Expert Systems With Applications 55 (2016) 87–105

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A data driven anonymization system for information rich online social network graphs

David F. Nettleton[a,1,*], Julián Salas[b,1]

[a] Department of Information Technology and Communications, Universitat Pompeu Fabra, c/Roc Boronat, 138, 08018 Barcelona, Spain
[b] School of Engineering (ETSE), Universitat Roviri i Virgili, Avenue Països Catalans, 26, 43007 Tarragona, Spain

ARTICLE INFO

Keywords:
Data privacy
Anonymization
Graphs and networks
Online social networks
Synthetic data generator
Information loss

ABSTRACT

In recent years, online social networks have become a part of everyday life for millions of individuals. Also, data analysts have found a fertile field for analyzing user behavior at individual and collective levels, for academic and commercial reasons. On the other hand, there are many risks for user privacy, as information a user may wish to remain private becomes evident upon analysis. However, when data is anonymized to make it safe for publication in the public domain, information is inevitably lost with respect to the original version, a significant aspect of social networks being the local neighborhood of a user and its associated data. Current anonymization techniques are good at identifying risks and minimizing them, but not so good at maintaining local contextual data which relate users in a social network. Thus, improving this aspect will have a high impact on the data utility of anonymized social networks. Also, there is a lack of systems which facilitate the work of a data analyst in anonymizing this type of data structures and performing empirical experiments in a controlled manner on different datasets. Hence, in the present work we address these issues by designing and implementing a sophisticated synthetic data generator together with an anonymization processor with strict privacy guarantees and which takes into account the local neighborhood when anonymizing. All this is done for a complex dataset which can be fitted to a real dataset in terms of data profiles and distributions. In the empirical section we perform experiments to demonstrate the scalability of the method and the improvement in terms of reduction of information loss with respect to approaches which do not consider the local neighborhood context when anonymizing.

1. Introduction

Data Privacy in graphs has recently become a topic of renewed interest by researchers, partially due to the emergence of online social networks (OSN), which can be represented and analyzed as graphs. OSN data is of great potential for data analysts from different disciplines, but also represents a threat to data privacy if information that users wish to remain private inadvertently becomes public. The recent release by Yahoo (Martin, 2016) of a 13.5 TB dataset of users' news interaction data raises issues of personal privacy risks, although the company declared the data release had followed their data privacy and anonymization practices. Thus, there is a need to anonymize the data before publishing it in the public domain and making it available to data analysts. However, a consequence of data anonymization is information loss. Hence, it is of interest to establish an equilibrium between information loss and privacy level. This is one of the focuses of our present work, especially in terms of what we call the 'local neighborhood' of a user.

On the other hand, data anonymization is still a process which requires specialist knowledge and calibration in order to achieve a result which is outside the skill set of a typical data analyst, or which is complex and time consuming even for a data analyst specialized in data privacy. Also, many data privacy analysts are presented with the difficulty of the lack of access to detailed and diverse datasets (especially previous to anonymization) describing online social network users, in order to carry out empirical testing. Sophisticated expert systems may be a solution to allow novice and experienced users to manage complex processes, however few systems of this type currently exist which can help the data analyst in these tasks.

Hence, in the present work we have developed a system which proposes to cover these issues: **(i)** the availability of data for testing, **(ii)** anonymizing data to a given privacy guarantee while preserving key local neighborhood information of interest in online social networks, and **(iii)** facilitating the anonymization process for

* Corresponding author. Tel.: +34 93 542 14 33.
E-mail address: david.nettleton@upf.edu (D.F. Nettleton).

[1] Part of this work was done when the authors were in the Universitat Oberta de Catalunya.

