IEICE TRANS. INF. & SYST., VOL.E97–D, NO.6 JUNE 2014

LETTER

Analyzing Network Privacy Preserving Methods: A Perspective of Social Network Characteristics*

Duck-Ho BAE[†], **Jong-Min LEE**[†], **Sang-Wook KIM**[†a)], **Youngjoon WON**[††], ***and*** **Yongsu PARK**[†], ***Nonmembers***

SUMMARY A burst of social network services increases the need for in-depth analysis of network activities. Privacy breach for network participants is a concern in such analysis efforts. This paper investigates structural and property changes via several privacy preserving methods (anonymization) for social network. The anonymized social network does not follow the power-law for node degree distribution as the original network does. The peak-hop for node connectivity increases at most 1 and the clustering coefficient of neighbor nodes shows 6.5 times increases after anonymization. Thus, we observe inconsistency of privacy preserving methods in social network analysis.

key words: *network privacy preserving, privacy breach, structural disparity*

1. Introduction

A burst of social network services (SNS) leads to a heavy competition in the market. SNS providers focus on analyzing their networks and extracting valuable information to differentiate the service quality from each other [1]–[4]. They often publish their networks to outside experts for in-depth analysis; however, a risk of revealing individual and relationship information is increasing. Social network adversary can use such information, especially relationship information, for privacy breach [1].

Recently, there are various studies of privacy preserving for social network participants [5]–[14]. These studies first built an attack model for privacy and then proposed their own preserving method against it. A common approach is to add phony or delete legitimate relationships, resulting in structural change. Thus, the adversary cannot track down the identity of participant and correctly observe its relationships to the others.

More structural changes can promise a higher level of privacy preserving. For the impact of privacy preserving, we question ourselves with the following.

- Are the original network and the network applying privacy preserving methods (the anonymized network) sharing similar structural properties?
- If they show differences, how much are they apart? Can we measure the disparity between?
- Can we rely on the analysis result from the anonymized networks?

In this paper, we investigate the impact of applying privacy preserving methods to the social network. We use three features to measure a structural disparity: (1) degree distribution, (2) hop plot, and (3) average clustering coefficient. We observe that the degree distribution does not follow the power-law after anonymization whereas the original network does [15]. There is 1-hop change in peak-hop count and showing 6.5 times higher in average clustering coefficient. To measure a difference in the analysis results, we employ link-based similarity and ranking algorithms on the original and the anonymized networks, respectively. Both show low mean average precision and precision values. Overall, anonymization leads to a massive structural change.

This paper is organized as follows. Section 2 summarizes the existing privacy preserving methods. Section 3 describes the dataset and analysis features. Section 4 discusses the differences between the original and anonymized networks. Finally, we conclude our paper in Sect. 5.

2. Privacy Preserving Methods

This section explains several privacy preserving methods for social network participants. Adversary relies on the *background knowledge* to identify the target individual where it refers to the structural information of node (individual) [16], [17]. The background knowledge includes such as node degree, connectivity, sub-networks of the target, and etc.

Naïve anonymization for tabular data randomizes the node ID [7], [18]. However, we can easily track down the true identity of node via node degree information, which is preserved regardless of ID change. Figure 1 (a) and (b) show the original network and its corresponding ID-anonymized network, respectively. Assuming the target node is ID2, the node degree of ID2 leads to Bob in the original network because it is the only node of degree 4. It is difficult to protect the privacy in social networks with this approach.

The k-anonymity approach via edge modification changes the network structure to satisfy k-candidate anonymity [8]–[10]. Such manipulation of network can

Manuscript received December 16, 2013.

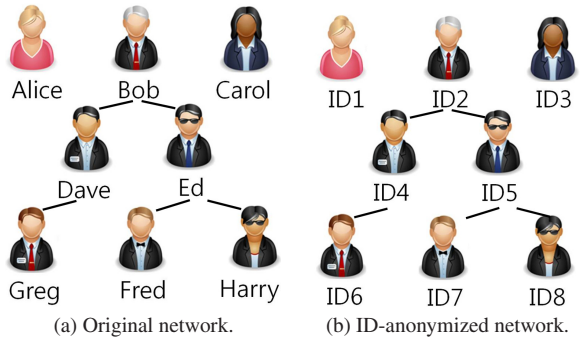
[†]The authors are with the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea.

[††]The author is with the Department of Information System, Hanyang University, Seoul, Korea.

