Defeses , sep 2010.

[6] S. Fortunato. Community detection in graphs. Physics Reports, 486:75, 2010.

[7] M Rowe, Fabio C. Disambiguating Identity through Social Circles and Social Data.

[8] L Zhang, Z Qin, The improved Pagerank in web crawler.

[9] M. Castro, P. Druschel, A. J. Ganesh, A.I.T. Rowstron, and D.S. Wallach. Secure Routing for Structured Peer-to-Peer Overlay Networks. OSDI, Boston, MA, Dec. 2002.

[10] R. Rodrigues and B. Liskov. Rosebud: A Scalable Byzantine-Fault-Tolerant Storage Architecture. MIT CSAIL, Technical Report TR/932, Dec. 2003.

[11] J.R. Douceur. The Sybil Attack. IPTPS, Cambridge, MA, Mar. 2002.

[12] B.N. Levine, C. Shields, and N.B. Margolin. A Survey of Solutions to the Sybil Attack. University of Massachusetts Amherst, Amherst, MA, 2006.

[13] K. Thirunarayan and P. Anantharam. Trust Networks: Interpersonal, Sensor, and Social. Ohio Center of Excellence in Knowledge-enabled Computing.

[14] M. Rowe, F. Ciravegna. Disambiguating identity web references using Web 2.0 data and semantics. Web Semantics: Science, Services and Agents on the WWW8(2010) 125-142.

[15] A. Cheng, E. Friedman. Manipulability of Page Rank under Sybil Strategies.

[16] S.Qiao, T. Li, H. Li, Y. Zhu, J. Peng, J. Qiu. Simrank: A Page Rank Approach based on Similarity Measure.

[17] L. Zhang, Z. Qin. The improved Page Rank in Web Crawler. ICISE 2009.

[18] A.K.Pal, D.Nath,S. Chakraborty. A Discriminatory Rewarding Mechanism For Sybil Defense with Application to Tor. WASET 63.ET 63, 2010.