Information Loss Evaluation based on Fuzzy and Crisp Clustering of Graph Statistics

David F. Nettleton
Data Privacy Research Group[1]
IIIA-CSIC[1] and Pompeu Fabra University[2]
Bellaterra, Spain
dnettleton@iiia.csic.es

***Abstract*— In this paper we apply different types of clustering, fuzzy (fuzzy c-Means) and crisp (k-Means) to graph statistical data in order to evaluate information loss due to perturbation as part of the anonymization process for a data privacy application. We make special emphasis on two major node types: hubs, which are nodes with a high relative degree value, and bridges, which act as connecting nodes between different regions in the graph. By clustering the graph's statistical data before and after perturbation, we can measure the change in characteristics and therefore the information loss. We partition the nodes into three groups: hubs/global bridges, local bridges, and all other nodes. We suspect that the partitions of these nodes are best represented in the fuzzy form, especially in the case of nodes in frontier regions of the graphs which may have an ambiguous assignment.**

Keywords-clustering; graphs; data privacy; perturbation; hubs; bridges; fuzzy; crisp

I. INTRODUCTION

Data Privacy has become a key issue in recent years due on the one hand to the ever greater need for disseminating data to the general public and specialized investigators, and on the other hand to the requirement of protecting the privacy of individuals and restricting confidential information. In order to publish data and information, we need to be able to guarantee that it does not infringe current data protection laws. Typically the original data is perturbed to produced an anonymized version while still maintaining its informational utility. In the case of graph data, the perturbation is potentially more damaging to the original data, and special care has to be taken not to destroy the topological characteristics, connectivity, and so on, of the graph.

In this paper we take a 'classic' graph dataset, 'Karate club', and we perturb it with two different node aggregation methods, using local node pair selection and global node pair selection, respectively. Then we evaluate the results in terms of graph structural change and information loss. We apply two different clustering techniques, one crisp and one fuzzy, in order to evaluate the information loss. As input data we have prepared statistical data which represent each node and the key topological characteristics of the graph. In order to do this, we identify two key node types in the graph, hubs and bridges, where a hub is a node with many connections to other nodes and a bridge is a connector between different regions of the graph. This information is given to the clustering algorithms, which group the nodes accordingly. If we apply a perturbation to the graph we assume that the statistical and structural information will change. This is especially critical for the hub and the bridging nodes. We will use the HITS algorithm to quantify the hub value for each node, and for the bridge nodes we will use two different metrics which measure their "local" bridging and "global" bridging coefficient, respectively. We distinguish between two types of bridging nodes because the 'global' bridges are often confused with the hubs, whereas the local bridges are not. We recalculate the statistics for the perturbed graph and perform the clustering on this data, which allows us to compare it with the clustering before perturbation. Using different measures for the cluster results, we can quantify the change and hence the information loss. The typical cluster measures we can use are the centroid values, then we inspect the assignment of the nodes to the clusters, and we evaluate the 'fuzziness' of the clusters before and after perturbation.

The structure of the paper is as follows: in Section II we briefly present related work, in Section III we define the metrics and techniques that are used in the paper: hub and bridge node statistics, clustering methods and perturbation methods (local and global), in Section IV we present the similarity measure used for node pair selection, in Section V we present the dataset used, in Section VI we present the results of the analysis of perturbation on graph structure, using first visualization and then clustering, followed by a presentation of information loss as quantified values, and finally in Section VII we conclude the paper.

II. RELATED WORK

We present 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, Sweeney's paper on k-anonymity [1], and more recently [2], give key definitions for information loss and risk of disclosure.

Firstly, in the field of the statistical analysis of online social networks, some key authors are: Kumar [3], Ahn [4], Klienberg [5][6], Mislove [7], Shetty [8] and Viswanath [9]. In [8], Shetty presented 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

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

removing the given node. In [7], Mislove defined some of the key metrics which characterize a social network. Viswanath in [9] performed a statistical analysis of the New Orleans Facebook dataset, using the degree, clustering coefficient and average path length statistics to evaluate social network evolution over time. Klienberg, in [5][6], considered the data mining of online social networks, defining different possible topologies within OSNs and making considerations about the computational cost of data processing.

Secondly, in the field of data privacy analysis applied to online social networks, two key authors relevant to the present work are Hay [10][11] and Zhou [12]. Also, in [13] a study was presented of the anonymization of the Facebook and Enron datasets represented as graphs, using 'edge addition' as the perturbation method.

Hay [11] presented a simple graph anonymization based on random addition and deletion of edges. 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. Nodes are chosen for nodes for aggregation based on isomorphisms. Although an efficient data representation and search method (DFS on tree representation) is used, it still requires checking each graph element with every other.

In [10], Hay presented a different approach in which nodes are grouped (aggregated) into partitions based on isomorphic properties. Entropy is calculated for the entire graph, which incurs in the problem of high computational cost. In contrast to Hay, our method pre-calculates a unique factor for each node topology which is then inserted into a hash-table, and then passed to a vector which is sorted on the values of the factor. The closest node can then be found in the vector by a binary search on the factor value.

Zhou, in [12], also presented a method which grouped nodes into partitions, based on isomorphic properties. Zhou [12] presented 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.

