Multimed Tools Appl (2019) 78:29731–29745
https://doi.org/10.1007/s11042-018-6217-9

A traffic analysis attack to compute social network measures

Alejandra Guadalupe Silva Trujillo[1] · Ana Lucila Sandoval Orozco[2] · Luis Javier García Villalba[2] · Tai-Hoon Kim[3]

Received: 8 February 2018 / Revised: 28 April 2018 / Accepted: 25 May 2018
Published online: 14 June 2018

Abstract The development of digital media, the increasing use of social networks, the easier access to modern technological devices, is perturbing thousands of people in their public and private lives. People love posting their personal news without consider the risks involved. Privacy has never been more important. Privacy enhancing technologies research have attracted considerable international attention after the recent news against users personal data protection in social media websites like Facebook. It has been demonstrated that even when using an anonymous communication system, it is possible to reveal user's identities through intersection attacks or traffic analysis attacks. Combining a traffic analysis attack with Analysis Social Networks (SNA) techniques, an adversary can be able to obtain important data from the whole network, topological network structure, subset of social data,

✉ Luis Javier García Villalba
javiergv@fdi.ucm.es

Alejandra Guadalupe Silva Trujillo
asilva@uaslp.mx

Ana Lucila Sandoval Orozco
asandoval@fdi.ucm.es

Tai-Hoon Kim
taihoonn@daum.net

[1] Facultad de Ingeniería, Universidad Autónoma de San Luis Potosí (UASLP), Zona Universitaria Poniente, San Luis Potosí 78290, México

[2] Group of Analysis, Security and Systems (GASS), Faculty of Computer Science and Engineering, Department of Software Engineering and Artificial Intelligence (DISIA), Universidad Complutense de Madrid (UCM), Office 431, Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, Spain

[3] Department of Convergence Security, Sungshin Women's University, 249-1 Dongseon-Dong 3-ga, Seoul 136-742, Korea

revealing communities and its interactions. The aim of this work is to demonstrate how intersection attacks can disclose structural properties and significant details from an anonymous social network composed of a university community.

Keywords Social networks analysis · Privacy · Statistical disclosure attack · Anonymous network communications

1 Introduction

Society is every day more attached to technology. Threatens, virus, technological risks always have existed but with our technological dependence, security has becomes an increasing serious matter. In the last few years, online social networks have changed the dynamics of our day to day lives. Several studies show how people are most interested in maintain more friends and being well-liked, than keep his personal information restricted [7]. Millions of people no matter age, nationality, or education are part of online social networks and expose highly personal information about them in exchange for a service to communicate with their friends, family and colleagues. Companies like Facebook, Instagram, Twitter, or Tinder have sought to coordinate an attitudinal shift of privacy value.

Privacy in social networks is an open field of research; data is the source of decision makers and analysts. It has been shown how much personal information people disclose voluntary, like full names, photos, mobile numbers, address, etc. Many of them are unaware of risks and price associated to their personal information [39]. When using online social networks, it is quite important to understand and recognize the privacy risks involved [20]. The majority of people are unaware of the fact that their privacy has been endanger and they don't do anything to protect themselves. For example, if someone posts personal information online, it is no longer private, and this could fall into the wrong hands. Even when it was posted with the highest possible security measures, some of the users' acquaintances such as friends, colleagues and companies interacting with them, can expose their personal information. Even when there are no intentional purposes, it can be used to deliberate diminish their security or privacy. These users can become victims of identity theft, harassment, cyber bullying or illegal practices. The privacy paradox argues that individuals express security and privacy concerns about information sharing in online social network, but their behavior states the opposite to publish all kind of personal data [7]. Experts recommend evaluate our online social networks profile to pay close attention about the way each profile permit to protect personal information. Furthermore, take advantage of the enhanced privacy tools available to block personal targeting. In spite of privacy as a human right and necessary condition for the goods that are part of our well-being like freedom and security, it is important to pursue the goal of make better platforms for communicate but without exchange our privacy value. We have observed in recent years the implementation of technologies oriented to different approaches like anonymous communications, identity management, digital credentials, e-voting, privacy engineering, and others [16, 21, 22]. In this sense, privacy enhancing technologies aim to build mechanisms to provide tools to safely interact with technology keeping privacy's users. This topic is also developed in k-anonymity [41], l-diversity [31] and t-closeness [34] methods. These are privacy preserving techniques which aim to protect databases scrambling and swapping values or adding noise to keep information usable. The challenge is to release data and maintain it interesting for analysis and statistical research [25]. In [23] exhibit the effect of combining k-anonymity with unlinkable systems like mix servers.

In this work we are interested in intersection attacks or traffic analysis attacks research. The family of statistical disclosure attacks belongs to it [13]. Combining a statistical disclosure attack with Analysis Social Networks (SNA) techniques, one adversary can be able to obtain important data from the whole network, topological network structure, subset of social data, revealing communities and its interactions [1]. For practical purposes the study of SNA is a field well known in different areas. Community detection recognizes groups of interest based on their behavior [29]. SNA may help to know who is a leader person in order to influence other users to shop specific products. Link prediction research infers new interactions in a social network based on analyzing several measures of their nodes [3, 20, 38]. Sociologists and history researchers want to know the correlation about political and social actors [2, 24]; epidemiologists study disease transmission and the influence of personal and social networks on health behavior; anthropologists measure the evolution of sociocultural systems, trying to understand what is going on, what went on before, and what the future prospects are [8, 20, 42, 45]. Collective inference techniques are used for online blog analysis in order to predict entity behavior through its connections. Using automatic learning techniques or natural language models it is desired to identify a text author by carrying out an analysis of his writing and the vocabulary used [4, 32]. In the ethological field SNA is used to study behavior in animals to learn about members of the same species [47]. These areas of study include SNA, but are not limited to economics, biology, anthropology, information science, social psychology, sociolinguistics, sociology, and so on.

