A Comparison of Two Different Types of Online Social Network from a Data Privacy Perspective

David F. Nettleton[1,2], Diego Sáez-Trumper[2], and Vicenç Torra[1]

[1] Artificial Intelligence Research Institute, IIIA
Spanish National Research Council, CSIC
Campus Universitat Autònoma de Barcelona 08193 Bellaterra, Catalonia, Spain
{Dnettleton,vtorra}@iiia.csic.es
[2] Pompeu Fabra University, c/Tanger 122-140
08018 Barcelona, Spain
diego.saez01@upf.edu

Abstract. We consider two distinct types of online social network, the first made up of a log of writes to wall by users in Facebook, and the second consisting of a corpus of emails sent and received in a corporate environment (Enron). We calculate the statistics which describe the topologies of each network represented as a graph. Then we calculate the information loss and risk of disclosure for different percentages of perturbation for each dataset, where perturbation is achieved by randomly adding links to the nodes. We find that the general tendency of information loss is similar, although Facebook is affected to a greater extent. For risk of disclosure, both datasets also follow a similar trend, except for the average path length statistic. We find that the differences are due to the different distributions of the derived factors, and also the type of perturbation used and its parameterization. These results can be useful for choosing and tuning anonymization methods for different graph datasets.

Keywords: Social network, data privacy, descriptive statistics, risk of disclosure, information loss.

1 Introduction

Data Privacy in Social Network logs is now an important issue, given that millions of users worldwide are generating high volume data logs of their online social network activity and relations. This data offers a great analysis opportunity to data miners, but on the other hand, it may represent a threat to an individual's data privacy if it falls into the wrong hands. However, if we can sufficiently protect the data by anonymization techniques, then we can publish the social network log data for commercial and academic use.

In the current work we statistically compare and anonymize two real datasets represented as a graph, from a data privacy perspective: the Enron emails dataset [1] and the Facebook New Orleans dataset [2]. We calculate descriptive statistics for the graphs: degree, clustering coefficient and average path length. Then we anonymize/

V. Torra et al. (Eds.): MDAI 2011, LNAI 6820, pp. 223–234, 2011.

perturb the datasets by randomly adding links to the nodes and calculate the information loss and risk of disclosure for different degrees of perturbation.

The structure of the paper is as follows: in Section 2 we present the state of the art and related work; in Section 3 we define the basic data and derived factors used to describe the graphs; in Section 4 we present the statistics calculated for both graphs and make comments and comparisons; in Section 5 we calculate the information loss and risk of disclosure for both datasets, for different degrees of perturbation; finally, in Section 6 we summarize the present work.

2 State of the Art and Related Work

Privacy in on-line social networks is a relatively new area of research which however has a solid base in classic graph theory and data privacy concepts.

We will consider the state of the art from two main perspectives: the statistical analysis of online social networks, and data privacy analysis of online social networks. In terms of data privacy in general, we can cite Sweeney's paper on k-anonymity[3], and more recently [4], in which key definitions are given for information loss and risk of disclosure.

In the field of the statistical analysis of online social networks, some key authors are: Kumar[5], Ahn[6], Klienberg[7,8], Mislove[9], Shetty[1] and Viswanath[2]. In [1], Shetty et al. present some concepts related to 'graph entropy' and the identification of 'important' or 'interesting nodes. The study is specifically applied to the Enron email dataset. The basic idea is to measure the effect of removing a node from a graph, as the difference between the 'entropy' of the graph before and after removing the given node. In [9], Mislove defines some of the key metrics which characterize a social network. Viswanath in [2] performs a statistical analysis of the New Orleans Facebook dataset (the dataset we use in the present work), using the degree, clustering coefficient and average path length statistics to evaluate social network evolution over time. Klienberg[7,8] considers data mining of online social networks, defining different possible topologies within OSNs and making considerations about the computational cost of data processing.