In [14], Girvan and Newman present an analysis of 'community structure' in OSNs, defining different key metrics which characterize these type of structures. This is a theme which we consider in Section VI of this paper, with respect to the effects of perturbation.

Finally, two references of relevant recent work in fuzzy and possibilistic clustering are [15] and [16]. In [15], fuzzy k-means and hard k-means clustering are applied to data which represents 3D images, in order to detect lines, outliers and perform segmentation. The lowest least square mean errors were given by the fuzzy methods. One feature of the method is that of finding the optimal number of clusters by evaluating inter and intra cluster distances. In our present work we fixed the number of clusters at three by conducting different empirical tests and analyzing the resulting cluster assignments and least mean square error. In [16], possibilistic aggregation functions are applied to the Epinions online social network in order to evaluate "trust" for access control methods. In contrast to the Karate graph, which is the result of a sociological study of the real relations between people in a group, the authors state that the Epinions graph manifests the general difficulty of evaluating relations in online social networks, because many of the network participants do not really know each other personally.

III. METRICS AND TECHNIQUES USED

In this Section we define the metrics and techniques that are used in the remainder of the paper: hub and bridge nodes, clustering methods and perturbation methods.

A. Hub and Bridge Nodes

1) Hub nodes: Intuitively, 'hub' nodes can be considered as those which have a high number of links with other nodes. However, in order to quantify the relative importance of the links that a node has, and not just their number, hub metrics take into account the characteristics of the neighbors. Two algorithms which quantify this are HITS[18] and PageRank[19]. Klienberg's HITS algorithm[18], can be considered a precursor to the more well know Page Rank[19]. In contrast to Page Rank, it calculates two values for each node, a hub value and an authority value which are defined in terms of one another. As we are only interested in calculating a hub metric for the current study, HITS is more adequate than Page Rank, as the latter combines the hub and authority values. The authority value of a node *x* is the sum of the normalized hub values of the nodes that point to *x*. The hub value of a node *x* is the sum of the normalized authority values of the nodes that *x* points to. The process is as follows: Initially, $\forall$p , auth(p) = 1 and hub(p) = 1. There are two types of updates: Authority Update Rule and Hub Update Rule. The hub/authority scores of each node are calculated by repeated iterations of the Authority Update Rule and the Hub Update Rule. This calculation gives an approximation of the scores whose precision depends on the number of iterations and the convergence/cut-off.

a) Authority update rule: $\forall$p, auth(p) is updated by:

$$\sum_{i=1}^{n} \mathrm{hub}(i) \tag{1}$$

where *n* is the total number of pages connected to *p* and *i* is a page connected to *p*. Hence, the Authority score of a page is the sum of all the Hub scores of pages that point to it.

b) Authority update rule: $\forall$p, hub(p) is updated by:

$$\sum_{i=1}^{n} \mathrm{auth}(i) \tag{2}$$

where *n* is the total number of nodes which *p* is connected to and *i* is a node which *p* is connected to. Hence, a node's Hub score is the sum of the Authority scores of all its linking pages.

2) *Bridge nodes:* These are nodes which may not necessary have a high degree but which are "strategically" placed between other nodes such that they form a key part of the graph's connectivity. That is, their removal would cause a major disruption to the graph. Bridge nodes can be quantified by different metrics. One of the most commonly used is 'betweenness centrality', which is calculated in terms of the number of critical paths which go through a given node, from/to other nodes. However, 'hub' (high degree) nodes also tend to have a relatively high 'betweenness centrality', and often obfuscate the presence of low degree bridging nodes. Hence, in this paper we also consider a second metric, called 'bridging centrality', published by Hwang et al. in [17], which is effective in distinguishing bridge nodes, and differentiating them from 'hub' nodes. Hwang defines the 'bridging centrality' of a node as the product of the 'betweenness centrality' C_B and the bridging coefficient (BC). In this paper we refer to 'betweenness centrality' as identifying 'global' bridges, whereas 'bridging centrality' identifies 'local' bridges.

The bridging coefficient of a node measures how well the node is located between hub (high degree) nodes. For a given node v:

$$BC(v) = \frac{d(v)^{-1}}{\sum_{i \in N(v)} \frac{1}{d(i)}} \tag{3}$$

where d(v) is the degree of node v, and N(v) is the set of neighbors of node v. This value embodies local characteristics of node v.

On the other hand, the betweenness centrality of node v is defined as:

$$C_B(v) = \sum_{\substack{x \neq v \neq t \\ x,v,t \in V}} \frac{p_{st}(v)}{p_{st}} \tag{4}$$

where P_{st} is the number of shortest paths from node s to t, $P_{st}(v)$ is the number of shortest paths from s to t that pass through the node v. This value embodies global characteristics of node v. Finally, the bridging centrality of node v is defined as:

$$C_R(v) = BC(v) \times C_B(v) \tag{5}$$

which combines the 'local' and 'global' bridging characteristics of v.

B. Clustering Methods

We use a 'hard' clustering technique, k-Means [20], and a 'soft' clustering technique, Fuzzy c-Means [21], which allows us to contrast the results. The standard version of k-Means is used from the Weka software program [22]. In the case of Fuzzy c-Means, we have used our own implementation written in the 'C' programming language.

