ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

A Survey on Attack Prevention and Handling Strategies in Graph Based Data Anonymization

S.Charanyaa[1], T.Shanmugapriya[2]

M.Tech., Student, Dept of Information Technology, S.N.S. College of Technology, Coimbatore, TamilNadu, India[1]

Assistant Professor, Dept of Information Technology, S.N.S. College of Technology, Coimbatore, TamilNadu, India[2]

Abstract: Current days have seen a steep rise in the need for the social network data to be published in public. With the rapid growth of web application development the need for private data to be published has grown multi fold. Representing the individual sensitive labels in a graphical structure for the ease of access and quick re-identification is an important issue to be addressed. In this paper, we have made a detailed surveyed about the existing techniques that preserve the sensitive data in social network data. It is pragmatic that preserving the graph structure and label re-identification by adding some noise nodes to the graph makes difference in degree is inferred from literature. Several techniques dealt in literature to deal with abnormal graph structure are studied in detail. The various possible attacks on social network data and techniques to prevent and handle the same are studied in detail.

Keywords: Privacy, Graphical Data, Re-identification, Attack, Data Security, Anonymization

I. INTRODUCTION

Need for publishing sensitive data to public has grown extravagantly during recent years. Though publishing demands its need there is a restriction that published social network data should not disclose private information of individuals. Hence protecting privacy of individuals and ensuring utility of social networ data as well becomes a challenging and interesting research topic. Considering a graphical model [35] where the vertex indicates a sensitive label algorithms could be developed to publish the non-tabular data without compromising privacy of individuals. Though the data is represented in graphical model after KDLD sequence generation [35] the data is susceptible to several attacks such as homogeneity attack, background knowledge attack, similarity attacks and many more. In this paper we have made an investigation on the attacks and possible solutions proposed in literature and efficiency of the same. The remainder of the paper is organized as follows. Techniques for detecting abnormal graph structures are dealt in detail in Section 2. Section 3 gives the survey on attacks on social network data. Section 4 briefs about techniques to prevent and handle the attacks on social network data. Section 5 concludes the paper and outlines the future work.

II. TECHNIQUES FOR DETECTING ABNORMAL STRUCTURES IN GRAPHS

W. Eberle and L. Holder(2007)[9] dealt in detail about discovering structural anomalies in graph-based data. Authors presented graph-based approaches to expose anomalies in domains where the anomalies consist of entity/relationship alterations that closely resemble non-anomalous activities. Further they introduced three new algorithms for detecting anomalies in all three types of possible graph changes namely label modifications, vertex/edge insertions and vertex/edge deletions. Each of the three proposed algorithms focuses on one of these anomalous types and uses the minimum description length principle to discover those substructure instances that contain anomalous entities and relationships. Authors evaluated the effectiveness of each of these algorithms in terms of each of the types of anomalies using synthetic and real-world data for the purposes of detecting fraud.

C.C. Noble and D.J. Cook (2003) [22] dealt about graph-based anomaly detection. Authors introduced two techniques for graph-based anomaly detection. In addition, a new method for calculating the regularity of a graph, with applications to anomaly detection was also proposed. Authors hypothesized that these methods established usefulness for finding anomalies, and for determining the likelihood of successful anomaly detection within graph-based data. The work was supported with experimental results generated using both real-world network intrusion data and artificially-generated data.

III. ATTACKS ON SOCIAL NETWORK DATA

There's another type of attack on social networks, which is called "active attack." "Active attack" (2013) [35] is to actively implant special subgraphs into a social network

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

when this social network is collecting data. An attacker can attack the users who are connected with the embedded subgraphs by reidentifying these special subgraphs in the published graph. Backstrom et al.(2007) [1] described active attacks based on randomness analysis and showed that an attacker may fix some constructed substructures linked with the target entities. One technique to stop the active attack is to distinguish the fake nodes appended by attackers and eliminate them before releasing the data. Authors also described a set of attacks such that even from a single anonymized version of a social network, it is possible for an adversary to gain knowledge whether edges exist or not between explicit targeted pairs of nodes.