In the field of data privacy analysis applied to online social networks, we can cite Hay[10], Zhou[11], Wondracek[12] and Liu[13]. Hay[10] presents a simple graph anonymization based on random addition and deletion of edges. The attack method attempts re-identification using two types of queries, vertex refinement and sub-graph knowledge. The risk measure is considered as the percentage of nodes whose equivalent candidate set falls into one of a given set of buckets (1 node, 2-4 nodes, 5-10 nodes, ...). The information loss measure calculates some common graph metrics (clustering coefficient, path length distribution, degree distribution, ...) in the graph before and after anonymization. The information loss is considered from the point of view of an analyst who consults these statistical properties. Zhou[11] presents a more sophisticated anonymization algorithm which firstly generalizes vertex labels and secondly adds edges. One of the precepts of the approach is to create local topologies which are isomorphic with other local topologies, achieved by adding edges to them. Wondracek[12] presents a different approach, in that the attacker uses a malicious website to obtain information about users of an on-line social network. Finally, Liu, in

[13], presents a defense method which is k-anonymous, that is, it produces k-degree anonymized degree sequences.

3 Definition of Basic Data and Derived Factors

In this Section we present the datasets used and their data format. We also define the derived statistical factors which we later use to calculate the information loss and risk of disclosure for the graphs.

The Enron email dataset[1] consists of a collection of 150 folders corresponding to the emails to and from senior management and others at Enron, collected over a period between 1998 to 2002. The total number of emails sent/received between users is approx. 1.5 million. We filtered the records so as to only include users with mutual links for which at least one email was sent and received along the link. This gave us a subset of 10630 users, which we used for all the analysis in the current work. Each email sender/recipient represents a node in the graph and the activity is represented by the number of emails sent-received along the edges which connect the users. We consider the email corpus as an extension of the idea of an "online social network", useful for comparison purposes with the Facebook data.

The Facebook New Orleans dataset was generated by Viswanath et al[2] by crawling the Facebook New Orleans regional network, and consists of approx. 63,000 users, 1.5 million links between users, and 800,000 logged interactions over a two year period. We filtered the records so as to only include users with mutual links for which at least one write to wall was sent in each direction. This gave us a subset of 31720 users, which we used for all the analysis in the current work. In contrast to the Enron dataset, for which a link between users is established when an email is sent/received between them, in the case of the Facebook users, a link is established by the explicit solicitation and acceptance of friendship. Also, in the Facebook dataset, 'writes to wall' is the activity indicator.

Basic Data - Facebook: the available data consists of one file which contains writes to wall between users and their corresponding timestamps. The format of the write to wall data is {user-id 1, user-id 2, timestamp}, where the user ids are anonymous numbers between 1 and 63000. For example, {1, 2, 3-4-2010} would signify that user 1 wrote on user 2's wall on the 3rd of April, 2010. All links are reciprocal, therefore, in the dataset there will be a corresponding record: {2, 1, …..}. This is assured by only including users who reciprocally wrote on each others' walls, at least once.

Basic Data - Enron: the available data consists of separate sender and recipient files which we merged into one file and used as input to create the graph. We anonymized the emails to sequential integers. In the original files, the 'to' and 'cc' type recipients are not distinguished, following Shetty's [15] approach. This gives us a unique file with two columns of anonymized id's, the first id is that of the sender and the second is that of the recipient. In order to construct the graph and the edges, we select unique id's between sender and recipient.

Note that we consider both graphs (Facebook and Enron) as undirected in the current study, that is the degree (total number of links to a node) is considered as the in-degree (number of incoming links) + the out-degree (number of outgoing links).

Derived Factors: in order to calculate the statistics, we have implemented the algorithms which process the graphs in Java. In the case of the 'apl' (average path length) statistic, we have used Dijkstra's algorithm[14]. The following basic statistics have been calculated to describe the graph:

(i) **Degree:** number of immediate neighbors which a node has.

