RESEARCH ARTICLE OPEN ACCESS

A Two-stage Deanonymization Attack Against Anonymized Social Networks

Ms.K.Manasa
M.Tech scholar, CMR Engineering College

Ms.K.PriyaDarshini
Asst.Professor, Dept. of CSE, CMR Engineering College

Abstract—Digital traces left by users of online social networking services, even after anonymization, are susceptible to privacy breaches. This is exacerbated by the increasing overlap in user-bases among various services. To alert fellow researchers in both the academia and the industry to the feasibility of such an attack, we propose an algorithm, Seed-and-Grow, to identify users from an anonymized social graph, based solely on graph structure. The algorithm first identifies a seed sub-graph, either planted by an attacker or divulged by a collusion of a small group of users, and then grows the seed larger based on the attacker's existing knowledge of the users' social relations. Our work identifies and relaxes implicit assumptions taken by previous works, eliminates arbitrary parameters, and improves identification effectiveness and accuracy. Simulations on real-world collected datasets verify our claim.

Index Terms—social networks, anonymity, privacy, attack, graph.

I.Introduction:

Internet-based social networking services are prevalent in modern societies: a lunch-time walk across a uni-versity campus in the United States provides enough evidence. As Alexa's Top 500 Global Sites statistics (re-trieved on May 2011) indicate, Facebook and Twitter, two popular online social networking services, rank at 2nd and 9th place, respectively.

One characteristic of online social networking services is their emphasis on the users and their connections, in addition to the content as seen in traditional Inter-net services. Online social networking services, while providing convenience to users, accumulate a treasure of user-generated content and users' social connections, which were only available to large telecommunication service providers and intelligence agencies a decade ago.

Online social networking data, once published, are of great interest to a large audience: Sociologists can verify hypotheses on social structures and human behavior patterns; third-party application developers can produce value-added services such as games based on users' contact lists; advertisers can more accurately infer a user's demographic and preference profile and can thus issue targeted advertisements. As the December 2010 re-vision of Facebook's Privacy Policy phrases it: "We allow advertisers to choose the characteristics of users who will see their advertisements and we may use any of the non-personally identifiable attributes we have collected (including information you may have decided not to show to other users, such as your birth year or other sensitive personal information or preferences) to select the appropriate audience for those advertisements."

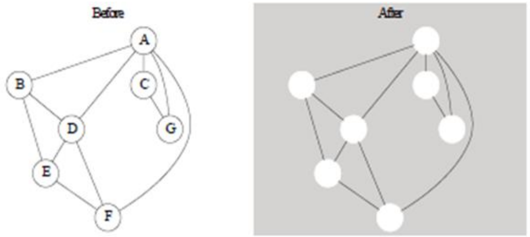

Fig. 1. Naive anonymization removes the ID, but retains the network structure.

Due to the strong correlation to users' social identity, privacy is a major concern in dealing with social network data in contexts such as storage, processing and pub-lishing. Privacy control, through which users can tune the visibility of their profile, is an essential feature in any major social networking service.

A common practice in publishing social network is anonymization, i.e., removing plainly identifying labels such as names, social security numbers, postal or e-mail addresses, but retaining the network structure. Fig-ure 1 illustrates this process.

Can the aforementioned "naive" anonymization tech-nique achieve privacy preservation in the context of privacy-sensitive social network data publishing? This interesting and important question was posed only re-cently [3]. A few privacy attacks have been proposed to circumvent the naive anonymization protection [2, 3]. Meanwhile, more sophisticated anonymization techniques have been proposed to provide better privacy protection [4, 5, 6, 7, 8]. Nevertheless, research in this area is still in its infancy and a lot of work, both in attacks and defenses, remains to be done.

II.System Design

i.Existing System

Digital traces left by users of online social networking services, even after anonymization, are susceptible to privacy breaches. This is exacerbated by the increasing overlap in user-bases among various services. To alert fellow researchers in both the academia and the industry to the feasibility of such an attack.