*This work was supported by (1) the National Research Foundation of Korea (NRF) (No. 2011-0029181), (2) Business for Cooperative R&D between Industry, Academy, and Research Institute in 2013 (No. C0006278), (3) Ministry of Culture, Sports and Tourism (MCST) and Korea Copyright Commission in 2013, and (4) the Ministry of Science, ICT and Future Planning (MSIP) of Korea, under the ITRC program (NIPA-2013-H0301-13-4009).

a) E-mail: wook@agape.hanyang.ac.kr (Corresponding author)

DOI: 10.1587/transinf.E97.D.1664



(a) Original network. (b) ID-anonymized network.

Fig. 1 Naïve anonymization for social networks.

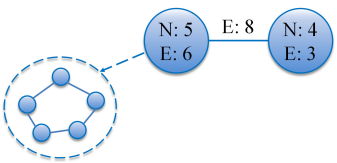


Fig. 2 Graph generalization.

have more than k different responses to a single structural query by the adversary. Thus, it guarantees the adversary's chance for a successful privacy breach below the probability of $1/k$. There are two methods adapting this approach: k-degree anonymization (KD) [8] and k-automorphism anonymization (KA) [10]. KD inserts extra edges to create at least k nodes with the same degree. Meanwhile, KA adds edges to create at least k sub-networks sharing the identical structure to avoid the adversary's attempt.

The adversary keeps trying to identify the existence of any edges between two target nodes. A randomization approach is to reduce the ratio of successful privacy breach incidents and to provide fake responses to the adversary by inserting, deleting, or switching arbitrary edges. The random deletion (RD) method regulates the probability of edge existence among the groups of nodes clustered by node degrees [11]. It leaves a large room for error in predicting relationship between the target nodes. The random switching (RS) method is to hide edges of the target and its connected sub-networks [12]. Thus, the adversary can no longer rely on the edge information in the anonymized network.

Finally, the cluster-based graph generalization method aggregates nodes and edges into groups and makes the group statistics available for reconstruction [13]. Figure 2 shows the node and edge counts of each group and also the edge counts among the groups themselves. However, this method cannot guarantee a consistent network reconstruction. Therefore, in this paper, we do not consider this method as a privacy preserving choice.

3. Dataset and Features

Our dataset is collected from the Epinion website [19] having 3,094 nodes, 9,680 edges, and 3.13 node degrees on average. Each node represents a single user and edge does a recommendation between two users.

We analyze the structural change and differences in analysis results between the original and anonymized networks. For analyzing structural change, we use the following features: node degree distribution, hop-plot, and average clustering coefficient. To spot differences in the analysis results, we use mean average precision and precision.

- Node degree distribution: It presents distribution of nodes having same degree [15]. In general, the degree distribution of social networks follows the power-law. The power-law is expressed as $y = ax^e$ where x is a node degree and y is the number of nodes having the corresponding degree.
- Hop-plot: Hop is a distance between two nodes, and hop-plot shows the distribution of hop counts for all connected node pairs in the network [15]. Hop-plot is closely related to the average node degree. If the average node degree is high, then most node pairs are reachable to each other within a small hop counts.
- Average clustering coefficient (CC): It measures strength of connectivity between node and its neighbors [15]. In Eq. (1), $CC(i)$ shows the ratio of edges among the neighbors of node i to all nodes. N_i is a set of the neighbors of node i and e_{jk} is an edge between node j and k. The average CC distribution represents the average value of CC of the nodes sharing the same degrees. If the average CC is close to 1, it means that we have strong connectivity among its neighbor nodes. It becomes opposite towards 0. Thus, if the average node degree becomes high, then the average CC value increases.

$$CC(i) = \frac{|\{e_{jk}\}|}{|N_i|(|N_i| - 1)/2} \qquad (j, k \in N_i) \tag{1}$$

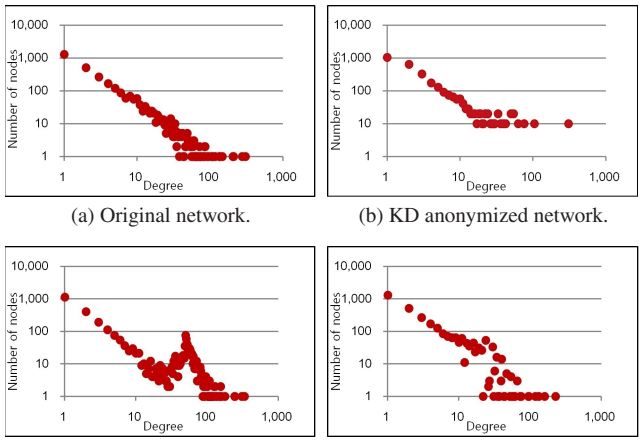
- Precision & mean average precision (MAP): We use the analysis result of the original network as the ground truth and compare it to the analysis result of the anonymized networks. We rely on the precision value for ranking analysis and MAP for similarity measure. Precision describes the ratio of correct match in ranking analysis between the anonymized result and the ground truth [20]. MAP is an average precision value of node similarity measures [20]. The precision & MAP values are towards 1 if the responses to the query in the anonymized networks are similar to those in the original network.