(ii) **Clustering Coefficient:** is an indicator of how many of the "friends" of a user, are friends of each other.

$$CC = \frac{Number\_of\_mutual\_friends\_of\_user\_i}{Total\_number\_of\_friends\_of\_user\_i} \tag{1}$$

Example: if user 1 has 30 friends, and of those 30 friends, 7 have links between them, independently of the link with user 1, then the CC for this "group" will be 7 / 30 = 0.233. For the New Orleans Facebook dataset an average CC value of 0.0257 was reported., and for Enron, 0.15.

(iii) **Average path length:** For each node *x* this is the average of the sum of the shortest number of hops required to reach every other node *y* in the graph:

$$APL(x) = \frac{\sum_{i=1}^{n}(shortest\_path\_length\_from\_node\_x\_to\_node\_y_i)}{n} \tag{2}$$

4 Descriptive Statistics for Derived Factors

In this Section we present the descriptive statistics for the Facebook and Enron datasets, and compare the two.

Firstly we will comment the Enron and Facebook correlation statistics for 'degree', 'cc' (clustering coefficient) and 'apl' (average path length). For Enron, the highest correlation was between "degree" and "apl" (-0.49), some correlation between "degree" and "cc" (-0.12), and a negligible correlation between "cc" and "apl" (-0.001). With reference to the Facebook correlation statistics, the highest correlation was between "degree" and "apl" (-0.14), followed by the correlation between "degree" and "cc" (0.12), and a negligible correlation between "cc" and "apl" (-0.037).

In Table 1 (Enron) we observe a high standard deviation of degree with respect to the average value (2 times more than its average value), whereas "cc" shows a lesser deviation and "apl" shows a significantly smaller relative deviation (7.3 times less than its average value). In Table 1 (Facebook) we observe a high standard deviation of "cc" with respect to its average value (more than twice), whereas "degree" shows a deviation slightly greater than its average value and "apl" shows a significantly smaller relative deviation (3.37 times less than its average value).

In terms of the distributions, the degree displays a typical "power law" distribution for both datasets, with just a few nodes having a very high degree. The distribution of the clustering coefficient for Facebook and Enron have different characteristics: for Facebook, In the first two quartiles and half the third quartile, all the nodes have a "cc" equal to zero, which means that none of the neighbors are interconnected between each other. The distribution of the *average path length* for both datasets shows a characteristic 'S' pattern, but in the case of Facebook the left hand ascent is displaced to the right, which implies there are more nodes with a small average path length.

Table 1. Averages and standard deviations of statistical factors for Enron and Facebook datasets

		degree	**cc**	**apl**
Enron	average	31.035	0.1556	3.1516
	standard dev.	63.384	0.1121	0.4275
Facebook	average	5.0815	0.0257	6.001
	standard dev.	6.4705	0.0528	1.7782

5 Data Privacy: Information Loss and Risk of Disclosure - Enron vs. Facebook

In this Section we present the results of Information Loss and Risk of Disclosure for the Enron and Facebook datasets, and compare the two.

5.1 Information Loss

The objective of this test is to introduce a given percentage of random perturbation into the graph data and observe the change in the graph statistics. We interpret information loss as the deviation from the original data which a data analyst (end user of the data) would perceive. We measure the information loss by calculating the correlations between the three key descriptive variables for the original graph (degree, clustering coefficient and average path length) and then for the perturbed graph. The difference will then be the information loss. That is, if $C_d o$, $C_d p$, $C_{cc} o$, $C_{cc} p$, $C_{apl} o$ and $C_{apl} p$ are the correlations of the degree, clustering coefficient and average path length, for the original graph and the perturbed graph, respectively, then:

$$Inf.Loss = \frac{|(|C_d o| - |C_d p|)| + |(|C_{cc} o| - |C_{cc} p|)| + |(|C_{apl} o| - |C_{apl} p|)|}{3} \qquad (3)$$

The correlation values are already normalized between -1 and 1, and we take the absolute value to obtain a number between 0 and 1. The difference between the correlation values is a typical statistic used in the data privacy literature. The perturbation method we have used, that of adding links to nodes, selected randomly in the graph, is also a common graph perturbation method used in the literature of graph privacy[10,11]. We add one link to each randomly selected node. Thus a perturbation of 25% means that we added one link between 25% of the nodes in the graph. Each node can only be selected once in any trial. We note that each trial (for each % perturbation) was repeated randomly three times as an experimental procedure to validate the results, and the average was taken.