However, there are still open problems associated when background knowledge is available to an attacker. Some of the results of identity disclosure in social network anonymous communications have been published in previous papers [36, 37]. The aim of this work is to extend and clarifies our intersection attack to divulge the structural properties of a social network. The foundation of this work has been presented in [37]. In this paper we have validated our algorithms distinguishing how many relationships can be inferred in an anonymous network when attacker gets partial information. We can obtain competitive advantages by using them and disclose important information in a real social network even when a mechanism of anonymization has been applied. In Section 2 we show the relevance of analysis social network and a state of art in statistical disclosure attack applied to real social networks. Also, it describes briefly several techniques used to mining data in social networks. The results and simulations are shown in Section 3, and the conclusions and future work are completed in Section 4.

2 Relevance of privacy in social network

2.1 Social network properties

A social network is a social structure made of individuals, which are connected by one or more types of relationships. Its representation can be made through a graph where the vertices represent individuals or entities and the edges the relations among them. Formally a simple social network is modeled as a graph $G = (V, E)$ where: $V = (v_1, v_n)$ is the set of vertices or nodes, represented as entities or individuals. E is the set of social relationships, represented as edges in the graph, where $E = (v_i, v_j)|v_i, v_j \epsilon V$.

In literature exist three levels of analysis within the Social Network Analysis (SNA) [26, 40, 46]: i) analysis of egocentric networks; ii) analysis focused on subgroups of actors; iii) analysis focused on the overall structure of the network. The objective of the analysis of egocentric networks is to study how a behavior actor evolves, taking into account that is focus

Table 1 Example of a representation of friendship

Users	U1	U2	U4	U5
U1	0	1	0	0
U2	0	0	1	0
U3	0	0	1	0
U4	0	0	0	0
U5	1	1	0	0

solely on that actor and his relationships with the rest of the participants. The second type of analysis allows understanding the logical of networks clustering and the existence of cooperation and competition patterns, which are adapted or maintained over time. Finally, in the analysis of overall structure of the network are considered the morphological characteristics adopted, the existence, role and subgroups interaction, the distribution of relationships between actors involved, the geodesic distance between actors, among others. According to the type of problem to solve some of the three levels of analysis is chosen.

The structural analysis of a social network is based on develop a matrix and a graph to represent the relationships among users. It is common to use and adjacency matrix M for a graph representation of n^2 size, where n is the number of nodes. If there is an edge between node i and node j, 1 is placed in the cell (i, j) and 0 otherwise. Let's imagine we want to examine friendship in a set of 5 people. Its representation is show in Table 1 with an adjacency matrix where 1 indicates the existence of friendship and 0 no relationship between user i and j. Figure 1 shows the same friend relationships through a directed graph composed of 5 nodes.

The graph can also be classified according to various topological measures. In SNA is important to know if it is possible to reach a node through another node. In this case, it is interesting to identify how many ways exist and which one is the best. Paths are used to calculate distance between two nodes. Path is a set of nodes and different lines. The path length is the number of lines in it, where the first node is called the origin and the final destination. A shortest path between two nodes is the minimal length path of all the possible paths between nodes.

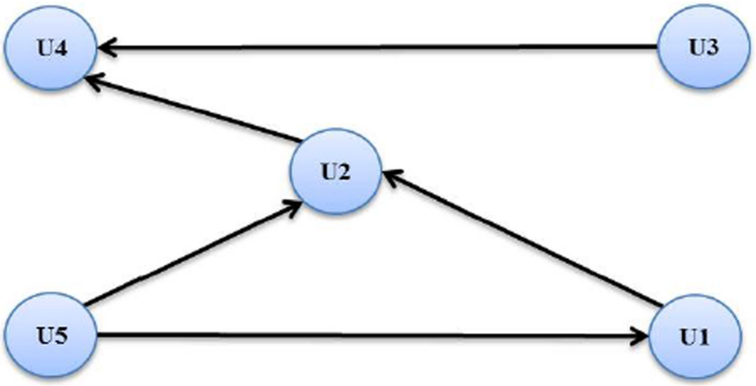

Fig. 1 Example of a directed graph

One of the most common paths is called geodesic path, which is the shortest path between two nodes. The length of a geodesic path is called geodesic distance and is denoted as $d(i, j)$, which is the distance between the nodes n_i and n_j. Both directed and undirected graphs, the geodesic distance is the number of relationships in the shortest possible path from one actor to another. Distances are mainly used in some of the centrality measures. One of the main uses of graph theory in SNA is to recognize the most important nodes. Centrality measures at node level are node degree, nodal transitivity degree, betweenness and closeness. Measures related to the entire network such as density, diameter and clustering coefficient allow comparison of the whole network structure.