Shrivastava et al. [25] and Ying et al.[28] throwed light on a special active attack named Random Link Attack. Shrivastava et al. [25] projected an algorithm that is capable of identifying fake nodes based on the triangle probability difference between normal nodes and fake nodes. Ying et al. [28] proposed an alternate method, which employs spectrum analysis to identify the fake nodes. To publish a graph that is potentially changed by Random Link Attack, the publisher can employ a twostep mechanism proposed by the authors. First, the graph is filtered by the methods introduced by Backstrom et al. [1] or Shrivastava et al. [25]. Second, the attacker can produce the published graph using the proposed model from the filtered graph.

X. Ying, X. Wu, and D. Barbara (2011) [28] worked on spectrum based fraud detection in social networks. Since social networks are susceptible to a variety of attacks such as spam emails, viral marketing etc., the authors developed a spectrum based detection framework to discover the perpetrators of these attacks. In specific, the authors focused on Random Link Attacks (RLAs) in which the malicious user introduces multiple false identities and interactions among those identities to later proceed to attack the regular members of the network. Through experimental results authors proved that RLA attackers can be filtered by using their spectral coordinate characteristics, which are hard to hide even after the efforts by the attackers of mimicking as much as possible the entire network, and the proposed method is found to be promising in detecting those attackers and outperforms various other techniques in the literature.

There are also two other works in literature that focus on the edge weight protection in weighted graphs. Liu et al. [19] treated weights on the edges as sensitive labels and proposed a method to preserve shortest paths between most pairs of nodes in the graph. L. Liu, J. Wang, J. Liu, and J. Zhang (2008) [19] considered preserving weights data privacy of certain edges, while trying to preserve close shortest path lengths and exactly the same shortest paths of certain pairs of nodes. Also the authors developed two privacy preserving strategies for this application. The first strategy is based on a Gaussian randomization multiplication, and the second one is a greedy perturbation algorithm which is based on the graph theory.

Das et al. [8] proposed a Linear Programming-based method to protect the edge weights while preserving the path of shortest paths. These two works focused on the protection of edge weights instead of nodes. Some authors studied other attacks besides the "passive attack" and "active attack." Zheleva [31] analyzed the ability of an attacker to learn the unpublished attributes of users in an online social network when the user uses the published attributes and relationships between users to do the data mining. E. Zheleva and L. Getoor (2009) [31] studied the illusion of privacy in social networks with mixed public and private user profiles. Authors depicted how an adversary can exploit an online social network with a mixture of public and private user profiles to predict the private attributes of users. This problem is mapped to a relational classification problem and authors proposed practical models that use friendship and group membership information to infer sensitive attributes. Authors' key novel idea is that in addition to friendship links, groups can be carriers of significant information. They proved that on several well-known social media sites, the information regarding private-profile users can easily and accurately be recovered. the information of private-profile users.

Narayanan and Shmatikov [21] (2009) proved the fact that from "seed nodes," a large quantity of othervnodes can be reidentified. Authors aimed to develop effective algorithms for de-anonymizing real-world social networks. In specific, authors focused on dual tasks: one is to align the networks of Flickr and Instagram and the other is to align Flickr and Twitter. Their work was motivated by the two parts of information that network data is composed of: network structure and node attributes. Preliminary tests conducted by the authors proved that de-anonymizing algorithm based merely on node attributes, is computationally effective but not adequately accurate. As against, algorithms that rely on network structures, which bring in more relationship information, might contribute to the precision of de-anonymization. Authors concluded that the structure of the real-world social networks be quite different, but also the computation costs will be significantly high because it is difficult to align two networks merely based on their structures without any assisting information. With these facts, authors decided to develop approaches that can combine network structure information and node attributes to do the alignment.

L. Page, S. Brin, R. Motwani, and T. Winograd (1999) [23] describes PageRank method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them. Authors compared PageRank to an idealized random Web surfer and illustrated techniques to efficiently compute PageRank for large numbers of pages and to apply PageRank to search and

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

to user routing. S.R.Ganta, S. Kasiviswanathan, and A. Smith (2008) [11] explored how an individual can reason about privacy in the presence of concrete sources of auxiliary information. Specifically, the authors investigated the effectiveness of current anonymization schemes in preserving privacy when multiple organizations independently release anonymized data about overlapping populations. Further investigations were also carried out by the author in composition attacks, in which an adversary uses independent anonymized releases to violate privacy. Authors clarify the reason behind the failure of recently proposed models of limited auxiliary information to capture composition attacks. For the proposed technique, authors demonstrated that even a simple instance of a composition attack can breach privacy. Optimistically, some randomization-based notions of privacy, usually resist composition attacks and, in fact, the use of arbitrary side information. This resistance allows "stand-alone" design of anonymization schemes, eliminating the need for explicitly keeping track of other releases. Authors also provide an accurate formulation of this property, and proved that an important class of relaxations of differential privacy also satisfy the property.