Primary and Secondary (collateral) perturbation. Given the interrelated nature of graph data, if we modify a given (primary) node, other (secondary) nodes may also be affected. Our perturbation measure refers only to the number of primary nodes modified. However it is worthwhile to comment the aspect of secondary node modification, how it may affect the results, and how we could measure it. In this context, we propose that the way in which the results are affected depends on the way we define "risk of disclosure", which in our case is in terms of statistical properties such as degree, clustering coefficient and average path length, with a "hit" margin of 1%.

Given that, in our current work, the perturbation operator is "add link", then the only statistical value which will be directly modified (and which cannot be modified indirectly), is the *degree*. On the other hand, the *clustering coefficient*, in some cases, could change as a secondary effect (of joining two neighbors together, for example). Finally, the *average path length* is the statistic which would be most likely to change, if we add links to the graph. However, in general, from empirical observation of the data values before and after perturbation for the same nodes, the values only register relatively small alterations.

In conclusion, we propose that it would be reasonable to consider that the risk of disclosure (the percentage of hits), within the defined attacker "hit" margin of 1%, is equivalent to one minus the percentage of nodes affected both primarily and secondarily with a margin greater than 1%. That is:

$$DR = 1 - A_{PS} \qquad (4)$$

For example, in Table 2 we show the relation between Risk of Disclosure and the total percentage of nodes affected, for the Enron dataset with 50% perturbation.

Table 2. Relation between Risk of Disclosure and Nodes affected for the Enron dataset and 50% perturbation

Attacker query	%Hits (Risk of Disc.)	%Nodes whose values are affected more than 1% (primary and secondary)
Degree	**0.54**	**0.46***
Degree, cc	**0.49**	**0.51**
Degree, cc, apl	**0.48**	**0.52**

We observe that the percentage of affected nodes is only 46% (with margin > 1%) for 50% perturbation. This is possible because there exist a percentage of nodes with more than 100 links, thus if we add just one link to one of these nodes, the change will be less than (or equal to) 1%, and thus with this criteria will not count as having being perturbed.

Enron. In Fig. 1a we see the information loss for different percentages of perturbation on the Enron graph. On the *y-axis* a value of 0.01 represents an information loss of 1%, and on the *x-axis* a value of 0.1 represents a grade of perturbation of 10%.

We observe a fairly linear relation between the two, with a slightly steeper gradient between 25% and 100% perturbation. We note that the maximum information loss is only 7.7% at 100% perturbation. This is correct given our definition of information loss and perturbation: adding just one link to a node in a graph when the average degree is 31.03 (see Table 1) will not have a great influence on the graph overall. This allows a comparison with the results of Facebook for which the average degree is much lower at 5.08 (Table 1) and therefore we would expect that adding one link to a node will have a significantly greater effect on the graph statistics.

In Table 3 (Enron) we see the results of the tests of perturbation versus information loss on the Enron graph dataset, which are also plotted in Fig. 1a. A clear increasing and linear trend for information loss is evident in relation with increasing perturbation values.