4. Observation

This section exploits changes in network properties via privacy preserving methods. First, we demonstrate structural changes after anonymization in degree distribution, hop-plot, and average CC. Second, we investigate any difference in the social networks analysis results between the original and anonymized networks. We use ranking [23] and link-based similarity measures [20].

Table 1 Parameter values and edge difference.

Method	Parameter value	Edge difference
KD	10 (k-candidate anonymity)	+1,514
KA	5 (k-candidate anonymity)	+8,366
RD	0.7 (probability of edge existence)	−447
RS	3,000 (#edge switching)	0



(a) Original network. (b) KD anonymized network.

(c) KA anonymized network. (d) RD anonymized network.

Fig. 3 Degree distributions.

Table 1 demonstrates the parameter values of privacy preserving methods and the numbers of edge difference between original and anonymized networks.

4.1 Structural Change

Figure 3 presents the degree distributions of each anonymization. The original network (Fig. 3 (a)) follows the power-law distribution with slope of −1.504. The KD anonymized network (Fig. 3 (b)) shows that a group of high degree nodes does not follow the power-law. In typical social networks, the high degree nodes (e.g., social hub) are rare. These nodes do not satisfy k-candidate anonymity; thus, KD adds edges in the original network to create more high degree nodes. We now have at least k nodes contributing to the break of the power-law (slope: −0.974). This break continues for KA where it adds more edges to the sub-networks of the nodes in the middle of distribution (e.g., 10∼100 degree nodes in Fig. 3 (c), slope: −1.282). It is likewise for RD; in fact, it removes the edges belonging to the high degree nodes. The RD network presents that the 10∼100 degree nodes do not satisfy the power-law (Fig. 3 (d)). The slopes of KD and KA are less steep than the original network's slope because these methods eventually increase the number of nodes with higher degrees.

Tables 2 and 3 illustrate the ratios of reachable node pairs to all pairs at each hop and the average CC among the nodes of degree under 50. The nodes under 50 show a clear distinction between the original and anonymized networks.

The ratio of reachable node pairs to all pairs at each hop reflects any change in the peak-hop count, which connects the most node pairs in the network. The KD and KA anonymized networks introduce stronger connectivity due to additional edges; thus, their peak-hop count is down to 3 from 4 in the original network. In RD, the proportion of reachable node pairs is down 7% in 3-hops and up 3% in 4-hops. This observation implies weaker node connectivity after anonymization. RS does not interfere with the edge density. However, the paths between the nodes are deviated. Its hop-plot distribution changes even if it shows the smallest change.

Table 2 Ratios of reachable node pairs to all pairs at each hop (unit: %).

	2-hop	3-hop	4-hop
Org	5.6	31.8	40.5
KD	8.6	43	33.1
KA	11.2	42.9	31.4
RD	4.9	24.5	43.4
RS	5.2	34.1	45.2

Table 3 Average CC of each degree interval.

	1∼10	11∼20	21∼30	31∼40	41∼50
Org	1	1	1	1	1
KD	1.29	1.44	1.96	2	2.64
KA	1.02	1.19	6.52	5.31	4.23
RD	0.82	0.76	0.72	0.57	0.52
RS	0.94	0.95	0.95	0.96	0.95

We provide the relative ratio of each anonymization by setting the average CC of the original network to 1. If the ratio is greater than 1, the connectivity of the anonymized network increases. If it is less than 1, the connectivity decreases. In KD, the average CC increases, especially for the nodes of degree 31∼50. It is likewise for KA; the average CC of degree 21∼40 increases more than 6.5 times compared to the original network. Meanwhile RD and RS did not show any significant change.

4.2 Difference in Analysis Results

Ranking is to measure the authority of each node in the network [22], [23]. We verify whether the highly authoritative nodes in the original network maintain a similar status in the anonymized. We use a well-known PageRank technique to compute the authority [23]. To compare the ranking results, we extract the top 30 authoritative nodes via PageRank and then calculate the precision values of each network.

Table 4 shows the low precision values of anonymization in overall. The value of RD is relatively high where it does not cause too much of structural change. The precision value of RS is the lowest. This is because RS initiates around 3,000 edge swaps in our experiments and results in 5,000∼6,000 edge changes among all 9,680 edges.

The link-based similarity measure is based on the proportion of common neighbors shared by the two nodes [20], [21]. So, the link-based similarity between two nodes increases when they have more common neighbors. For comparison, we did the following steps:

- **Step 1.** Calculate the similarity measures for all node

Table 4 Precision of ranking.