http://dx.doi.org/10.1016/j.eswa.2016.02.004
0957-4174/

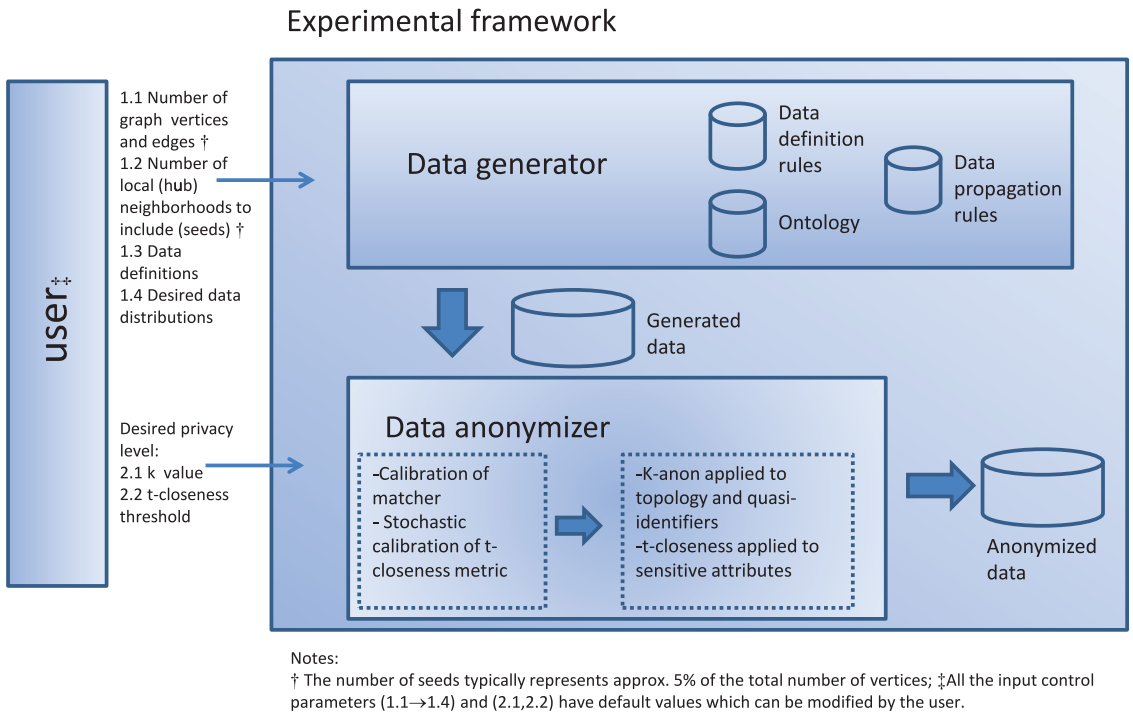

Fig. 1. Overall system architecture.

a user who is not a data anonymization specialist, or for expert users to reduce time dedicated to calibration and experimentation. Making this system available to the data analysis community will have a big impact in impulsing online social network analysis, while respecting user privacy.

In Fig. 1 we see a schematic representation of the overall system, which consists of two main components, a synthetic data generator and a data anonymizer. Both modules are assigned default input control parameters, which generates and anonymizes sunthetic online social network datasets which approximate a real datasets. Both the data definitions and how it is anonymized can then be changed by the user in an intuitive manner. With reference to Fig. 1, firstly, the data generation is defined in terms of the graph structure (number of vertices, edges and data propagation seeds) and the data to be assigned to the graph (data profile definitions and desired data distributions). The data is then generated using different rule sets for data propagation and matching (ontologies and distance functions). Secondly, the data anonymization is defined in terms of the privacy level (k-value and t-closeness threshold). In general, higher values of k and the t-closeness threshold mean that a greater anonymization is applied. Default values are available for all control parameters.

In this work we also address the challenge of stricter and stronger privacy guarantees, applying k-anonymity to the topology and quasi-identifiers and t-closeness to the sensitive attributes. This represents a major challenge for the complex data set which is used for testing. The t-closeness approach gives a stricter privacy guarantee than k-anonymity and even ℓ-diversity. However, there are few empirical implementations and to the best of our knowledge none for graph structured data.

Also, in order to obtain an optimum processing, we use a calibrated matching algorithm to choose k subgraphs for anonymization. In the literature, some authors have considered anonymization as a graph partitioning/clustering task based on an overall utility measure *(Hay, Miklau, Jensen, Towsley, & Weis, 2008)* or by modifying nodes using a cost function *(Zhou & Pei, 2008)*. However, to the best of our knowledge, all current methods have a high computational cost in the optimization step. We follow a different approach from the usual in that we optimize at a local level which avoids expensive global calculations (such as average path length). We also use a reduced (but representative) set of seed vertices to propagate the data locally. In this way, our system represents expert knowledge and embodies intelligent optimization and propagation techniques.