Zou et al. (2009) [34] proposed k-automorphism to protect against multiple structural attacks and develop an algorithm called KM that ensures k-automorphism. Authors also discussed an extension of KM to handle "dynamic" releases of the data. A graph is k-Automorphism if and only if for every node there exist at least k-1 other nodes do not have any structure difference with it. Extensive experiments prove that k-automorphism performs well in terms of protection. J.Cheng, A.W.c. Fu, and J. Liu (2010) [6] identified a new problem of enforcing k-security for protecting sensitive information concerning the nodes and links in a published network dataset. Their investigation lead to the invention of k-isomorphism where, the selection of anonymization algorithm depends on the adversary knowledge and the targets of protection. Authors addressed the information of protetion against structural attack if the target is only NodeInfo. Authors say that NodeInfo and LinkInfo are two basic sources of sensitive information in network datasets, and they call for special efforts for their security. X. Xiao and Y. Tao (2006) [27] developed a linear-time algorithm for computing anatomized tables that obey the l-diversity privacy requirement, and diminish the fault of reconstructing the microdata. Extensive experimental results of the authors confirm the fact that the proposed technique allows significantly more effective data analysis than the conventional publication method based on generalization. Especially, anatomy allows aggregate reasoning with average error below 10%, which is lower than the error obtained from a generalized table interms magnitude.

X. Ying and X. Wu (2008) [29] proposed a spectrum preserving approach of randomizing social networks. Authors investigated the consequence of various properties of networks due to randomization. They studied how random deleting and swapping edges change graph properties and proposed an eigenvalues oriented random graph change algorithm. All the edge editing- based models prefers to produce a published graph with as fewer edge change. A. Campan and T.M. Truta (2008) [4] proposed a clustering approach for data and structural anonymity in social networks. Authors discussed how to implement clustering when considering the lost of both node labels and structure information S.Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava (2009) [3] enunciated a Class based graph anonymization for social network data. Since social network data is more affluent in details about the users and their communications, loss of details due to anonymization confines the possibility for analysis. Authors present a new set of techniques for anonymizing social network data based on grouping the entities into classes, and masking the mapping between entities and the nodes that represent them in the anonymized graph. Their techniques allow queries over the rich data to be evaluated with high degree of accuracy while guaranteeing flexibility to certain types of attack. To prevent inference of interactions, there is a support on a critical "safety condition" when forming these classes. Authors demonstrate utility via empirical data from social networking settings and provide illustrations of complex queries that may be posed and proved that they can be answered over the anonymized data more efficiently and effectively.

IV. TECHNIQUES TO PREVENT AND HANDLE THE ATTACKS ON SOCIAL NETWORK DATA

K. Liu and E. Terzi (2008) [18] formally defined the graph-anonymization problem and devise simple and efficient algorithms for solving the problem. The algorithms are based on principles related to the realizability of degree sequences. The algorithms were applied to a huge range of synthetic and real datasets and their efficiency and practical utility is demonstrated. B.Zhou and J. Pei (2008) [32] identified a type of privacy attacks called neighborhood attacks. If an antagonist has certain knowledge about the neighbors of a target prey and the relationship among the neighbors, the prey may be re-identified from a social network even if the prey's identity is preserved by employing conventional anonymization techniques. They also proved that the problem is challenging and NP-hard and gave a practical solution to fight against neighborhood attacks. The empirical study indicated that their method of generation of anonymized social networks can be employed to answer aggregate network queries with peak accuracy.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis (2008) [15] categorized the entities connected by relations such as friendship, communication, or shared activity. They quantified the privacy risks associated with three different classes of attacks on the privacy of individuals in networks, based on the adversarial knowledge. They proved that network structure and size are the main root of the risks of these attacks. They also proposed a novel approach to anonymizing network data that models aggregate network structure and then allows samples to be drawn from that model, which guarantees anonymity for network entities while preserving the ability to estimate a wide variety of network measures with negligible bias.

