Preventing Private Information Leakage from Social Networks

K.Nagendran
Assistant Professor,
Selvam College of Technology, Namakkal, Tamilnadu

Abstract — In subsequent releases we intend to incorporate additional capabilities to capture ancillary threat models. From our initial results, we quantify the privacy risk attributed to friend relationships in Facebook. We show that for each user in our project a majority of their personal attributes can be derived from social contacts. This result denoting the number of friends contributing to a correctly inferred attribute. These unwanted inferences are related to the users' identity, current location and other personal information. After analyzing the problem and reviewing our risk estimation method, decision tree is needed to distinguish between high risk and normal situations.

Index Terms — Security, Mainteanance, Service

INTRODUCTION

DOMAIN INTRODUCTION

Online social networking sites like Orkut, YouTube, and Flickr are among the most popular sites on the Internet. Users of these sites form a social network, which provides a powerful means of sharing, organizing, and finding content and contacts. The popularity of these sites provides an opportunity to study the characteristics of online social network graphs at large scale. Understanding these graphs is important, both to improve current systems and to design new applications of online social networks. This project presents a large-scale measurement study and analysis of the structure of multiple online social networks. Data gathered from four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut. Crawled the publicly accessible user links on each site, obtaining a large portion of each social network's graph. Our data set contains over 11.3 million users and 328 million links. This is the first study to examine multiple online social networks at scale. This project results confirm the power-law, small-world, and scalefree properties of online social networks.

The degree of user nodes tends to match the outdegree; that the networks contain a densely connected core of high-degree nodes; and that this core links small groups of strongly clustered, low-degree nodes at the fringes of the network. The implications of these structural properties for the design of social network based systems. The Internet has spawned different types of information sharing systems, including the Web. Recently, online social networks have gained significant popularity and are now among the most popular sites on the Web. For example, MySpace (over 190 million users), Orkut (over 62 million), LinkedIn (over 11 million), and LiveJournal (over 5.5 million) are popular sites built on social networks.Unlike the Web, which is largely organized around content, online social networks are organized around users. Participating users join a network, publish their profile and (optionally) any content, and create links to any other users with whom they associate. The resulting social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users. An in-depth understanding of the graph structure of online social networks is necessary to evaluate current systems, to design future online social network based systems, and to understand the impact of online social networks on the Internet. For example, understanding the structure of on-line social networks might lead to algorithms that can detect trusted or influential users, much like the study of the Web graph led to the discovery of algorithms for finding authoritative sources in the Web. Moreover, recent work has proposed the use of social networks to mitigate email spam, to improve Internet search, and to defend against Sybil attacks. However, these systems have not yet been evaluated on real social networks at scale, and little is known to date on how to synthesize realistic social network graphs.

In this project a large-scale (11.3 million users, 328 million links) measurement study and analysis of the structure of four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut. Data gathered from multiple sites enables us to identify common structural proper ties of online social networks. The data obtained by crawling publicly accessible information on these sites, and make the data available to the research community. In contrast, previous studies have generally relied on propreitary data obtained from the operators of a single large network .In addition to validating the power-law, small-world and scale-free properties previously observed in offline social networks. A high degree of reciprocity in directed user links, leading to a strong correlation between user indegree and outdegree. This differs from content graphs like the graph formed by Web hyperlinks, where the popular pages (authorities) and the pages with many references (hubs) are distinct.

Online social networks contain a large,strongly connected core of high-degree nodes, surrounded by many small clusters of low-degree nodes. This suggests that high-degree nodes in the core are critical for the connectivity and the flow of information in these networks. More specifically, the properties of the large weakly connected component 2 (WCC) in the user graphs of four popular sites. The entire user community (which would include users who do not use the social networking features), information flow, workload, or evolution of online social networking sites. While these topics are important, they are beyond the scope of this project.

OVERVIEW OF PROJECT

Online social networks have existed since the beginning of the Internet. For instance, the graph formed by email users who exchange messages with each other forms an online social network. However, it has been difficult to study this network at large scale due to its distributed nature.Popular online social networking sites like Flickr, You-Tube, Orkut, and LiveJournal rely on an explicit user graph to organize, locate, and share content as well as contacts. In many of these sites, links between users are public and can be crawled automatically to capture and study a large fraction of the connected user graph. These sites present an opportunity to measure and study online social networks at a large scale.