The primary contributions of the paper are:

- System which allows a non-expert user (that is a data analyst not specialized in data anonymization) and/or an expert user (facilitating his/her work) to create multiple online social network datasets that mimic a given social network and perform anonymization experiments on them. This may be used by varying the parameters k and t for multiple synthetic graphs and evaluate the trade-off between information loss (cost) and privacy level with respect to the original network, all this without having direct access to it (the original network is generated from its statistics: profile and attribute distributions).
- A key innovation in terms of the anonymization process is the preservation of the "local neighborhoods" of nodes, rather than just considering nodes as individuals, thus maintaining the social context of the users.
- Strong privacy guarantee for complex graph structured (social network) dataset: k-anonymity for the topology and quasi-identifiers and t-closeness for multiple sensitive attributes. The anonymization of the sensitive attributes is optimized using an intelligent temperature optimization process (simulated annealing) to find the minimum perturbation which obtains the threshold $t=0.20$ used by the t-closeness algorithm.

Secondary or auxiliary contributions of the paper are:

- An integrated sophisticated synthetic data generator which given a set of user specified data profiles and distributions, creates a data set which approximates a real online social network.
- A new approximation to the problem of graph anonymization consisting of a dense set of seeds with non-overlapping neighborhood subgraphs.
- A comparison of non-overlapping neighborhood subgraphs (our method) with (i) the case when they overlap and overlapping

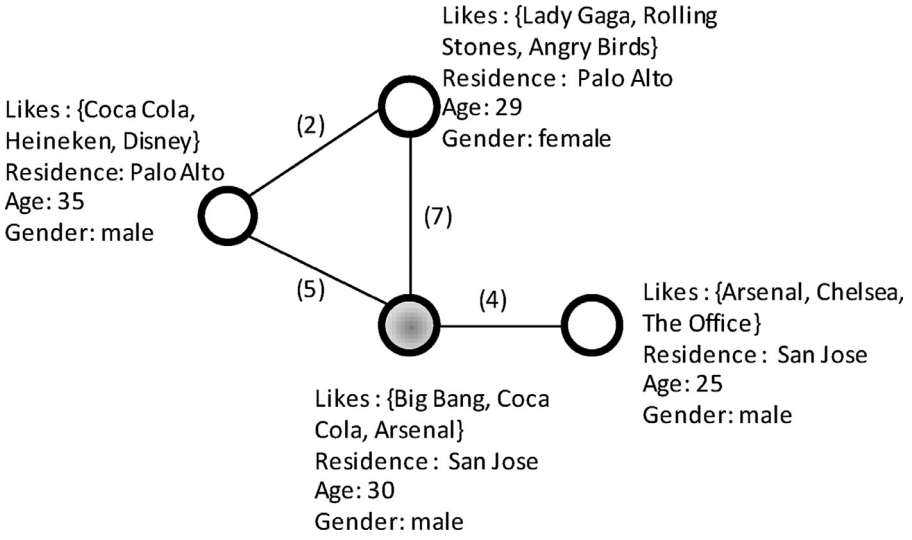

Fig. 2. Example local neighborhood for a node (grey shaded) in an online social network.

(ii) no local neighborhood subgraphs. In the latter case, represented by the method of Zhou and Pei (2011), individual nodes are anonymized and overall information loss is minimized in contrast to our method which minimizes information loss at a local level.

- Efficient matching algorithm to create k-group partitions each consisting of k non-overlapping sub-graphs for anonymization with minimum information loss.

The key innovation in terms of the anonymization process is that we anonymize "local neighborhoods" of nodes, and not individual nodes. In Fig. 2 we see an example of a local neighborhood in which the grey shaded node is considered as the reference user which has three (friend) neighbors, each with its descriptive attributes. The numbers on the links represent communications (such as average number of emails per week) between users. In social networks it is often the case that mutual friends have things in common such as (in general) living geographically close to each other, having similar ages and likes. A local neighborhood is considered as a reference node (which we call "seed") and its immediate neighbors. To the best of our knowledge, this is new to the state of the art. Other methods, such as Zhou and Pei (2011) have considered local neighborhoods for matching (that is, from the adversary point of view to detect the risk of disclosure), but have then anonymized on a node by node basis. Thus, in order to evaluate our method, we have designed experiments to measure the improvement of anonymizing local neighborhoods versus **(i)** anonymizing on an individual node basis, and **(ii)** a hybrid with partial overlap of the local neighborhoods.