C. Perturbation Methods

Hay[10][11] and Zhou[12] consider generalizing nodes into partitions in order to anonymize them. However, the effect of the distance between the node pairs is not explicitly considered. Hence, in the present work, we will evaluate this aspect. We consider two perturbation methods, local pair selection and global pair selection. In local pair selection we first randomly choose a node N from the whole graph, and then we choose one neighbor N' of N (from its adjacency list) using a similarity criteria. That is, we choose the neighbor N' of N which is most similar to N. Contrastingly, in global pair selection, we first randomly choose a node N from the whole graph, and then we choose a second node N' again from the whole graph, using a similarity criteria to choose the node N' most similar to node N. In both local and global selection we then proceed to aggregate the pair {N, N'} in just one node, that is, the links of N' are deleted and its neighbors are linked to node N, and then node N' itself is deleted. Finally, the graph statistics are recalculated where necessary.

IV. SIMILARITY MEASURE FOR NODE SELECTION AND PAIRING

We use a similarity based selection approach, which has a significantly lower computational cost than isomorphism based selection methods such as [10][12]. However, the similarity approach can be considered a good approximation, whereas the isomorphic similarity is more precise. We calculate descriptive statistical factors which are sufficiently representative of a nodes topology to permit accurate identification of similar pairs in the whole graph. We now present two metrics which will calculate a characteristic value for the nodes in a graph and which will enable us to calculate the 'distance' between two given nodes and hence their 'similarity'. The following formula and weighting factors have been calibrated through diverse empirical tests:

$$\begin{aligned} \textbf{\textit{Sim(}}\textit{node}\textbf{\textit{)}} = \\ \text{LOG}\ \big(\ \big(\ \text{degree}^2 \times \gamma^3\ \big) + \\ \big(\ \text{degree} \times \text{normalized average degree of adjacent nodes} \times \gamma^2\ \big) + \\ \big(\ \text{degree} \times \text{normalized standard deviation of degree of adjacent nodes} \times \gamma\ \big) \\ \big) \end{aligned} \tag{6}$$

where γ is equal to the maximum degree of any node found in the graph. So for any relevant graph, $\boldsymbol{\gamma} > 1$, or $\boldsymbol{\gamma} >> 1$. We take the log to manage the magnitude of the resulting value. We observe that *Sim* includes information about the immediate neighbors, the average and standard deviation of their degree values.

V. TEST DATASET

In this Section we present the 'Karate' test dataset[23] and visualize the graph before perturbation with the Gephi 0.8 alpha software (http://gephi.org/) and using the 'Force Atlas' layout technique [24] [25]. The 'Force Atlas' technique is particularly efficient at showing the 'community structure' of a graph, that is, hub nodes for each community and boundaries and linking nodes between the communities.

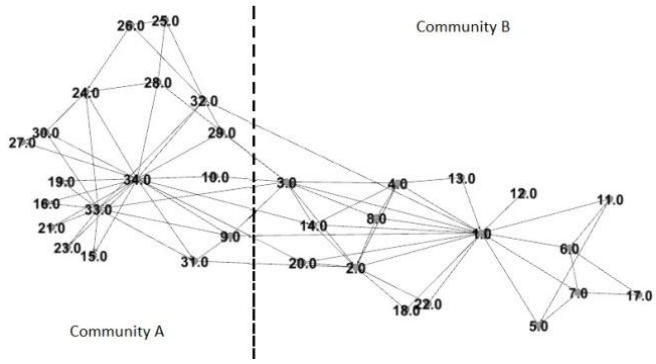

Figure 1. Graph representation of the 'Karate' dataset (generated using Gephi 0.8 alpha and Force Atlas layout option[24][25])

In this way, we can see if the perturbation technique, although having maintained the global statistical values of the graph, has 'destroyed' some or all of the community structure. We use the Karate graph dataset because it enables us to clearly see the effect and results of perturbation and clustering, given that the role of each node and the interrelations are exactly known apriori [23].

We now present some simple examples of graph statistical summaries of the 'Karate' dataset, and how we could interpret this information, especially for the identification of community structure. The 'Karate' dataset consists of 34 nodes, which represent real persons who were members of a Karate club. The members were polarized around two key figures, the owner of the club, represented by node 34, and the club trainer/teacher, represented by node 1. Over time, the club experienced a rupture, in which the followers of the trainer left to form a new club. The two communities, as can be seen from Fig. 1 (identified as Community A and Community B), are made up of the following nodes: Community A {34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 21, 19, 16, 15, 10, 9} and Community B {1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22}. From Zachary's original description, as a point of curiosity, node 9 was originally a supporter of the owner (node 34), but later defected to support the trainer (node 1).

In Table I we see the summarized statistics for the Karate dataset graph, where the top ranked nodes are indicated by the grey indicated cells, corresponding to the ranking statistic. We are particularly interested in the last three columns (hub, betweenness centrality and bridging centrality) because we will use these values later (Section VI) for the clustering. For the hub value (fourth column), we see that nodes 34, 1, 33, 3 and 2 are the top five ranked, in that order. In general, for 'hub' nodes, we would expect them to have high values for degree and 'betweenness centrality'. The clustering coefficient does not have such a clear correlation, although it tends to be lower for higher degree nodes because of the lower possibility that a high number of neighbor nodes have mutual links. 'Betweenness centrality' is a measure of the relative positioning of a vertex within the complete graph. respectively.