A network can be an extremely complex structure; the connections between nodes may have complicated patterns. One challenge at studying complex networks is to develop simple metrics that capture structural elements in an understandable form. One such simplification is to ignore any pattern between different nodes, and observe each node separately. Node degree in an undirected network is the number of its connections. By counting the number of nodes that each degree, it can be established the grade distribution $P_{deg}(k)$ defined as the percentage of nodes in the graph with degree k.

An example of the distribution of degrees of an undirected graph shown in Fig. 2.

Where degrees are $k_1 = 1, k_2 = 3, k_3 = 1, k_4 = 1, k_5 = 2, k_6 = 5, k_7 = 3, k_8 = 3, k_9 = 2$, and $k_10 = 1$. The grade distribution is $P_{deg}(1) = \frac{4}{10}, P_{deg}(2) = \frac{2}{10}$, and $P_{deg}(3) = \frac{3}{10}, P_{deg}(5) = \frac{1}{10}$

Distribution degrees gives important clues within the structure of a network. For example, in the simplest types of networks, it is common to find most nodes in the network have similar degrees. Real-world networks usually have very different degrees distribution. In such networks, most nodes have a relatively small degree, but there are few nodes with a very high degree.

There are several works in the literature that suggest real-world social networks have very particular characteristics. Complex networks as www or social networks do not have an organized architecture, but rather have been promoted organized themselves according to the actions of many individuals. From these interactions global phenomenon, can emerge for example, properties of small world or free scale distribution. These two global properties have considerable implications for network behavior under attack, as well as dissemination of information or epidemiological issues. In late 1950, Erdos and Renyi [19] marked a precedent in classical mathematical theory to model problems of complex networks describing a network using a random graph, defining the foundations of the theory of random networks.

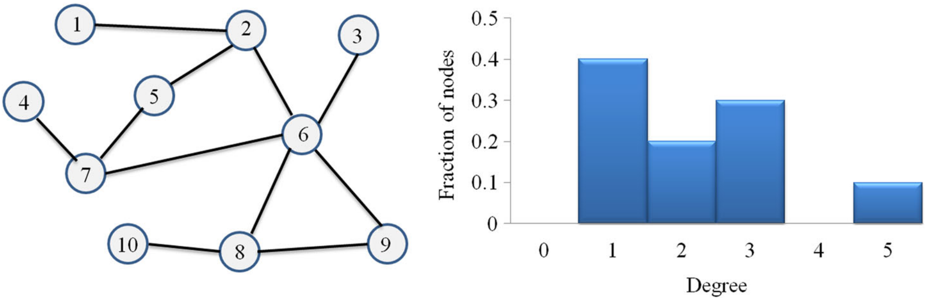

Fig. 2 Distribution degrees example

Networks composed of people connected through the exchange of emails exhibit characteristics of small world networks and scale-free networks. The “scale-free network” definition describes the kind of networks that exhibit a power-law distribution [6]. The characteristic of such networks is distribution of links results in a straight line if plotted on a logarithmic scale twice, as we can see in Fig. 3. The power law is a member of the family of distributions skewed toward the extremes, so describing events in which a random variable reaches high values infrequently, while medium or low values are much more common. Seen from another angle, the power law probability of occurrence of small events is relatively high, while the probability of occurrence of large events is relatively low.

In literature have analyzed the structural properties of email networks, the results have concluded that traffic from a legitimate email system results in small-world networks and scale-free [33]. On the other hand, it is also argued that considering an email system as a single whole, does not display a scale-free behavior completely antisocial behavior as spam (Fig. 3).

One of the previous works about email networks consider the study of the structure of emails networks observing university log files [18]. Taking into account the network topologies of email address where emails are nodes and edges are the communications among them. The resulting network also shows a distribution of links or relationships with pronounced free scale and small-world behavior. We have contemplate the features of real social networks in order to achieve our goal to infer the most possible relationships in an email social network. There are several papers that review the evolution of different types of real networks [27]. Other works utilize communication patterns in the dataset Enron email to: detect social tensions [11]; discover structures within the organization [10]; identify the most relevant actors in the network over time [44]. A more detailed work studied more than 100 real-world networks to reveal clusters or communities, the authors note that large networks have a very different structure compared to the small-world networks [28]. And there is an inverse relationship between the size of the community and the high quality of the community. The largest networks of 100 nodes do not show good conductivity which can be translated as not having the ability to be a good community; the best communities are quite small, in the range of 10 to 100 nodes.