L. Zou, L. Chen, and M.T. Ozsu (2009) [34] proposed k-automorphism to protect against multiple structural attacks and developed an algorithm called KM that ensures k-automorphism. Authors also discussed an extension of KM to handle "dynamic" releases of the data and proved that the algorithm performs well in terms of protection it provides. A significant work of defining a k-degree anonymity model to prevent degree attacks that employ node degree is done by Liu and Terzi (2008) [18]. A graph is termed to be k-degree anonymous if and only if for every node in the graph, there exist at least k-1 other nodes with the same degree

V. CONCLUSION AND FUTURE WORK

In this paper, we have made an study on several approaches of anonymization and knowledge hiding of graphical data. Several algorithms pertaining to ensuring database privacy have been studied along with the computation overhead involved in implementing the algorithms for real and synthetic data sets. Techniques for detecting abnormal graph structures such as label modifications, vertex/edge insertions and vertex/edge deletions are dealt in detail. Various types of attacks possible on social network such as active attack based on randomness, random link attack, passive attack, seed node de-identification, multiple structural attack are surveyed in detail. Various techniques to prevent and handle the attacks on social network data such as prey identification, friendship communication, shared activity and dynamic release of data are analyzed . In future we planned to develop novel efficient algorithms to defend against homogeneity attack, background knowledge attack and similarity attack.

REFERENCES

[1] L. Backstrom, C. Dwork, and J.M. Kleinberg, "Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography," Proc. Int'l Conf. World Wide Web (WWW), pp. 181-190, 2007.
[2] A.-L. Baraba´si and R. Albert, "Emergence of Scaling in Random Networks," Science, vol. 286, pp. 509-512, 1999.
[3] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava, "Class-Based Graph Anonymization for Social Network Data," Proc. VLDB Endowment, vol. 2, pp. 766-777, 2009.
[4] A. Campan and T.M. Truta, "A Clustering Approach for Data and Structural Anonymity in Social Networks," Proc. Second ACM SIGKDD Int'l Workshop Privacy, Security, and Trust in KDD (PinKDD '08), 2008.
[5] A. Campan, T.M. Truta, and N. Cooper, "P-Sensitive K-Anonymity with Generalization Constraints," Trans. Data Privacy, vol. 2, pp. 65-89, 2010.
[6] J. Cheng, A.W.-c. Fu, and J. Liu, "K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks," Proc. Int'l Conf. Management of Data, pp. 459-470, 2010.
[7] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang, "Anonymizing Bipartite Graph Data Using Safe Groupings," Proc. VLDB Endowment, vol. 1, pp. 833-844, 2008.
[8] S. Das, O. Egecioglu, and A.E. Abbadi, "Privacy Preserving in Weighted Social Network," Proc. Int'l Conf. Data Eng. (ICDE '10), pp. 904-907, 2010.
[9] W. Eberle and L. Holder, "Discovering Structural Anomalies in Graph-Based Data," Proc. IEEE Seventh Int'l Conf. Data Mining Workshops (ICDM '07), pp. 393-398, 2007.
[10] K.B. Frikken and P. Golle, "Private Social Network Analysis: How to Assemble Pieces of a Graph Privately," Proc. Fifth ACM Workshop Privacy in Electronic Soc. (WPES '06), pp. 89-98, 2006.
[11] S.R. Ganta, S. Kasiviswanathan, and A. Smith, "Composition Attacks and Auxiliary Information in Data Privacy," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 265- 273, 2008.
[12] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast Data Anonymization with Low Information Loss," Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07), pp. 758-769, 2007.
[13] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A Framework for Efficient Data Anonymization Under Privacy and Accuracy Constraints," ACM Trans. Database Systems, vol. 34, pp. 9:1-9:47, July 2009.
[14] J. Han, Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc., 2005.
[15] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting Structural Re-Identification in Anonymized Social Networks," Proc. VLDB Endowment, vol. 1, pp. 102-114, 2008.
[16] E.M. Knorr, R.T. Ng, and V. Tucakov, "Distance-Based Outliers: Algorithms and Applications," The VLDB J., vol. 8, pp. 237-253, Feb. 2000.
[17] N. Li and T. Li, "T-Closeness: Privacy Beyond K-Anonymity and L-Diversity," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE '07), pp. 106-115, 2007.
[18] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," SIGMOD '08: Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 93-106, 2008.
[19] L. Liu, J. Wang, J. Liu, and J. Zhang, "Privacy Preserving in Social Networks against Sensitive Edge Disclosure," Technical Report CMIDA-HiPSCCS 006-08, 2008.
[20] A. Machanavajjhala, D. Kifer, J. Gehrke, and M.Venkitasubramaniam, "L-Diversity: Privacy Beyond K-Anonymity," ACM Trans. Knowledge Discovery Data, vol. 1, article 3, Mar. 2007.
[21] A. Narayanan and V. Shmatikov, "De-Anonymizing Social Networks," Proc. IEEE 30th Symp. Security and Privacy, pp. 173-187,2009
[22] C.C. Noble and D.J. Cook, "Graph-Based Anomaly Detection," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '03), pp. 631-636, 2003.
[23] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing Order to the Web," Proc. World Wide Web Conf. Series, 1998.
[24] K.P. Puttaswamy, A. Sala, and B.Y. Zhao, "Starclique: Guaranteeing User Privacy in Social Networks Against Intersection Attacks," Proc. Fifth Int'l Conf. Emerging Networking Experiments and Technologies (CoNEXT '09), pp. 157-168, 2009.
[25] N. Shrivastava, A. Majumder, and R. Rastogi, "Mining (Social) Network Graphs to Detect Random Link Attacks," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE '08), pp. 486-495, 2008.
[26] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertain. Fuzziness Knowledge-Based Systems, vol. 10, pp. 557- 570, 2002.
[27] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. 32nd Int'l Conf. Very Large Databases (VLDB '06), pp. 139-150, 2006.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