TABLE I. STATISTICS OF THE 'KARATE' DATASET

Node id	Degree	Clust. Coef.	Hub ranking	Between-ness Centrality	Bridging Centrality
1	16	0.15	0.089	0.4376	0.0053
2	9	0.33	0.053	0.0539	0.0025
3	10	0.24	0.058	0.1437	0.0067
4	6	0.67	0.037	0.0119	0.0016
5	3	0.67	0.021	0.0006	0.0003
6	4	0.50	0.026	0.0300	0.0065
7	4	0.50	0.026	0.0300	0.0065
8	4	1.00	0.026	0.0000	0.0000
9	5	0.50	0.032	0.0559	0.0202
10	2	0.00	0.016	0.0008	0.0027
11	3	0.67	0.021	0.0006	0.0003
12	1	0.00	0.011	0.0000	0.0000
13	2	1.00	0.016	0.0000	0.0000
14	5	0.60	0.032	0.0459	0.0184
15	2	1.00	0.016	0.0000	0.0000
16	2	1.00	0.016	0.0000	0.0000
17	2	1.00	0.016	0.0000	0.0000
18	2	1.00	0.016	0.0000	0.0000
19	2	1.00	0.016	0.0000	0.0000
20	3	0.33	0.021	0.0325	0.0466
21	2	1.00	0.016	0.0000	0.0000
22	2	1.00	0.016	0.0000	0.0000
23	2	1.00	0.016	0.0000	0.0000
24	5	0.40	0.032	0.0176	0.0036
25	3	0.33	0.021	0.0022	0.0010
26	3	0.33	0.021	0.0038	0.0018
27	2	1.00	0.016	0.0000	0.0000
28	4	0.17	0.026	0.0223	0.0081
29	3	0.33	0.021	0.0018	0.0018
30	4	0.67	0.026	0.0029	0.0009
31	4	0.50	0.026	0.0144	0.0079
32	6	0.20	0.037	0.1383	0.0191
33	12	0.20	0.068	0.1452	0.0032
34	17	0.11	0.095	0.3041	0.0031

Vertices that form part of many shortest paths between other vertices have a higher betweenness value than those that do not. If we observe the values for nodes 1 and 34, they have the following values for degree and 'betweenness centrality': {16, 0.4376} and {17, 0.3041}, respectively. We observe from Table I that these values represent the maximums for all the nodes in the graph, confirming our hypothesis that they are 'hubs'. However, their 'bridging centrality' is relatively low, being 0.0053 and 0.0031, respectively. That is because 'bridging centrality' is designed to identify nodes which link hub nodes, but which are not necessarily hub nodes themselves (see Section III for the definitions).

Returning to Fig. 1, we can see that the 'Force Atlas' layout option has successfully distinguished the two communities, which we will call 'Community A' and 'Community B', with respective core vertex sets of {33, 34} and {1, 2, 3}. Suppose that we are particularly interested in the boundary between the two communities, and connections which form "bridges" between them. A "bridge" can be usefully considered as a node or an edge which connects "modular" regions in a graph. By simple observation, we can identify which nodes in community "A" have links with nodes in group "B", and vice versa. Such nodes in community "A" are: {10, 31, 9, 32, 29, 33 and 34 (2 links)}. All have just one link unless otherwise stated. Two nodes of special interest are node 34, which also

happens to be the hub node of community "A", and node "9", which has direct links from nodes 33 and 34 of its own community, and a direct link to node 1 of the other community (node 1 being the hub node of community B). If we apply the same process to community B, this gives the following cross community boundary linking nodes: {3 (4 links), 14, 20, 1 (2 links)}. Again, all have one link unless otherwise stated. Two nodes of special interest are 1, the hub node of community B, which has links to nodes 9 and 32 in community A; and node 3, which has a high relative number of cross community links, with a total of 4 links. This number of links is only one less than the number of links which node 3 has to its own community, and background information [14][23] is needed to place it in community B.

However, a quicker way to identify the top boundary linking nodes is to use the 'bridging centrality' statistic, which is shown in the rightmost column of Table I. The cells shaded in grey are the three top ranked nodes in terms of 'bridging centrality', identified as nodes 20, 9 and 32, with values of 0.0466, 0.0202 and 0.0191, respectively. These are key nodes for the graph's structures (see Fig. 1), but they are not as evident as the hub nodes. In the cluster analysis of Section VI, we will see that both hard and fuzzy clustering techniques form a partition to distinguish them from the hub nodes and the rest of the nodes which are neither hubs nor bridges.

VI. ANALYSIS OF THE EFFECT OF PERTURBATION ON GRAPH STRUCTURE

In this Section we present the results of the analysis of perturbation on graph structure, using first visualization and then clustering. We then present the information loss as quantified values.

A. Analysis by Visualization

In this subsection we perturb the Karate Graph dataset by selecting node pairs based on similarity and aggregating them. Then we visualize the results using the Gephi software, and summarize the graph statistics before and after perturbation for each method. We test two aggregation methods, the first aggregated similar nodes which must be immediate neighbors (local pairing) and the second aggregated similar nodes which can be anywhere in the graph (global pairing). Thus we can contrast the methods and their respective information loss.