2.2 Privacy in social network

The privacy concerns about social networks analysis has been considered for several years. There is no doubt that social networks are increasing interest in the database, mining and

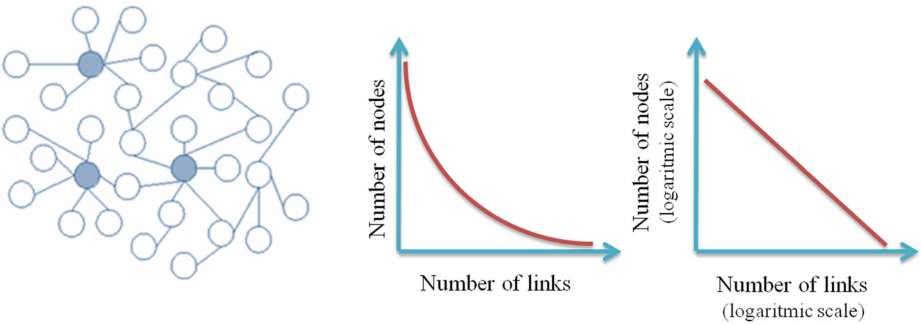

Fig. 3 Power law distribution

theory communities. Hiding identities of a social network members in order to maintain its privacy, has still a lot of open problems. One of them is the use of background knowledge to map from individuals with known identities to anonymized nodes. In order to clarify how an attacker can take advantage of context information, it is important to distinguish between passive attacks and active attacks. In the first case, attackers just observe data flow, while in the second case, attackers actively manipulate nodes before anoymization to reveal identities network users [17]. There are techniques on privacy preservation in social networks focused on different ways such as edge modification [30], randomization for network structure [48], prevent identity disclosure [5], among others. Besides, there is not warranty for users to protect their data from operators. Social networks' users require protection against malicious entities. It has been proposed architecture to protect personal information from the social network operators and other users [9].

The current problems of maintain privacy in social networks can be classified in three items: i) Each single method has been designed for a particular network; ii) There are a lot of network measures, so it is difficult to know if the network performance is optimal because there are not an standardized platform; iii) it must be considered the temporal information in order to obtain accurate results.

Traffic analysis is used to derive information from the patterns of a communication system. It has been shown that encryption by itself does not guarantee anony-mity. Even when the content of the communications is encrypted, the routing information must be clearly sent. These attacks get the most likely set of friends of a particular user, by carrying out the intersection of the anonymous sets receiving the messages user sends. One of the most widely used mechanisms for the protection of this type of attacks is the implementation of mixes.

In literature of statistical disclosure attacks, the hypotheses are overly demanding and unrealistic. For example, it were supposed scenarios in which messages had to be sent with uniform probability by all users, previous knowledge of the number of friends of a user or some network parameters, similar behaviors for all users like the average of messages sent or received. To our knowledge, it was the first time that a statistical disclosure attack was applied to email data or social networks data to detect relationships between users. The method presented in [35] leads to results in different dimensions: estimation of the number of messages sent by round or unit of time for each pair sender-receiver, ordering of the pairs from highest likelihood of communication to lowest, hard classification of pairs of users in communication-not communication. And, we have no restriction about the number of friends each user has [14, 15]. Later, a second version of the procedure, including a second pass on the data using the EM algorithm was presented in [36]. This improvement obtains better estimation of messages sent by users and detects which users really communicate. Each user i sends messages in each round to user j according to a Poisson distribution with rate $\lambda_{i,j}$. Users who do not communicate with each other will have a rate $\lambda_{i,j} = 0$. It has been shown a classification rate using three different methods: (i) uniform distribution, (ii) EM algorithm with Poisson distribution, (iii) EM algorithm with discrete tabulated distribution. One of the major results derived is the occasional detection of some pairs that have certainly communicate (without any doubt, based on combinatorial deduction) and the detection of some pairs that did never communicate in the time horizon of the attack. In this work we have applied the last method of EM algorithm with discrete tabulated distribution. In each step t, where t is the number of rounds, the following two steps are performed:

1. The first step is known as the Expectation step, it calculates the hope under a distribution Z conditioned to values of X and θ. Where X are the values of the marginal and θ is the parameter vector of λ.

Table 2 Results of Faculty A for 3 months observations

	Batch	Nodes	Edges	Avarage degree	Density	Clustering coefficient
	10	85	406	4.776	0.057	0.335
Estimation	30	85	406	4.776	0.057	0.335
	50	85	403	4.741	0.056	0.334
Real data	–	85	406	4.776	0.057	0.335

2. In the second step, known as the Maximization step, a new θ value is obtained. It was considered that attacker knows the number of messages sent and received by each user.

We observed e-mail data patterns are very specific. Details about a statistical disclosure attack to estimate the network and node characteristics of an anonymized email network such as power law coefficient, centrality and clustering measures, degree distribution and small-world-ness are described in [37]. The estimations allow identifying the evolution of the networks, evaluating differences between networks, and knowing who the most influential users are. This was an innovation, because there was not previous statistical disclosure attack utilized to estimate global network characteristics or node based measures.

3 Application of the method to the estimation of email network characteristics

It has been shown that an attacker can reveal the identities of mix users by analyzing network traffic, watching the flow of incoming and outgoing messages. An attacker can get partial information to study an anonymous social network, taking into account the vulnerability to attacks capture path [12, 43]. Such attacks using the vulnerability of the network traffic to compromise the identity of users to compromise the network.

We have applied our algorithm to data provided by the Data Center of the Universidad Complutense of Madrid which were previously anonymized. Such information is divided into 32 sub domains or faculties that composed the email system. We do not consider any further information like time-stamps or content email. We assume the attacker only gets traffic information, it means the number of messages each user sends and receives in every period of time, which we call a round. A batch is the number of messages sent and received in a round. In each round not all users participate, the sender set and the receiver set are not always the same. Only a small fraction of them are active, sending and receiving messages. Figure 1 represents a round with 5 users.