[28] X. Ying, X. Wu, and D. Barbara, "Spectrum Based Fraud Detection in Social Networks," Proc. IEEE 27th Int'l Conf. Very Large Databases (VLDB '11), 2011.
[29] X. Ying and X. Wu, "Randomizing Social Networks: A Spectrum Preserving Approach," Proc. Eighth SIAM Conf. Data Mining (SDM '08), 2008.
[30] E. Zheleva and L. Getoor, "Preserving the Privacy of Sensitive Relationships in Graph Data," Proc. First SIGKDD Int'l Workshop Privacy, Security, and Trust in KDD (PinKDD '07), pp. 153-171, 2007.
[31] 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," Proc. 18th Int'l Conf. World Wide Web (WWW '09), pp. 531-540, 2009.
[32] B. Zhou and J. Pei, "Preserving Privacy in Social Networks Against Neighborhood Attacks," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE '08), pp. 506-515, 2008.
[33] B. Zhou and J. Pei, "The K-Anonymity and L-Diversity Approaches for Privacy Preservation in Social Networks against Neighborhood Attacks," Knowledge and Information Systems, vol. 28, pp. 47-77, 2011.
[34] L. Zou, L. Chen, and M.T. O¨zsu, "K-Automorphism: A General Framework for Privacy Preserving Network Publication," Proc. VLDB Endowment, vol. 2, pp. 946-957, 2009.
[35] Mingxuan Yuan, Lei Chen, Philip S. Yu, Ting Yu, "Protecting Sensitive Labels in Social Network Data Anonymization", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 3, March 2013
[36] S.Balamurugan, P.Visalakshi, "Modified Partitioning Algorithm for Privacy Preservation in Microdata Publishing with Full Functional Dependencies", Australian Journal of Basic and Applied Sciences, 7(8): 316-323, July 2013

BIOGRAPHIES

S.Charanyaa obtained her B.Tech degree in Information Technology from Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India. She is currently pursuing her M.Tech degree in Information Technology at S.N.S. College of Technology, Coimbatore, Tamilnadu, India. Her areas of research interest accumulate in the areas of Database Security, Privacy Preserving Database, Object Modeling Techniques, and Software Engineering.

Prof.T.Shanmugapriya is currently working as Assistant Professor in the Department of Information Technology at S.N.S. College of Technology, Coimbatore, Tamilnadu, India. She has 6 years and 2 months of teaching experience. She has published a number of research papers which include 4 International Journals, 5 National Conferences and 3 International Conferences. Her areas of research interest accumulate in the area of Computer Networks.