Firstly we will enumerate the node aggregations performed by each pairing method, local and global. This information is necessary in order to interpret the results shown in Figs. 2a and 2b, and the following analysis. We recall the first node N_1 of each pair is chosen randomly, and the second node N_2 is chosen as the most similar node to N_1. In the case of local aggregation, N_2 will be the most similar of N_1's immediate neighbors. In the case of global aggregation, N_2 will be the node most similar to N_1, anywhere in the graph. N_1 and N_2 are fused into a new node which is given the id of N_1, and the neighbors of N_2 are reconnected to N_1. Once a node has been aggregated, it is excluded from being selected again for pairing. Thus, some nodes may not be paired during the processing of the complete graph, for example if their only neighbor is a node which has already been paired (local), or if there are no unpaired nodes remaining (global and local).

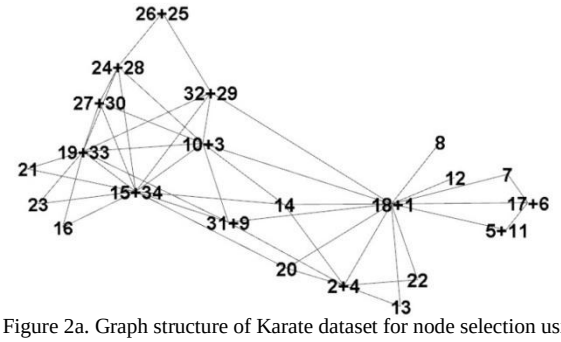

Figure 2a. Graph structure of Karate dataset for node selection using similarity metric with *local* pairing

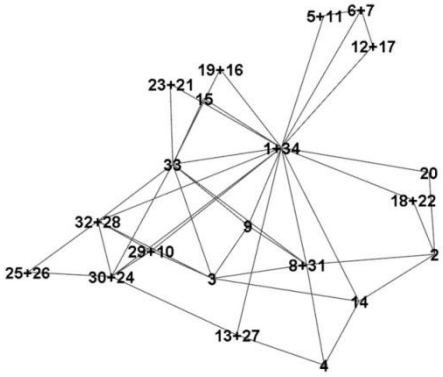

Figure 2b. Graph structure of Karate dataset for node selection using similarity metric with *global* pairing

Local pairing: the following pairing were made: 27+30, 17+6, 18+1, 10+3, 2+4, 5+11, 19+33, 24+28, 32+29, 31+9, 15+34, 26+25.

Global pairing: the following pairing were made: 8+31, 25+26, 6+7, 23+21, 13+27, 19+16, 12+17, 18+22, 29+10, 1+34, 32+28, 5+11, 30+24.

We note that in the case of the global method, the 'pathological case' of pairing nodes 1 and 34 has occurred, which causes a major disruption to the graph because nodes 1 and 34 are the respective hubs of the two major partitions of the Karate dataset. This is possible because the global method has chosen the two most similar nodes at the given point in the processing, and no restrictions are placed on including nodes based on their characteristics or relative position in the graph. We also note that the selection of the first node in each pair, N_1, is random, therefore in execution of the perturbation process, the node pairings will be different.

With reference to Figs. 2a and 2b, pairs are indicated thus, '26+25', which means that a new node has been formed by pairing nodes 25 and 26 (see Fig.1 for original nodes). If we compare Fig. 1 (original graph) with Figs. 2 (perturbed graphs), we see the effect of anonymization (node aggregation) on the original graph structure. We especially note what has occurred to the two communities, and to the corresponding hub nodes '1' and '34', of those communities. We see that the global pairing method has caused the greater distortion of the graph (Fig. 2b). On the other hand, local pairing (Fig. 2a) has maintained the two community structures to a greater extent.

TABLE II. GRAPH STATISTICS FOR DIFFERENT PERTURBATION METHODS- KARATE DATASET

	HITS (hub)	Betweenness Centrality	Bridging Centrality
Original dataset	0.027 ± 0.017*	0.036 ± 0.083	0.005 ± 0.009
Local node pairing	0.045 ± 0.029	0.057 ± 0.124	0.005 ± 0.009
Global node pairing	0.045 ± 0.029	0.049 ± 0.133	0.006 ± 0.007

*Values represent the average value and standard deviation, respectively.

In Fig. 2b we see that the connectivity of hub nodes 1 and 34 has been combined, thus significantly distorting the overall graph structure. However, in both Figs. 2a and 2b, Community A is still approximately represented by nodes towards the left, and Community B by the nodes to the right.

In Table II we see the summarized statistics for the perturbed graphs of Figs. 2a and 2b, together with the statistics of the original dataset. For bridging centrality, the global method gives the greatest difference when compared with the average value for the original dataset (0.006 with respect to 0.005). For betweenness centrality, on the other hand, the local method gives the greatest difference (0.057 with respect to 0.036). This implies that the bridging coefficient of the nodes is less affected by global aggregation. Finally, for the hub value, both methods (local and global) give the same result of 0.045.

B. Analysis by Clustering

In this subsection we present the results of applying two contrasted clustering techniques to the graph data, represented by the three statistics 'hub', 'betweenness centrality' and 'bridging centrality'. Then, we derive an information loss measure by calculating the difference between the cluster centers, for the data before and after perturbation. The use of a 'hard' technique, k-Means, and a 'fuzzy' technique, Fuzzy c-Means, allows us to contrast the results.