Table 3 Results of Faculty A for 12 months observations

	Batch	Nodes	Edges	Avarage degree	Density	Clustering coefficient
	10	116	929	8.009	0.070	0.482
Estimation	30	116	923	7.957	0.069	0.490
	50	116	924	7.966	0.069	0.479
Real data	–	116	929	8.009	0.070	0.482

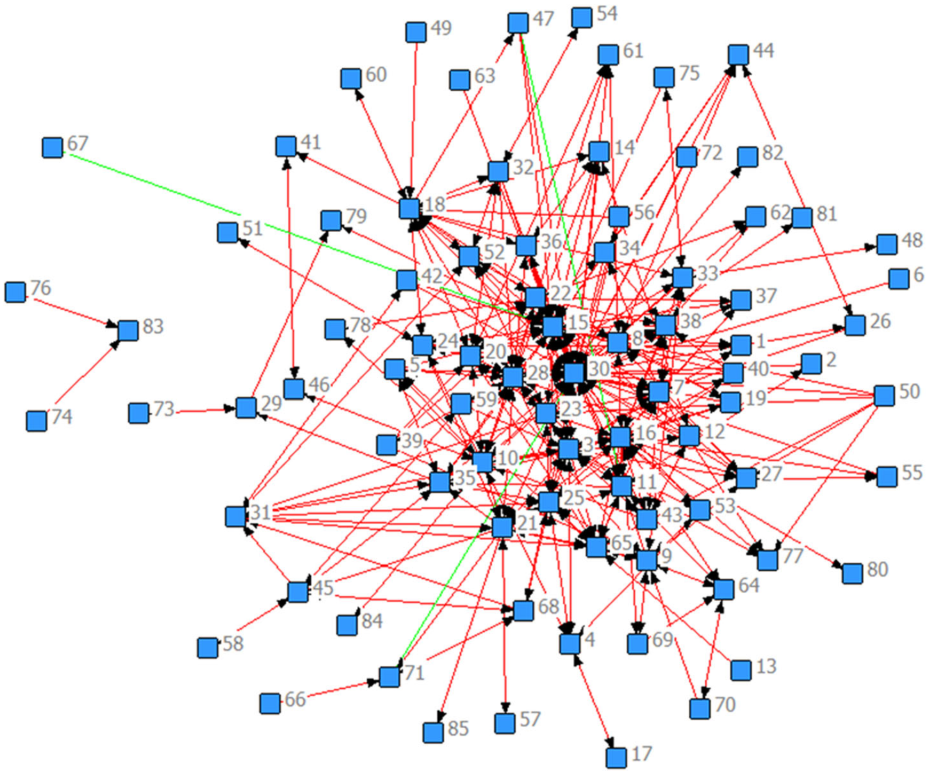

Fig. 4 Simulated vs. real graph of Faculty A for 3 months

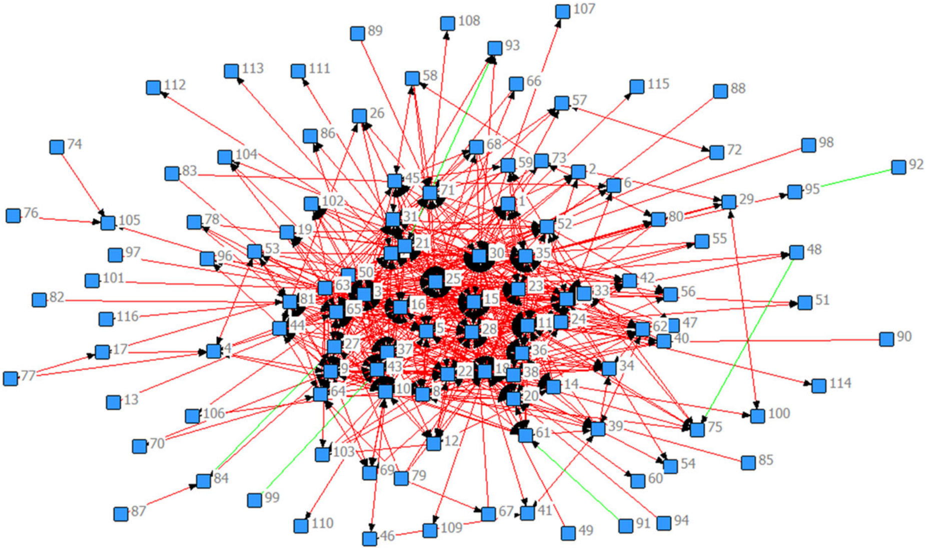

Fig. 5 Simulated vs. real graph of Faculty A for 12 months

Table 4 Five highest degree centrality nodes of Faculty A

Batch 10	Batch 30	Batch 50	Real
0.286	0.286	0.286	0.286
0.214	0.214	0.214	0.214
0.190	0.190	0.190	0.190
0.167	0.167	0.167	0.167
0.167	0.167	0.167	0.167

For demonstration purposes we have chosen only the Faculty A. Faculty A is a network composed of 85 users or nodes. In Table 2 we present the results obtained after applying our algorithm to Faculty A of 3 month data and Table 3 for 12 months data. We can see that estimations about smaller batches of messages are closer to the real values of the network. Since the information obtained consists only on the number of messages sent and received by the users in each round, the size of the rounds (batch size) is an important parameter that affects seriously the results. The first three rows show the estimated results for a batch size of 10, 30 and 50 messages. Faculty A has 85 users represented as nodes in social network, and edges are the relationships between them. The last row exhibits the real values of the network. So, we can notice smaller batches accomplish better estimations, batch sizes 10 and 30 calculate 406 edges. In Table 3 the result is similar, with a 10 batch size we have gotten better estimations of the social network real values.