Facebook. In Fig. 1b we see the information loss for different percentages of perturbation on the Facebook graph. On the *y-axis* a value of 0,01 represents an information loss of 1%, and on the *x-axis* a value of 0.1 represents a grade of perturbation of 10%. Similarly to the Enron dataset (Fig. 1a), we observe a fairly linear relation between the two, with a slightly steeper gradient between 25% and 100% perturbation. However, in contrast to the Enron results, the information loss is significantly greater, ranking from 2.5%, for 10% perturbation, to 27.4% for 100%

Table 3. Results of tests of perturbation *versus* information loss on the Enron and Facebook graph datasets

	Perturbation				
	10%	**25%**	**50%**	**75%**	**100%**
Enron	0.00702	0.02056	0.04009	0.06010	0.07742
Facebook	0.02519	0.06387	0.14156	0.21435	0.27491

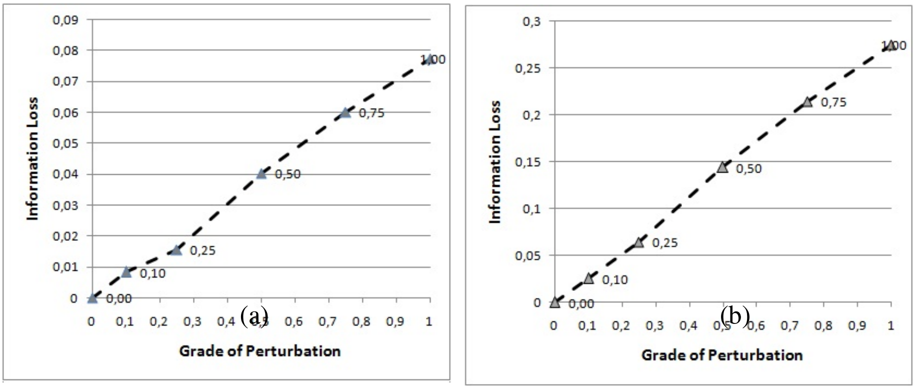


Fig. 1. Information Loss *versus* Grade of Perturbation: (a) Enron, (b) Facebook. The marker labels indicate the grade of perturbation.

perturbation. This is primarily due to the greater impact of adding one link to each node, given the different statistical characteristics of Enron with respect to Facebook, especially the smaller average degree of the nodes in Facebook (ratio of degree in Facebook Vs degree in Enron is 6.1 to 1).

In Table 3 (Facebook) we see the results of the tests of perturbation versus information loss on the Facebook graph dataset, which are also plotted in Fig. 1b. Again, a clear increasing and linear trend for information loss (2%, 6%, 14%, 21%, 27%) is evident in relation with increasing perturbation values (10%, 25%, 50%, 75% and 100%).

5.2 Risk of Disclosure

The risk of disclosure is calculated by launching a query on the graph to find a given sub-graph topology (node and its immediate neighborhood) in the complete graph, with a % margin. A check is made to determine if the target node is in the subset S returned, and how many nodes are in S (value equivalent to that given by k anonymity). We perceive the attacker as statistically knowledgeable and whose objective is to identify specific nodes and their immediate neighbors, in a simply anonymized graph.

Consider that if we do not consider the 'apl' statistic, then there are many low risk users, whose 'cc' is equal to zero and/or whose 'degree' is equal to one. The 'apl' statistic is much more expensive and difficult to obtain, because it needs access to the whole graph dataset, thus we have considered the risk with and without the 'apl' statistic. Thus, we have three different measures for the risk of disclosure, defined by three queries:

- Q_1 which searches for a given node based on 'degree'
- Q_2 which searches for a given node based on 'degree' and 'cc'
- Q3 which searches for a given node based on 'degree', 'cc' and 'apl'.

For Q_1 we only consider users with degree > 1, and for Q_2 and Q_3 we only consider users with degree > 1 and cc > 0.0. All queries are allowed a 1% margin of error.