After different empirical tests we fixed the number of clusters to three, for both methods, which gave the lowest 'least square error' and the most coherent node assignments for the given input variables, which were the 'hub', 'betweenness centrality' and 'bridging centrality' metrics. For processing with Fuzzy c-Means, we used the Euclidean norm, and a weight exponent of 2.0.

1) k-Means: In Table III we summarize the clustering results in terms of the cluster centroids for each of the input attributes. Then in Table IV we show the results for the node assignments for each cluster and dataset. With reference to Table III, the grey cells indicate the maximum values for each data attribute's centroid value. In this way we distinguish that *Cluster A* contains the nodes with high values for hub and between centrality, *Cluster B* contains the nodes with high values for bridging centrality, and *Cluster C* contains the remaining nodes.

TABLE III. K-MEANS CLUSTERING RESULTS (CLUSTER CENTROIDS) FOR ORIGINAL DATASET, DATASET PERTURBED USING LOCAL PAIRING, AND DATASET PERTURBED USING GLOBAL PAIRING

ORIGINAL DATASET

	Complete Dataset (34)	Cluster A (2, 6%)	Cluster B (7, 21%)	Cluster C (25, 74%)
HITS (hub)	0.0294	0.0921	0.0429	0.0206
Betw. Cen.	0.0440	0.3709	0.0879	0.0056
Bri. Cen.	0.0049	0.0042	0.0167	0.0017

DATASET PERTURBED USING LOCAL PAIRING

	Complete Dataset (22)	Cluster A (4, 18%)	Cluster B (4, 18%)	Cluster C (14, 64%)
HITS (hub)	0.0455	0.0970	0.0453	0.0308
Betw. Cen.	0.0597	0.2508	0.0616	0.0046
Bri. Cen.	0.0063	0.0078	0.0245	0.0007

DATASET PERTURBED USING GLOBAL PAIRING

	Complete Dataset (34)	Cluster A (2)	Cluster B (7)	Cluster C (25)
HITS (hub)	0.0476	0.0923	0.0396	0.0360
Betw. Cen.	0.0514	0.2043	0.0365	0.0066
Bri. Cen.	0.0061	0.0064	0.0167	0.0016

TABLE IV. K-MEANS CLUSTERING RESULTS (NODE ASSIGNMENTS TO CLUSTERS) FOR ORIGINAL DATASET, DATASET PERTURBED USING LOCAL PAIRING, AND DATASET PERTURBED USING GLOBAL PAIRING

ORIGINAL DATASET

	Nodes assigned
Cluster A	34, 1
Cluster B	33, 32, 20, 14, 9, 3, 2
Cluster C	31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 19, 18, 17, 16, 15, 13, 12, 11, 10, 8, 7, 6, 5, 4

DATASET PERTURBED USING LOCAL PAIRING

	Nodes assigned
Cluster A	{19+33}, {18+1}, {15+34}, {10+3}
Cluster B	{32+29}, {31+9}, 20, 14
Cluster C	{27+30}, {26+25}, {24+28}, 23, 22, 21, {17+6}, 16, 13, 12, 8, 7, {5+11}, {2+4}

DATASET PERTURBED USING GLOBAL PAIRING

	Nodes assigned
Cluster A	33, {32+28}, {30+24}, {1+34}
Cluster B	20, {18+22}, 14, {13+27}, {8+31}
Cluster C	{29+10}, {25+26}, {23+21}, {19+16}, 15, {12+17}, 9, {6+7}, {5+11}, 4, 3, 2

With reference to Table IV, we have indicated in curly brackets the pairings, where the pair assumed the id of the first node. For example, {1+34} indicates that nodes 1 and 34 were paired, the new node assigned id=1. We observe that for the dataset perturbed using global pairing, node pair {1,34} was assigned to Cluster A (the hub nodes), which is consistent with the original dataset and the dataset perturbed using local pairing. We also confirm that Cluster B contains the nodes with highest bridging centrality.

TABLE V. FUZZY C-MEANS CLUSTERING RESULTS (CLUSTER CENTROIDS) FOR ORIGINAL DATASET, DATASET PERTURBED USING LOCAL PAIRING, AND DATASET PERTURBED USING GLOBAL PAIRING

ORIGINAL DATASET			
	Cluster A	Cluster B	Cluster C
HITS (hub)	0.0915	0.0539	0.0217
Betw. Cen.	0.3837	0.1380	0.0090
Bri. Cen.	0.0044	0.0099	0.0037
DATASET PERTURBED USING LOCAL PAIRING			
	Cluster A	Cluster B	Cluster C
HITS (hub)	0.1206	0.0766	0.0315
Betw. Cen.	0.5397	0.1344	0.0082
Bri. Cen.	0.0074	0.0132	0.0036
DATASET PERTURBED USING GLOBAL PAIRING			
	Cluster A	Cluster B	Cluster C
HITS (hub)	0.1532	0.0651	0.0331
Betw. Cen.	0.6070	0.0651	0.0069
Bri. Cen.	0.0069	0.0084	0.0047

Finally, it can be seen that Cluster C contains the remaining nodes, those which are not in Clusters A or B, that is, which have neither a relatively high hub, 'betweenness centrality' or 'bridging centrality' value.