In Figs. 4 and 5 we present the results obtained where: i) red edges represent the relationships disclosure among users in the anonymous network; ii) green edges are the links that our algorithm has not detected. We have placed the two overlapping graphs for three and twelve months, because of small differences. Figure 4 shows the estimated and real graph of Faculty A with a time horizon of three months. Figure 5 shows the results of the same Faculty A, but for a twelve months period. We also noted that both networks exhibit small world and scale-free characteristics.

We have utilized different batch sizes to estimate the most important nodes of each graph and we have gotten almost the same results. The incidents show that our algorithm is able to recognize who are the most influential nodes within a network despite increasing the number of nodes. The schema can be abstracted to other contexts; for example, repeated polls or elections in small populations, where the attack can be used to obtain an ordering of the likelihood users vote to some political groups or anti terrorist research, where the method can use phone calls information in repeated contexts to link senders with recipients.

Table 4 presents the five highest degrees of centrality calculated for Faculty A where the three first columns were estimated with a batch of 10, 30 and 50. The last column shows the five highest degrees of the real network. In Table 5 we show the five lowest centrality degrees. Also, the three first columns were estimated with size batches of 10, 30 and 50,

Table 5 Five lowest degree centrality nodes of Faculty A

Batch 10	Batch 30	Batch 50	Real
0.012	0.012	0.012	0.012
0.012	0.012	0.012	0.012
0.012	0.012	0.012	0.012
0.012	0.012	0.012	0.012
0.012	0.012	0	0.012

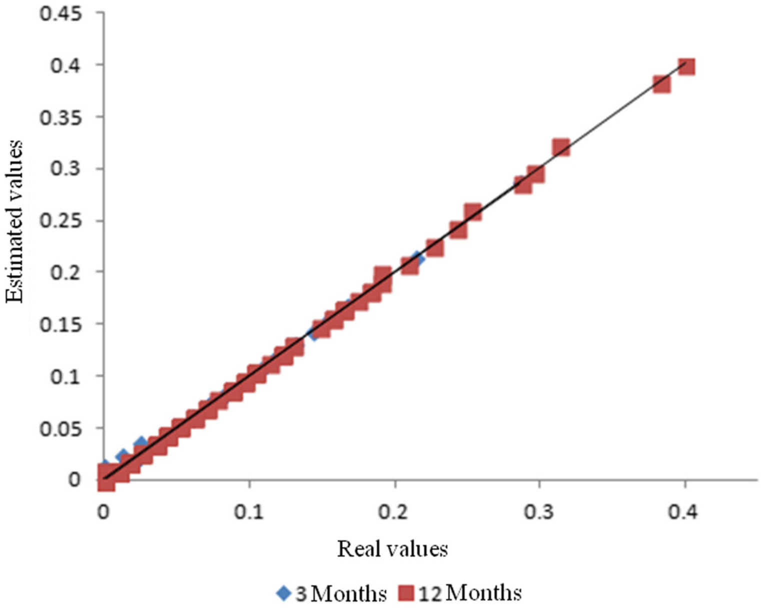

Fig. 6 Simulated vs. real graph of Faculty A for 12 months

and last column exhibit the five lowest centrality degrees of the real network. We can see how our algorithm estimates nodes more connected better.

In Fig. 6 we present the comparison of estimated and actual degrees of the Faculty A for 3 to 12 months; the closer to the diagonal point is better estimate. Otherwise, the points are above or below the diagonal.

4 Conclusions and future work

Disclosing if there exists communication or not between a pair of users in a network communication system is the object of the attacker in the present work. In this paper we have described the characteristics and metrics of social networks. Using social network analysis techniques and getting several social network measures we were able to know user's centrality to detect which are the most influential users in a network. We have applied an attack to disclosure identities on a university anonymous email system, representing such system as a social network. We showed that analysis of social networks helps to know the user's centrality to detect the most important elements in the network. From the results we found the attack performs better with small batches and that estimated graph is very similar to the real one. For future work, we have considered social network data can be used to further investigate the performance of the strategy developed here. There are also other standard databases that could be used as benchmark (Enron email data, for example). Other applications in the field of disclosure of public data could be considered.

Acknowledgements This research work was supported by Sungshin Women's University. In addition, this project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700326. Website: http://ramses2020.eu. Also, this work was funded by grants provided from PRODEP (Apoyo para gastos de publicación SEP-23-007-B).