Disadvantages of Existing System

1. Although a trade-off between utility and privacy is necessary, it is hard, if not impossible, to find a proper balance overall. Besides, it is hard to prevent attackers from proactively collecting intelligence on the social network.
2. It is especially relevant today as major online social networking services provide APIs to facilitate thirdparty application development. These programming interfaces can be abused by a malicious party to gather information about the network.

ii. Proposed System

We propose an algorithm, Seed-and-Grow, to identify users from an anonymized social graph, based solely on graph structure. The algorithm first identifies a seed sub-graph, either planted by an attacker or divulged by a collusion of a small group of users, and then grows the seed larger based on the attacker's existing knowledge of the users' social relations. Our work identifies and relaxes implicit assumptions taken by previous works, eliminates arbitrary parameters, and improves identification effectiveness and accuracy. Simulations on real-world collected datasets verify our claim.

Advantages of Proposed System

1. This algorithm automatically finds a good balance between identification
effectiveness and accuracy.
2. Although a trade-off between utility and privacy is necessary, it is hard, if not impossible, to find a proper balance overall. Besides, it is hard to prevent attackers from proactively collecting intelligence on the social network.

III.Implementation

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Module Description

User Module :

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

Initial Seed Size :

Recent literature on interaction-based social graphs (e.g., the social graph in the motivating scenario) singles out the attacker's interaction budget as the major limitation to attack effectiveness . The limitation translates to 1) the initial seed size and 2) the number of links between the fingerprint graph and the initial seed. Our seed algorithm resolves the latter issue by guaranteeing unambiguous identification of the initial seed, regardless of link numbers. As shown below, our grow algorithm resolves the former issue by working well with a small initial seed.

Grow Algoritham :

At the core of the grow algorithm is a family of related metrics, collectively known as the *dissimilarity* between a pair of vertices from the target and the background graph, respectively. In order to enhance the identification accuracy and to reduce the computation complexity and the false-positive rate, we introduce a *greedy heuristic* with *revisiting* into the algorithm. It is natural to start with those vertices in GT which connect to the initial seed VS because they are more close to the *certain* information, i.e., the already identified vertices VS. For these vertices, their neighboring vertices can be divided into two groups.

Re-Visiting:

The dissimilarity metric and the greedy search algorithm for optimal combination are heuristic in nature. At an early stage with only a few seeds, there might be quite a few mapping candidates for a particular vertex in the background graph; we are very likely to pick a wrong mapping no matter which strategy is used in resolving the ambiguity. If left uncorrected, the incorrect mappings will propagate through the grow process and lead to large-scale mismatch. We address this problem by providing a way to reexamine previous mapping decisions, given new evidences in the grow algorithm; we call this *revisiting*. More concretely, for each iteration, we consider all vertices which have at least one seed neighbor, i.e., those pairs of vertices on which the dissimilarity metrics in are well-defined. We expect that the revisiting technique will increase the accuracy of the algorithm. The greedy heuristic with revisiting is summarized in Algorithm.

SEED-AND-GROW: THE ATTACK

This section describes an attack that identifies users from an anonymized social graph. Let an undirected graph $G_T = \{V_T , E_T \}$ represent the *target* social network after anonymization. We assume that the attacker has an undirected graph $G_B = \{V_B , E_B \}$ which models his *background knowledge* about the social relationships among a group of people, i.e., V_B are labeled with the identities of these people. The motivating scenario demonstrates one way to obtain G_B .
The attack concerned here is to infer the identities of the vertices V_T by considering *structural similarity* between the target graph G_T and the background graph G_B : Nodes that belong to the same users are assumed to have similar connections in G_T and G_B . Although sporadic connections between who would otherwise be strangers may exist in an online social network (and, thus, affect the similarity between G_T and G_B), such links can be removed by, for example, quantifying the strength of these connections [13]; the residual network consists of the stable, strong connections that reflect the users' real-world social relationships, which give rise to the similarity between G_T and G_B . Additionally, auxiliary knowledge about the target graph G_T (such as the source and nature of the graph) may help in choosing a background graph G_B with similar structures.