2) Fuzzy c-Means: In Table V we summarize the clustering results in terms of the cluster centroids for each of the input attributes. Then in Table VI we show the results for the node assignments for each cluster and dataset.

With reference to Table V, the grey cells again indicate the maximum values for each data attribute's centroid value. In concordance with the results of k-Means shown in Table III, *Cluster A* again contains the nodes with high values for hub and between centrality, *Cluster B* contains the nodes with high values for bridging centrality, and *Cluster C* contains the remaining nodes.

With reference to Table VI, the results shown for the node assignments of Fuzzy c-Means have the same format as those of Table IV (k-Means)., with the exception of the rightmost column in which we show the nodes whose principal fuzzy membership grade is less than 0.9.

If we consider the node to cluster assignments of Fuzzy c-Means (Table VI) with respect to those of k-Means (Table IV), we observe that for the local and global methods, a significant number of nodes have passed from Cluster A to Cluster B. Also, for the same methods, some nodes have passed from Cluster B to Cluster C, such as node 20. However, it is out of the scope of the current paper to discuss in detail the differences between the hard and fuzzy assignments.

TABLE VI. FUZZY C-MEANS CLUSTERING RESULTS (NODE ASSIGNMENTS TO CLUSTERS) FOR ORIGINAL DATASET, DATASET PERTURBED USING LOCAL PAIRING, AND DATASET PERTURBED USING GLOBAL PAIRING

ORIGINAL DATASET		
	Node assignment (Cluster with membership grade > 0.5)	Nodes with significant fuzzy membership (principal membership grade < 0.9)
Cluster A	34, 1	34 (0.7778)
Cluster B	33, 32, 3	
Cluster C	31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4	20 (0.8371), 9 (0.7284), 14 (0.8344), 2 (0.6923)
DATASET PERTURBED USING LOCAL PAIRING		
Cluster A	{18+1}	
Cluster B	{32+29}, {31+9}, {19+33}, {10+3}	{32+29} (0.8925), {31+9} (0.5169)
Cluster C	{27+30}, {26+25}, {24+28}, 23, 22, 21, 20, {17+6}, 16, 15, 14, 13, 12, 8, 7, {5+11}, {2+4}	15 (0.8281), {2+4} (0.8539),
DATASET PERTURBED USING GLOBAL PAIRING		
Cluster A	{1+34}	
Cluster B	33, {32+28}, {30+24}, 14, {8+31}, 3	33(0.8750), 14 (0.7441), 3 (0.6041)
Cluster C	{29+10}, {25+26}, {23+21}, 20, {19+16}, {18+22}, 15, {13+27}, {12+17}, 9, {6+7}, {5+11}, 4, 2	{13+27} (0.8152), 2 (0.8686)

C. Quantification of Information Loss

In order to quantify information loss for the clustering methods, we simply calculate the Euclidean distance between the corresponding cluster centroids, for each data attribute (hub, betweenness centrality and bridging centrality), for the original data and for the perturbed data. The data attributes are given equal weighting. Then the sum of the differences will be a measure of the information loss, where an identical clustering result would give a zero value thus indicating no information loss. The higher the value the greater the information loss. As all the data values are normalized, we can compare the absolute differences for the different perturbation methods and clustering techniques.

TABLE VII. INFORMATION LOSS IN TERMS OF DIFFERENCES IN CENTROIDS

	k-Means	Fuzzy c-Means
Local node pairing	0.1773	0.2284
Global node pairing	0.2402	0.3876

TABLE VIII. INFORMATION LOSS IN TERMS OF FUZZINESS OF CLUSTERS (ONLY FUZZY C-MEANS)

	Fuzziness
Original dataset	1.4719
Local node pairing	1.3042
Global node pairing	1.5065

With reference to Table VII, we see that the dataset corresponding to global node pairing gives the highest difference between centroid values for both k-Means and Fuzzy c-Means, 0.2402 and 0.3876, respectively. In the case of the results of Fuzzy c-Means, we can also quantify the change in "fuzziness" of the perturbed dataset, by calculating S_f as the sum of (1 - N_if), where N_if indicates the principal membership grade for node *i*. If the principal membership grades are closer to 1, the clustering assignments will be less fuzzy and S_f will tend to zero. On the other hand, as the clustering assignments become increasingly more fuzzy, S_f will tend to $1 \times N = N$. In Table VIII we see the results corresponding to this calculation, in which it can be seen that the dataset produced by global node pairing is closest to the original dataset in terms of fuzziness, with a difference of +0.0346 for global with respect to -0.1677 for local. This result differs from the information loss result calculation based on centroid values in which local gave the lower value.

VII. SUMMARY AND CONCLUSIONS

The difference between the clustering metrics can be used as a measure for information loss. However, as information loss depends on the use of the perturbed data, it is clear that this information loss measure is most relevant when the user wishes to perform clustering on the perturbed data and wishes that the results be as close as possible to those of the original data. With respect to the local and global option for pair selection, it is clear from the results that local pair selection gives the lowest information loss and causes the least disruption to the overall graph structure. The similarity function has shown to be precise for selecting similar nodes. We have found that the 'bridging centrality' statistic is a better measure for identifying bridging nodes with respect to the 'betweenness centrality' measure. This is because the 'hub' nodes are also often highly correlated with the 'betweenness centrality' statistic, whereas 'bridging centrality' is not.