References

1. AlFalahi K, Atif Y, Abraham A (2014) Models of influence in online social networks. Int J Intell Syst 29:161–183. https://doi.org/10.1002/int.21631
2. Ahmed HS, Faouzi BM, Caelen J (2013) Detection and classification of the behavior of people in an intelligent building by camera. SIGKDD Explor Newsl 6:1317–1342
3. Al Hasan M, Zaki MJ (2011) A survey of link prediction in social networks. In: Social network data analytics, pp 243–275
4. Anderson A, Corney M, Vel O, Mohay G (2001) Identifying the authors of suspect e-mail. In: Communications of the ACM
5. Anderson J, Diaz C, Bonneau J, Stajano F (2009) Privacy-enabling social networking over untrusted networks. In: Proceedings of the 2nd ACM workshop on Online social networks. ACM, New York, pp 1–6, https://doi.org/10.1145/1592665.1592667
6. Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512. https://doi.org/10.1126/science.286.5439.509
7. Barnes S (2006) A privacy paradox: social networking in the United States First Monday 11(9)
8. Bekkerman C, McCallum A (2004) Extracting social networks and contact information from email and the web. In: Proceedings of CEAS-1
9. Bonchi F, Castillo C, Gionis A, Jaimes A (2011) Social network analysis and mining for business applications. In: ACM Trans. Intell. Syst. Technol., vol 2, pp 1–37. https://doi.org/10.1145/1961189.1961194
10. Chapanond A, Krishnamoorthy MS, Yener B (2005) Graph theoretic and spectral analysis of enron email data. Comput Math Organ Theory 11:265–281. https://doi.org/10.1007/s10588-005-5381-4
11. Collingsworth B, Menezes R (2009) Identification of social tension in organizational networks, vol 207. Complex Networks Studies in Computational Intelligence
12. Culotta A, Bekkerman R, McCallum A (2004) Extracting social networks and contact information from email and the web. In: Proceedings of the first conference on email and anti-spam (CEAS)
13. Danezis G, Serjantov A (2004) Statistical disclosure attack or intersection attacks on anonymity systems. Inform Hiding 3200:293–308
14. Danezis G, Troncoso C (2009) Vida: how to use Bayesian inference to de-anonymize persistent communications. In: Proceedings of the 9th international symposium of privacy enhancing technologies. Seattle, pp 56–72
15. Danezis G, Diaz C, Troncoso C (2007) Two-sided statistical disclosure attack. In: Proceedings of the 7th international conference on privacy enhancing technologist workshop. Ottawa, pp 30–44
16. Danezis G, Díaz C, Syverson P (2009) System for anonymous communication
17. Ding X, Zhang L, Wan Z, Gu M (2013) De-anonymization of dynamic social network. Inf Technol J 12(19):4882–4888. https://doi.org/10.3923/itj.2013.4882.4888
18. Ebel H, Mielsch LI, Bornholdt S (2002) Scale-free topology of e-mail networks. Phys Rev E - Stat Nonlinear, Soft Matter Phys, 66. https://doi.org/10.1103/PhysRevE.66.035103
19. Erdos P, Rényi A (1959) On random graphs. Publ Math 6:290–297. https://doi.org/10.2307/1999405
20. Garg V, Jean Camp L (2015) Cars, condoms, and facebook. In: Information Security, lecture notes in computer science, vol 7807. Springer
21. Gross R, Acquisti A (2005) Information revelation and privacy in online social networks. In: ACM workshop on privacy in the electronic society. ACM, New York, pp 71–80. https://doi.org/10.1145/1102199.1102214
22. Hansen M, Jensen M, Rost M (2015) Protection goals for privacy engineering. In: IEEE Security and Privacy Workshops, San José, pp 159–166, https://doi.org/10.1109/SPW.2015.13
23. Hopper N, Vasserman EY (2006) On the effectiveness of k;-anonymity against traffic analysis and surveillance. In: Proceedings of the 5th ACM workshop on privacy in electronic society, WPES ’06. ACM, New York, pp 9–18. 10.1145/1179601.1179604