The *risk of disclosure* for a given node *Ng* in the original graph is calculated by multiplying the % of correct hits on the perturbed dataset for node *Ng*, by the % of nodes which are returned by the query within a given margin with respect to node *Ng*. We apply a margin of 1% in all cases. That is, if the degree of node *Np* in the perturbed dataset is within 1% of the degree of node *Ng* in the original dataset, then it is returned by the query. The same margin of 1% applies to the 'cc' and 'apl' values. Finally, a 'hit' is considered when the unique id of a node *Np* returned by the query has the same unique id as the node *Ng* in the original graph.

Facebook. With reference to Table 4, we see the results of the three query types and grades of perturbation, on the risk of disclosure. In Table 4 we see that for the degree query, the risk of disclosure reduces from 90.11% risk for 10% perturbation, to 0.009% risk for 100% perturbation, a significant reduction, for a simple query based only on degree. For progressively more complex queries, we observe a faster reduction in risk. In the case of 'degree, cc' the risk reduces from 84% to 0.007%, for 10% to 100% perturbation. In the case of the 'degree, cc, apl' query the reduction of risk occurs earlier: from 71% to 0.00026% for 10% to 50% perturbation. This is because the 'apl' (average path length) statistic of a node is very sensitive to change when one link is added to the node. The 'apl' value is also much more statistically diverse than the 'degree' and 'cc' values.

In Fig. 2a (Facebook) we see a sharp drop for the risk of the 'degree,cc, apl' query, for increasing percentages of perturbation, whereas the other two queries, 'degree' and 'degree, cc' show a more gradual drop, from 90% and 85% risk respectively, for 10% perturbation, to 50% and 35% risk respectively, for 50% perturbation.

Enron. The results shown in Table 5 have the same format and calculation method as we have described previously for the Facebook data of Table 4. We see the results of the three query types and grades of perturbation, for the risk of disclosure.

In Table 5 (Enron) we see that for the degree query, the risk of disclosure reduces from 94.34% risk for 10% perturbation, to 11.74% risk for 100% perturbation, with a similar decreasing tendency as for the Facebook data (Table 4), but leaving a greater residual risk. For the query 'degree, cc' the risk reduces from 83% to 1.6%, for 10% to 100% perturbation, again with a similar decreasing tendency as for the Facebook data, but leaving a greater residual risk. However, the risk reduction of the query 'degree, cc, apl' behaves in a different way to the Facebook query. The reduction of risk is much less pronounced: from 81% to 28% for 10% to 50% perturbation. This is due to two factors: (i) the sensitivity of the 'apl' value and (ii) the difference in the 'apl' values for Facebook and Enron (see Table 1): the average 'apl' for the Enron dataset is much smaller than that of Facebook (3.1516 and 6.001, respectively), and the other statistics related to 'apl' are also different if we compare the datasets.

In Fig. 2b (Enron) we see a sharper drop for the risk of the 'degree, cc' and 'degree,cc, apl' queries (relative to the 'degree' query), for increasing percentages of perturbation, from 82% and 81% risk for 10% perturbation, to 1.6% and 0.7% risk, respectively, for 100% perturbation. Both queries follow a very similar line. On the other hand, the 'degree' query shows a more gradual drop for increasing perturbation. These tendencies are similar to the results for the Facebook dataset, as seen in Figure 2a, with the exception of the query including 'apl', which shows a much more gradual descent, as we have already discussed with reference to Table 4.

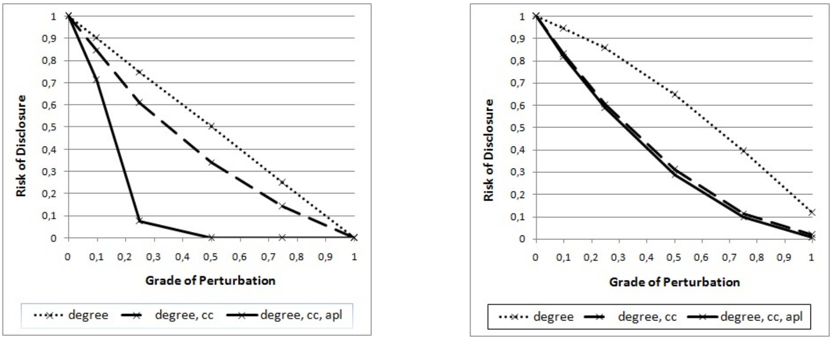
Information Loss vs Risk of Disclosure. With reference to Fig. 3, we see a plot of the results of Sections 5.1 and 5.2, for Information Loss and Risk of Disclosure for the Facebook (Fig. 3a) and Enron (Fig. 3b) datasets, respectively. Note that for information loss we have just one value for each degree of perturbation (see Sec. 5.1),