In the light of the results of the current work, we can consider improving the perturbation by taking into account ‘special’ nodes such as hubs and bridges. For example, the node aggregation process could exclude the nodes with the highest hub and bridge values, or perturb them following some special restrictions.

REFERENCES

[1] L. Sweeney, k-anonymity: a model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS). Vol. 10, Issue: 5, pp. 557-570, 2002.

[2] J. Domingo-Ferrer, J. and V. Torra, "Disclosure control methods and information loss for microdata, confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies. Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. eds., Elsevier, pp. 91-110.

[3] R. Kumar, J. Novak and J. Tomkins, "Structure and evolution of online social networks", Proceedings of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining", ACM, New York, USA, 2007.

[4] Y. Ahn, S. Han, H. Kwak, S. Moon, and S. Jeong, "Analysis of topological characteristics of huge online social networking services", Proc. 16th Int. Conf. on World Wide Web, New York, USA, 2007.

[5] J. Kleinberg, "Challenges in Mining Social Network Data", Proc. of the 13th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining (KDD '07), pp. 4 - 5, 2007.

[6] J. Kleinberg, L. Backstrom, C. Dwork and D. Liben-Nowell, "Algorithmic Perspectives on Large-Scale Social Network Data", Data-Intensive Computing Symposium (March 26, 2008 - Hosted by Yahoo! and the CCC). "research.yahoo.com/files/7KleinbergSocialNetwork.pdf"

[7] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks" Proc. 7th ACM SIGCOMM Conf. on Internet Measurement, San Diego, California, USA, 2007.

[8] J. Shetty and J. Adibi, "Discovering Important Nodes through Graph Entropy - The Case of Enron Email Database", KDD ’2005, Chicago, Illinois, 2005.

[9] B. Viswanath, A. Mislove, M. Cha and K.P. Gummadi, On the Evolution of User Interaction in Facebook. In Proc. 2nd ACM workshop on Online social networks (WOSN), (August 17,2009, Barcelona, Spain). "http://socialnetworks.mpi-sws.org/", 2009.

[10] M. Hay, G. Miklau, D. Jensen, D. Towsley and P. Weis, "Resisting structural re-identification in anonymized social networks", Proc. of the VLDB Endowment (SESSION: Privacy and authentication) Vol. 1 , Issue 1, pages 102-114, August 2008.

[11] M. Hay, G. Miklau, D. Jensen, P. Weis and S. Srivastava, "Anonymizing Social Networks", SCIENCE Technical Report 07-19 (2007) pp. 107--3, Vol. 245 (Computer Science Dept., Univ. Massachusetts Amherst)

[12] B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks", IEEE 24th International Conference on Data Engineering (ICDE), 2008, pp. 506 - 515.

[13] D.F. Nettleton, D. Sáez-Trumper and V. Torra. "A Comparison of Two Different Types of Online Social Network from a Data Privacy Perspective", MDAI 2011, Changsha, China, July, 2011. Proc. Springer LNAI, Vol. 6820/2011, p223-234.

[14] M. Girvan and M.E.J. Newman, "Community Structure in social and biological networks", Proc. National Academy of Sciences of the USA (PNAS), Vol. 99, No. 12, pp. 7821-7826, 2002.

[15] T.B. Nguyen, L. Sukhan, "Segmentation and outlier removal in 3D line identification based on fuzzy clustering," Proc. 2010 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2010), pp.1-8, 18-23 July 2010.

[16] J. Nin, V. Torra, "Possibilistic Reasoning for Trust-based Access Control Enforcement in Social Networks", Proc. 2010 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2010), pp.1-6, 18-23 July 2010.

[17] W. Hwang, T. Kim, M. Ramanathan and A. Zhang. "Bridging centrality: graph mining from element level to group level". In Proc. 14th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining (KDD ’08), pp. 336–344, New York, NY, USA.

[18] J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", Journal of the ACM (JACM), Volume 46 Issue 5, pp.604-632, Sept. 1999.

[19] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", In Computer Networks and ISDN Systems, Volume 30, Issues 1–7, April 1998, Pages 107–117.

[20] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm", Journal of the Royal Statistical Society. Series C (Applied Statistics) , Vol. 28, No. 1 (1979), pp. 100-108

[21] J.C.Bezdek, R. Ehrlich and W. Full, "FCM: The fuzzy c-means clustering algorithm", Computers & Geosciences, Volume 10, Issues 2–3, 1984, pp. 191–203.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1, 2009.

[23] W.W. Zachary, "An information flow model for conflict and fission in small groups", Journal of Anthrop. Research, 33, pp. 452-473, 1977.

[24] Y.F. Hu,. "Efficient and high quality force-directed graph drawing", Mathematica Journal 10, pp. 37–71, 2005.

[25] A. Noack. "Energy-based clustering of graphs with nonuniform degrees", Proc. 13th Int. Symposium on Graph Drawing (GD 2005), pp. 309–320, Springer Verlag, 2005.