24. Imizcoz J (2001) Introducción actores sociales y redes de relaciones: reflexiones para una historia global, pp 19–30
25. Kavitha S, Sivaraman E, Raja Vadhana P (2014) A suvery on k-anonymity generalization algorithms. Int J Adv Res Comput Commun Eng, 2(11)
26. Knoke D, Yang S (2008) Social networks analysis. Sage
27. Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. In: ACM Trans. Knowl. Discov. Data. https://doi.org/10.1145/1217299.1217301
28. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2011) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. In: Internet Mathematics, vol 6. https://doi.org/10.1080/15427951.2009.10129177
29. Liu Y, Kou Z (2007) Predicting who rated what in large-scale datasets. ACM SIGKDD Explor Newsl 9(2):62–65. https://doi.org/10.1145/1345448.1345462
30. Liu L, Liu J, Zhang J (2010) Privacy preservation of affinities in social networks. In: Proceedings of the International conference on information systems, pp 372–376
31. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) L-diversity: privacy beyond k-anonymity. In: ACM Trans. Knowl. Discov. Data, vol 1. https://doi.org/10.1145/1217299.1217302
32. McCallum A, Wang X, Corrada-Emmanuel A (2007) Topic and role discovery in social networks. J Artif Intell Res 30:249–272
33. Moradi F, Olovsson T, Tsigas P (2012) Towards modeling legitimate and unsolicited email traffic using social network properties. In: Proc. Fifth Work. Soc. Netw. Syst. - SNS ’12, pp 1–6. https://doi.org/10.1145/2181176.2181185
34. Ninghui L, Tiancheng L, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and l-diversity. In: IEEE 23rd International conference on data engineering 2007, pp 106–115. https://doi.org/10.1109/ICDE.2007.367856
35. Portela J, García-Villalba L, Silva Trujillo A, Sandoval A, Kim T (2015) Extracting association patterns in network communications. Sensors, 15(2)
36. Portela J, García-Villalba L, Silva Trujillo A, Sandoval A, Kim T (2016) Disclosing user relationships in email networks. J Supercomput 72:3787–3800
37. Portela J, García-Villalba L, Silva Trujillo A, Sandoval A, Kim T (2016) Estimation of anonymous email network characteristics through statistical disclosure attacks. Sensors, 16(11). https://doi.org/10.3390/s16111832
38. Socievole A, De Rango F, Marano S (2013) Link prediction in human contact networks using online social ties. In: International Conference on cloud and green computing, pp 305–312. https://doi.org/10.1109/CGC.2013.55
39. Spiekermann S, Korunovska J (2016) Towards a value theory for personal data. J Inf Technol 32(1):62–84. https://doi.org/10.1057/jit.2016.4
40. Streeter C, Gillespie D (1992) Social network analysis. J Soc Serv Res 16:201–222
41. Sweeney L (2002) K-anonymity: a model for protecting privacy. In: Int. J. Uncertain. Fuzziness Knowl.-Based Syst., pp 557–570. https://doi.org/10.1142/S0218488502001648
42. Tyler J, Wilkinson D, Huberman B (2003) Email as sprectroscopy: automated discovery of community structure within organizations. In: Proceedings of communities and technologies, pp 81–96
43. Tyler J, Wilkinson D, Huberman B (2005) E-mail as spectroscopy: automated discovery of community structure within organizations. Inf Soc 21(2):143–153
44. Uddin M, Murshed STH, Hossain L (2010) Towards a scale free network approach to study organizational communication network. In: Proceedings of Pacific Asia conference on information systems, vol 196, 415, pp 1937–1944
45. Van Alstyne M, Zhang J (2003) Emailnet: automatically mining social networks from organizational email communications. In: Proceedings of Annual conference of the North American association for computational social and organizational sciences. Pittsburg
46. Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press
47. Wey T, Blumstein T, Shen W, Jordan F (2008) Social network analysis of animal behaviour: a promising tool for the study of sociality. Anim Behav 75(2):333–344
48. Wu X, Ying X, Liu K, Chen L (2010) A survey of privacy-preservation of graphs and social networks. In: Managing and mining graph data, pp 421–453

Alejandra Guadalupe Silva Trujillo received a Computer Science degre from the Universidad Autónoma de San Luis Potosí (San Luis Potosí, México) in 2001 and holds a M.Sc. in Information Systems (2003) from Fundación Arturo Rosenblueth (México) and a Ph.D. in Computer Science (2016) from the Universidad Complutense de Madrid (Spain). She is currently a professor researcher at Universidad Autónoma de San Luis Potosí. Her main research interest are information security, anonymity and privacy enhancing technologies.

Ana Lucila Sandoval Orozco received a Computer Science Engineering degree from the Universidad Autónoma del Caribe (Colombia) in 2001. She holds a Specialization Course in Computer Networks (2006) from the Universidad del Norte (Colombia), and holds a M.Sc. in Research in Computer Science (2009) and a Ph.D. in Computer Science (2014), both from the Universidad Complutense de Madrid (Spain). She is currently a postdoctoral researcher at Universidad Complutense de Madrid (Spain). Her main research interests are coding theory, information security and its applications.

Luis Javier García Villalba received a Telecommunication Engineering degree from the Universidad de Málaga (Spain) in 1993 and holds a M.Sc. in Computer Networks (1996) and a Ph.D. in Computer Science (1999), both from the Universidad Politécnica de Madrid (Spain). Visiting Scholar at COSIC (Computer Security and Industrial Cryptography, Department of Electrical Engineering, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium) in 2000 and Visiting Scientist at IBM Research Division (IBM Almaden Research Center, San Jose, CA, USA) in 2001 and 2002, he is currently Associate Professor of the Department of Software Engineering and Artificial Intelligence at the Universidad Complutense de Madrid (UCM) and Head of Complutense Research Group GASS (Group of Analysis, Security and Systems) which is located in the School of Computer Science at the UCM Campus. His professional experience includes research projects with Hitachi, IBM, Nokia and Safelayer Secure Communications.

Tai-Hoon Kim received his M.S. and Ph.D. degrees in Electrics, Electronics & Computer Engineering from the Sungkyunkwan University, Korea. And he got his 2nd Ph.D. in Computer Engineering from Bristol University, United Kingdom. After working with Technical Institute of Shindoricoh for 2 years as a researcher and working at the Korea Information Security Agency as a senior researcher for 2 years and 6 months, he worked at the DSC (Defense Security Command) for about 2 years. After working with Hannam University for 4 years and 6 months as an associate professor, he is now an associate professor of GVSA in Australia, and fellow of UTAS in Australia. He has written 17 books about the software development, OS such as Linux andWindows 2000, and computer hacking & security. He has also published about 200 papers by 2012. He is a member of IEEE, ACM, KIIT and SERSC. He is a General Chair or Program Committee chair from more than 20 international conferences.