ONLINE SOCIAL NETWORKING SITES

Online social networking sites are usually run by individual corporations (e.g. Google and Yahoo!), and are accessible via the Web Users. To participate fully in an online social network, users must register with a site, possibly under a pseudonym. 3 Some sites allow browsing of public data without explicit sign-up. Users may volunteer information about themselves (e.g., their birthday, place of residence, or interests), which is added to the user's profile Links. The social network is composed of user accounts and links between users. Some sites (e.g. Flickr, LiveJournal)allow users to link to any other user, without consent from the link target. Other sites (e.g. Orkut, LinkedIn) require consent from both the creator and target before a link is created connecting these users. Users form links for one of several reasons. The nodes connected by a link can be real-world acquaintances, online acquaintances, or business contacts; they can share an interest; or they can be interested in each other's contribute d content. Some users even see the acquisition of many links as a goal in itself.

User links in social networks can serve the purpose of both hyperlinks and bookmarks in the Web. A user's links, along with her profile, are visible to those who visit the user's account. Thus, users are able to explore the social network by following user-to-user links, browsing the profile information and any contributed content of visited users as they go. Certain sites, such as LinkedIn, only allow a user to browse other user accounts within her neighborhood (i.e. a user can only view other users that are within two hops in the social network); other sites, including the ones, allow users to view any other user account in the system. Groups. Most sites enable users to create and join special interest groups. Users can post messages to groups and upload shared content to the group. Certain groups are moderated; admission to such a group and postings to a group are controlled by a user designated as the group's moderator. Other groups are unrestricted, allowing any member to join and post messages or content. Social networking sites are the portals of entry into the Internet for many millions of users, and they are being used both for advertisement as well as for the ensuing commerce. Many of
these applications, ranging from mail to auctions,implicitly rely on some form of trust .

For example, when a user accepts email from an unknown user, she trusts the other party not to send spam. When a user selects a winning bidder in an auction, she is trusting the other party to pay the winning amount, and the winning user is trusting the seller to produce the auctioned item.In a social network, the underlying user graph can potentially be used as a means to infer some level of trust in an unknown user, to check the validity of a public key certificate, and to classify potential spam. In all of these, trust is computed as a function of the path between the source and target user. Our findings have interesting implications for trust inference algorithms. The tight core coupled with link reciprocity implies that users in the core appear on a large number of short paths. Thus, if malicious users are able to penetrate the core, they can skew many trust paths (or appear highly trustworthy to a large fraction of the network).

However, these two properties also lead to small path lengths and many disjoint paths, so the trust inference algorithms should be adjusted to account for this observation. In particular, given our data, an unknown user should be highly trusted only if multiple short disjoint paths to the user can be discovered. The correlation in link degrees implies that users in the fringe will not be highly trusted unless they form direct links to other users. The "social" aspect of these networks is selfreinforcing: in order to be trusted, one must make many "friends", and create many links that will slowly pull the user into the core.

IMPLEMENTATION

In order to protect privacy, it sanitize both trait (e.g., deleting some information from a user's on-line profile) and link details (e.g., deleting links between friends) and explore the effect they have on combating possible inference attacks. Our initial results indicate that just sanitizing trait information or link information may not be enough to prevent inference attacks and comprehensive sanitization techniques that involve both aspects are needed in practice. While the specific accuracy reduction is varied by the number of details removed and by the specific algorithm used for classification, across a broad range of classifiers. The linear regression is affected the least, with approximately a 10 percent reduction in accuracy. To our knowledge this is the first comprehensive paper that discusses the problem of inferring private traits using real-life social network data and possible sanitization approaches to prevent such inference.

First, the modification of Naive Bayes classification is suitable for classifying large amount of social network data. Our modified Naive Bayes algorithm predicts privacy sensitive trait information using both node traits and link structure. Compare the accuracy of our learning method based on link structure against the accuracy of our learning method based on node traits. Please see extended version of this paper for further details of our modified Naive Bayes classifier there are a limited number of "groups" that are highly indicative of an individual's political affiliation. When removing details, these are the first that are removed.

PROBLEM DESCRIPTION

Security protection: It handles the following aspects :

- Availability (services are available to authorized users).
- Integrity (information is free from unauthorized manipulation).
- Confidentiality (only an intended user can access the respective information).
- Accountability (actions of any entity should be uniquely traceable).
- Assurance (guarantee that all security measures have been properly implemented).

The inference problem is mostly known as a security risk targeting system-based confidentiality. Two types of techniques have been proposed to identify and remove inference channels. One makes use of semantic data modeling methods to locate inference channels in the database design, in order to redesign the database for the removal of these channels. The other one evaluates database queries to understand whether they lead to unauthorized inferences. These techniques have been studied for statistical databases, multilevel secure databases and general purpose databases. A few researchers have also addressed the inference problem for data mining. Denning and Morgenstern employed classical information theory to measure the inference chance in the realm of multilevel databases. Our approach adapts their work for social computing environments.