Table 4. Results of tests of perturbation *versus* risk of disclosure on the Facebook graph dataset

		Perturbation				
		10%	25%	50%	75%	100%
Risk of Dis-closure	degree	0.9011	0.7495	0.5014	0.2495	0.00009
	degree, cc	0.84834	0.6076	0.3415	0.1407	0.00007
	degree, cc, apl	0.71068	0.07333	0.00026	0.0000	0.00000

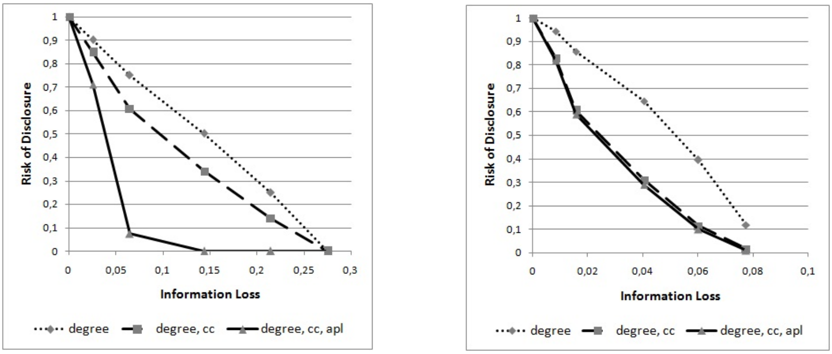
Table 5. Results of tests of perturbation *versus* risk of disclosure on the Enron graph dataset

		Perturbation				
		10%	25%	50%	75%	100%
Risk of Dis-closure	degree	0.9434	0.8570	0.6454	0.3943	0.11747
	degree, cc	0.82960	0.60519	0.31081	0.11363	0.01613
	degree, cc, apl	0.81819	0.58871	0.28832	0.09804	0.00798



(a) (b)

Fig. 2. Risk of disclosure *versus* Grade of Perturbation: (a) Facebook, (b) Enron



(a) (b)

Fig. 3. Information Loss *versus* Risk of Disclosure: (a) Facebook, (b) Enron

whereas for risk of disclosure we have three values (one for each query, see Sec. 5.2). We observe that the information loss of the Facebook dataset (Fig. 3a) ranges from 2.5% to 27.4% for a risk of disclosure which drops from 71% to 0.0%, in the case of the query 'degree, cc, apl'. On the other hand, the Enron dataset (Fig. 3b) has an information loss which rises from 0.7% to 7.7%, for a risk of disclosure which drops from 81.8% to 0.7%, in the case of the query 'degree, cc, apl'. Thus, we see that

Facebook has a greater reduction in risk of disclosure than Enron, especially for Q3 (degree, cc, apl). However, Facebook achieves this at a cost of four times the information loss, with respect to Enron. In summary, we can say that the information loss is relatively low for Enron (max. of 7.7%), whereas the Facebook result, with a maximum information loss of 27.4%, leaves room for improvement.

6 Conclusions

In this paper we have represented the Facebook and Enron user data and activity as a graph, which has allowed us to derive descriptive factors based on graph theory. We have introduced different percentages of perturbation into the data, by randomly adding links to the nodes. Then we have analyzed the information loss and risk of disclosure of the graphs from a data privacy point of view.