Thus, the two graphs G_T and G_B are syntactically (the social connections) similar but semantically (the meaning associated with such connections) different. By re-identifying the vertices in G_T with the help of G_B , the attacker associates the sensitive semantics with users on the anonymized G_T and, thus, compromise the privacy of such users. An example of sensitive semantics is the private chat sessions, and their associated timestamps, in the motivating scenario.

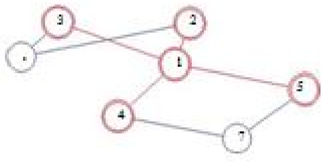

Fig. 2. A randomly generated graph G_F may be symmet-ric.

We assume that, *before* the release of G_T , the attacker obtains (either by creating or stealing) a few accounts and connects them with a few other users (the *initial seeds*) in G_T . The feasibility of doing this is the basis of the Sybil identity forgery attack studied in numerous previous works [14, 15, 16, 17, 18, 19, 20, 21, 22]. In-deed, experiments (Section 4) show that our algorithm is capable of identifying 10 times of anonymized users from as few as 5 initial seeds. Besides user IDs, the attacker knows nothing about the relationship between the initial seeds and other users in G_T . Furthermore, unlike previous works, we *do not assume that the attacker has complete control over the connections*: the attack only *knows* them before G_T 's release. This is more realistic. An example is a confirmation-based social network, in which a connection is established only if the two parties confirm it: the attacker *can decline but not impose* a connection.

The *seed* stage plants (by obtaining accounts and es-tablishing relationships) a small specially designed sub-graph $G_F = \{V_F , E_F \} \subseteq G_T$ (G_F reads as "fingerprint") into G_T before its release. After the anonymized graph is released, the attacker locates G_F in G_T . The neighboring vertices V_S of G_F in G_T are readily identified and serve as the *initial seeds* to be grown.

The *grow* stage is essentially comprised of a structure-based vertex matching, which further identifies vertices adjacent to the initial seeds V_S . This is a self-reinforcing process, in which the seeds grow larger as more vertices are identified.

IV.Conclusion

We propose an algorithm, *Seed-and-Grow*, to identify users from an anonymized social graph. Our algorithm exploits the increasing overlapping user-bases among services and is based solely on social graph structure. The algorithm first identifies a seed sub-graph, either planted by an attacker or divulged by collusion of a small group of users, and then grows the seed larger based on the attacker's existing knowledge of the users' social relations. We identify and relax implicit assump-tions for unambiguous seed identification taken by pre-vious works, eliminate arbitrary parameters in grow algorithm, and demonstrate the superior performance over previous works in terms of identification effective-ness and accuracy by simulations on real-world-collected social-network datasets.

References

[1] B. Krishnamurthy and C. E. Wills, “Characterizing privacy in online social networks,” in *Proc. ACM WOSN*, 2008.

[2] A. Narayanan and V. Shmatikov, “De-anonymizing social net-works,” in *Proc. IEEE S&P*, 2009.

[3] L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and struc-tural steganography,” in *Proc. ACM WWW*, 2007.

[4] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, “Anonymizing social networks,” Univ. Massachusetts, Amherst, Tech. Rep., 2007.

[5] E. Zheleva and L. Getoor, “Preserving the privacy of sensitive relationships in graph data,” in *Proc. ACM SIGKDD*, 2007.

[6] A. Korolova, R. Motwani, S. Nabar, and Y. Xu, “Link privacy in social networks,” in *Proc. ACM CIKM*, 2008.

[7] B. Zhou and J. Pei, “Preserving privacy in social networks against neighborhood attacks,” in *Proc. Intl. Conf. on Data Engineering (ICDE)*. IEEE, 2008.

[8] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, “Resisting structural re-identification in anonymized social networks,” *VLDB Endowment*, vol. 1, no. 1, pp. 102–114, 2008.