The challenge of social inferences cannot be addressed adequately enough by existing techniques because of the following reasons:

(a) an inferred user attribute may not be stored in the social application database;
(b) background knowledge available to the inferrer outside of this database is often the premise for inferences;
(c) information revealed in the past through this application can enable historical inferences;
(d) the sensitivity of user information may be of dynamic nature;
(e) social inferences do not necessarily result from deductive reasoning.

Our theoretical framework is based on this fact: project collect more information about a user, such as his/her contextual situation, our uncertainty about other attributes, such as his/her identity may be reduced; consequently, this process increases the probability of our correctly guessing some user attributes. This uncertainty can be measured by information entropy. Information, as used in information theory for telecommunications, is a measure of the decrease of uncertainty in a signal value at the receiver site. In the realm of the inference problem under study, as the inferrer collects more information about an entity or attribute (such as another user or a location), the number of possible entities that match known sets of attributes decreases; this results in reduced information entropy.

PERFORMANCE EVALUATION

To probably infeasible in maintaining the use of social networks None of the remaining traits are as highly indicative as the initial two, instead see a gradual decrease in the accuracy over the tested parameters. Unsurprisingly, the Links Only classifier is only slightly affected by the removal of traits. Report the additional experimental results that show the impact of link removal, collective inference and varying labeled vs unlabeled nodes ratios.

It handles the following aspects

• Availability (services are available to authorized users).
• Integrity (information is free from unauthorized manipulation).
• Confidentiality (only an intended user can access the respective information).
• Accountability (actions of any entity should be uniquely traceable).
• Assurance (guarantee that all security measures have been properly implemented).

CONCLUSION

Our research benefits from these works and many others. However, the opinion that our work provides novel contributions to the field. Consider a wider array of attributes which are not limit in this research to a set of structured data types. For example, derive inferences based on values such as employer and educational institution. Consequently, forced to address challenging problems such as data disambiguation and named entity recognition .It present solutions to help mitigate the results from our findings. The different algorithms to reduce information loss, explore their corresponding runtime complexities, and suggest actions to users to reduce their privacy risk. When the results are combined from the collective inference implications with the individual results, are begin to see that removing details and friendship links together is the best way to reduce classifier accuracy. This is probably infeasible in maintaining the use of social networks.Project greatly reduce the accuracy of local classifiers, which give us the maximum accuracy that are able to achieve through any combination of classifiers.

REFERENCES

[1] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava,"Anonymizing Social Networks," Technical Report 07-19, Univ.of Massachusetts Amherst, 2007.

[2] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08), pp. 93-106, 2008.

[3] J. He, W. Chu, and V. Liu, "Inferring Privacy Information from Social Networks," Proc.Intelligence and Security Informatics,2006.

[4] E. Zheleva and L. Getoor, "Preserving the Privacy of Sensitive Relationships in Graph Data," Proc. First ACM SIGKDD Int'l Conf. Privacy, Security, and Trust in KDD, pp. 153-171, 2008.

[5] R. Gross, A. Acquisti, and J.H. Heinz, "Information Revelation and Privacy in Online Social Networks," Proc. ACM Workshop Privacy in the Electronic Soc. (WPES '05), pp. 7180,http://dx.doi.org/10.1145/1102199.1102214, 2005.

[6] H. Jones and J.H. Soltren, "Facebook: Threats to Privacy," technical report, Massachusetts Inst. of Technology, 2005.

[7] P. Sen and L. Getoor, "Link-Based Classification," Technical Report CS-TR-4858, Univ. of Maryland, Feb. 2007.

[8] B. Tasker, P. Abbeel, and K. Daphne, "Discriminative Probabilistic Models for Relational Data," Proc. 18th Ann. Conf. Uncertainty in Artificial Intelligence (UAI '02), pp. 485-492, 2002.

[9] A. Menon and C. Elkan, "Predicting Labels for Dyadic Data," Data Mining and Knowledge Discovery, vol. 21, pp. 327-343, 2010.

[10] E. Zheleva and L. Getoor, "To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private user Profiles," Technical Report CS-TR-4926, Univ. of Maryland,College Park, July 2008.

[11] N. Talukder, M. Ouzzani, A.K. Elmagarmid, H. Elmeleegy, and M. Yakout, "Privometer: Privacy Protection in Social Networks,"Proc. IEEE 26th Int'l Conf. Data Eng. Workshops (ICDE '10), pp. 266-269, 2010.

[12] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Inferring Private Information Using Social Network Data,"Proc. 18th Int'l Conf. World Wide Web (WWW), 2009.