	Precision
Org	1.00
KD	0.37
KA	0.47
RD	0.60
RS	0.30

Table 5 MAP of link-based similarity.

	Org	KD	KA	RD	RS
High degree nodes	1.00	0.66	0.61	0.82	0.38
Mid degree nodes	1.00	0.70	0.37	0.90	0.22
Low degree nodes	1.00	0.63	0.22	0.65	0.20

pairs in both the original and anonymized networks.

- **Step 2.** Classify them into the groups of 'high', 'mid', and 'low' according to the node degree. Each group has the same number of nodes.
- **Step 3.** Select top 10 nodes having highest similarity among arbitrary 100 nodes from each section.
- **Step 4.** Calculate the MAP values for each section.
- **Step 5.** Repeat 1∼4 for 10 times to remove random effect.

Table 5 shows the MAP values of each anonymization. The results show low MAP values in all anonymization methods. In KA, MAP values of mid and low groups are relatively low compared to that of high group. This is because KA adds edges to the nodes having 10∼100 degrees in priority as shown in Fig. 3 (c). RD gets the best of the others where it caused minimal structural change. RS results in the smallest MAP value because the edge switching deviates the structures in the original network.

5. Conclusions

This paper investigated network structural and property changes via several anonymization techniques for social networks. The contributions are following:

- The anonymized social networks do not follow the power law for degree distribution. The peak-hop increases at most 1 and the average CC shows a maximum 6.5 times increase after anonymization.
- The MAP value was 0.53. The ranking shows the precision value of 0.45 on average. These lower values imply that the analysis result of the anonymized networks is not consistent with that of the original network.

For future work, we plan to develop more measures to quantify the stability of privacy preserving results.

References

[1] J. Kleinberg, "Challenges in mining social network data: Processes, privacy, and paradoxes," ACM SIGKDD, pp.4–5, 2007.
[2] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, "A framework for community identification in dynamic social networks," ACM SIGKDD, pp.717–726, 2007.
[3] T. Berger-Wolf and J. Saia, "A framework for analysis of dynamic social networks," ACM SIGKDD, pp.523–528, 2006.
[4] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," ACM SIGKDD, pp.611–617, 2006.
[5] X. Liu, B. Wang, and X. Yang, "Efficiently anonymizing social networks with reachability preservation," ACM CIKM, pp.1613–1618, 2013.
[6] L. Hermansson, T. Kerola, F. Johansson, V. Jethava, and D. Dubhashi, "Entity disambiguation in anonymized graphs using graph kernels," ACM CIKM, pp.1037–1046, 2013.
[7] M. Hay et al., "Anonymizing social networks," Technical Report 07-19, University of Massachusetts Amherst, 2007.
[8] K. Liu and E. Terzi, "Towards identity anonymization on graphs," ACM SIGMOD, pp.93–106, 2008.
[9] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," IEEE ICDM, pp.506–515, 2008.
[10] L. Zou, L. Chen, and M. Ozsu, "*k*-automorphism: A general framework for privacy preserving network publication," VLDB Journal, vol.2, no.1, pp.946–957, 2009.
[11] L. Zhang and W. Zhang, "Edge anonymity in social graphs," CSE, pp.1–8, 2009.
[12] X. Ying and X. Wu, "Randomizing social networks: A spectrum preserving approach," SDM, pp.739–750, 2008.
[13] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting structural re-identification in anonymized social networks," VLDB Journal, vol.1, no.1, pp.102–114, 2008.
[14] E. Zheleva and L. Getoor, "Preserving the privacy of sensitive relationships in graph data," ACM PinKDD, pp.153–171, 2007.
[15] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," ACM SIGCOMM Computer Communication Review, vol.29, no.4 pp.251–262, 1999.
[16] K. Liu, K. Das, T. Grandison, and H. Kargupta, "Privacy-preserving data analysis on graphs and social networks," Next Generation Data Mining, 2008.
[17] B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," ACM SIGKDD Explorations, vol.10, no.2, pp.12–22, 2008.
[18] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural steganography," WWW, pp.181–190, 2007.
[19] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the semantic Web," ISWC, pp.351–368, 2003.
[20] S. Yoon, S. Kim, and S. Park, "A link-based similarity measure for scientific literature," WWW, pp.1213–1214, 2010.
[21] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," ACM SIGKDD, pp.538–543, 2002.
[22] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," Technical Report, Stanford University, 1999.
[23] H. Tong, C. Faloutsos, and J. Pan, "Fast random walk with restart and its applications," IEEE ICDM, pp.613–622, 2006.