Lessons learnt: *firstly*, the perturbation method should be calibrated for each dataset. In our case, the perturbation method was 'add one link to node', and we could calibrate by varying the number of links added, based on the average degree value, for example; *second*, the risk of disclosure has to take into account the number of hits achieved within the subset of nodes returned by a query, rather than just the number of nodes returned (we note that this is distinct from k-anonymity); *thirdly*, it is important to filter the data, due to the presence of many nodes with just one link or with cc=0.0, in the graph. We filter these nodes because they are not interesting for a potential attacker because of their lack of interrelations (poor topology) and because they cannot be distinguished without the 'apl' (average path length) value, which is much more expensive and difficult to obtain. Also many values of 'degree' equal to one and 'cc' equal to zero would distort the graph statistics.

Future work: It would be interesting to try different perturbation methods on the graph, such as 'node aggregation' and compare this with 'add link'. For 'node aggregation' we could then consider 'k-anonymity' as a risk disclosure measure. Also it would useful to contrast the results for more online social network datasets, such as Twitter and a synthetic small-world graph.

Acknowledgements. This research is partially supported by the Spanish MEC (projects ARES CONSOLIDER INGENIO 2010 CSD2007-00004 -- eAEGIS TSI2007-65406-C03-02 -- and HIPERGRAPH TIN2009-14560-C03-01).

References

1. Shetty, J., Adibi, J.: Discovering Important Nodes through Graph Entropy - The Case of Enron Email Database. In: KDD 2005, Chicago, Illinois (2005)
2. Viswanath, B., Mislove, A., Cha, M., Gummadi, K.P.: On the Evolution of User Interaction in Facebook. In: Proc. 2nd ACM Workshop on Online Social Networks (WOSN), Barcelona, Spain, August 17 (2009), http://socialnetworks.mpi-sws.org/
3. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems (IJUFKS) 10(5), 557–570 (2002)

4. Domingo-Ferrer, J., Rebollo-Monedero: Measuring Risk and Utility of Anonymized Data Using Information Theory. In: Int. Workshop on Privacy and Anonymity in the Information Society, PAIS 2009 (2009)
5. Kumar, R., Novak, J., Tomkins: Structure and evolution of online social networks. In: Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. ACM, New York (2007)
6. Ahn, Y., Han, S., Kwak, H., Moon, S., Jeong: Analysis of topological characteristics of huge online social networking services. In: Proc. 16th Int. Conf. WWW 2007, USA (2007)
7. Kleinberg, J.: Challenges in Mining Social Network Data. In: Proc. of the 13th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining (KDD 2007), pp. 4–5 (2007)
8. Kleinberg, J., Backstrom, L., Dwork, C., Liben-Nowell, D.: Algorithmic Perspectives on Large-Scale Social Network Data. In: Data-Intensive Computing Symposium, (March 26, 2008 - Hosted by Yahoo! and the CCC), http://research.yahoo.com/files/7KleinbergSocialNetwork.pdf
9. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, California, USA (2007)
10. Hay, M., Miklau, G., Jensen, D., Weis, P., Srivastava, S.: Anonymizing Social Networks. SCIENCE Technical Report 07-19, vol. 245, pp. 107–3, Computer Science Department, University of Massachusetts Amherst (2007)
11. Zhou, B., Pei, J.: Preserving Privacy in Social Networks against Neighborhood Attacks. In: IEEE 24th International Conference on Data Engineering (ICDE), pp. 506–515 (2008)
12. Wondracek, G., Holz, T., Kirda, E., Kruegel, C.: A Practical Attack to De-Anonymize Social Network Users. In: Proc. IEEE Symp. on Security and Privacy, pp. 223–238 (2010)
13. Liu, K., Terzi, E.: Towards Identity Anonymization on Graphs. In: SIGMOD 2008 (2008)
14. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numer. Math. 1, 269–271 (1959)