[9] J. Scott, *Social network analysis: a handbook*. SAGE Publications, 2000.

[10] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Incognito: efficient full-domain k-anonymity,” in *Proc. ACM ICMD*, 2005.

[11] B. Zhou, J. Pei, and W. Luk, “A brief survey on anonymization techniques for privacy preserving publishing of social network data,” *ACM SIGKDD Explorations Newsletter*, vol. 10, no. 2, pp. 12–22, 2008.

[12] A. Mislove, H. Koppula, K. Gummadi, P. Druschel, and B. Bhat-tacharjee, “Growth of the flickr social network,” in *Proc. WOSN*. ACM, 2008.

[13] R. Xiang, J. Neville, and M. Rogati, “Modeling relationship strength in online social networks,” in *Proc. ACM WWW*, 2010.

[14] J. Douceur, “The sybil attack,” *LNCS*, vol. 2429, pp. 251–260, 2002.

[15] N. Tran, B. Min, J. Li, and L. Subramanian, “Sybil-resilient online content voting,” in *Proc. USENIX NSDI*, 2009.

[16] S. Park, B. Aslam, D. Turgut, and C. Zou, “Defense against sybil attack in vehicular ad hoc network based on roadside unit support,” in *Proc. IEEE MILCOM*, 2009.

[17] C. Lesniewski-Laas and M. Kaashoek, “Whanau: A sybil-proof distributed hash table,” in *Proc. USENIX NSDI*, 2010.

[18] C. Chen, X. Wang, W. Han, and B. Zang, “A robust detection of the sybil attack in urban vanets,” in *Proc. IEEE ICDCS*, 2009.

[19] H. Yu, M. Kaminsky, P. Gibbons, and A. Flaxman, “Sybilguard: defending against sybil attacks via social networks,” *ACM SIG-COMM CCR*, vol. 36, no. 4, pp. 267–278, 2006.

[20] H. Yu, P. Gibbons, M. Kaminsky, and F. Xiao, “Sybillimit: A near-optimal social network defense against sybil attacks,” in *Proc. IEEE S&P*, 2008.

[21] B. Viswanath, A. Post, K. Gummadi, and A. Mislove, “An analysis of social network-based sybil defenses,” *ACM SIGCOMM CCR*, vol. 40, no. 4, pp. 363–374, 2010.

[22] W. Wei, F. Xu, C. Tan, and Q. Li, “SybilDefender: Defend against sybil attacks in large social networks,” in *Proc. IEEE INFOCOM*, 2012.

[23] S. Sorlin and C. Solnon, “Reactive tabu search for measuring graph similarity,” *LNCS*, vol. 3434, pp. 172–182, 2005.

[24] P. Erdos¨ and A. Renyi,´ “On random graphs,” *Publicationes Math-ematicae*, vol. 6, no. 26, pp. 290–297, 1959.

[25] R. Kumar, J. Novak, and A. Tomkins, “Structure and evolution of online social networks,” in *Proc. ACM SIGKDD*, 2006.

[26] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhat-tacharjee, “Measurement and analysis of online social networks,” in *Proc. ACM IMC*, 2007.

[27] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, “Statistical properties of community structure in large social and information networks,” in *Proc. ACM WWW*, 2008.

[28] M. Porter, J. Onnela, and P. Mucha, “Communities in networks,” *Notices of the AMS*, vol. 56, no. 9, pp. 1082–1097, 2009.

[29] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, “Micro-scopic evolution of social networks,” in *Proc. ACM SIGKDD*,

2008.

[30] A. Barabasi´ and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[31] D. Soares, C. Tsallis, A. Mariz, and L. Silva, “Preferential at-tachment growth model and nonextensive statistical mechanics,” *Europhysics Letters*, vol. 70, p. 70, 2005.

[32] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in *Proc. IEEE S&P*, 2008.

[33] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao, “User interactions in social networks and their implications,” in *Proc. ACM EuroSys*, 2009.