The structure of the paper is as follows: in Section 2 the state of the art and related work is discussed; in Section 3 some preliminary concepts are presented; in Section 4 the anonymization method is described; in Section 5 the metrics are described for information loss and the privacy model is defined; in Section 6 we describe the synthetic data generation processing and the empirical results are presented for the anonymization of a range of dataset configurations, using different privacy levels and methods; finally, in Section 7 we discuss and summarize the present work.

2. Related work

In the following section we will first briefly comment some examples of expert systems, followed by a more detailed review of related work in the data anonymization field, which is the principal focus of our current work; the section is then finalized by some brief examples of approaches for synthetic data generator systems. Throughout the following sections, we contrast existing methods with the novel aspects and characteristics of our method.

2.1. Relevant expert systems

In the following section we will briefly comment selected exemplary expert systems whose design and user objectives can be extrapolated to the current work.

In Roby, Phillips, Thomas, and Sprouse (2012) a project is described whose objective was to create an expert system process that would analyze and parse raw sequence data, including an advanced sequence analysis program to improve automation of the routine and repetitive tasks involved in the interpretation of mtDNA sequence analysis. A spreadsheet type interface was used in which the end user can define a rule which is then validated against the data. As a result, different alerts can be given which indicate if the data 'does', 'may' or 'does not' exhibit the characteristic defined by the rule. Different metrics are used to quantify the data quality. The matching algorithm is based on sequencing, which is a different proposal to our work which matches subgraphs. The user interface is well designed and as future work we would incorporate some of their ideas such as the color scheme for different degrees of privacy risk or information loss. Our current user interface is limited to the Java project execution in Eclipse.

In Prasser and Kohlmayer (2015) the ARX Data Anonymization Tool is presented. It is aimed at the anonymization of biomedical data in order for researchers in that field to have access to data which includes sensitive personal attribute-values. ARX has the following main features: (1) models for analyzing re-identification risks, (2) risk-based anonymization, (3) syntactic privacy criteria, such as k-anonymity, ℓ-diversity, t-closeness and d-presence, (4) methods for automated and manual evaluation of data utility. In contrast to our approach which is designed for graph-structured data, ARX processes only tabular structured data.

Rubio, De la Sen, Longstaff, and Fletcher (2013) present a modular expert rule based system which facilitates the automatic selection of cutting parameters in milling operations. Although this represents a different data modeling domain problem to our own, their system is exemplary in terms of modular design, parameter calibration and end user motivations, which can be generalized to the current work. The system defines a cost function in terms of five key control parameters, which are associated with the milling

process setup. In our case the control parameters are the data definitions and the desired privacy level, whereas the cost function measures the information loss due to anonymization.

2.2. Anonymization methods

In the context of data privacy in general, k-anonymity was first defined following concepts proposed by Samarati and Sweeney (1998), Samarati (2001) and Sweeney (2002). More recently the paper by De Capitani di Vimercati, Foresti, Livraga, and Samarati (2012), gave key definitions for privacy levels, information loss and risk of disclosure. The objective of k-anonymity is to guarantee that each object to be anonymized in the dataset is made to be indistinguishable from k-1 other objects. K-anonymity forms the basis of the anonymization process we apply in the current work.

Later, several shortcomings were identified in the privacy guarantee of k-anonymity. New proposals were made to correct these problems, such as ℓ-diversity, defined in (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006), and later t-closeness, defined in (Li, Li, & Venkatasubramanian, 2007). Our approach uses k-anonymity as the basis and combines it with t-closeness to augment the privacy guarantee for the sensitive attributes.

In the survey by Zhou, Pei, and Luk (2008), graph anonymization methods were conveniently divided into two groups: (a) node modification approaches and (b) node clustering approaches. Our approach belongs to the second group of graph anonymization methods.

In the context of graph anonymization, *Zhou and Pei (2011)* extended k-anonymity with ℓ-diversity for graph anonymization. Other variants which have been proposed are b-likeness (Cao & Karras, 2012) and p-sensitivity (Truta, Campan, & Meyer, 2007). In the case of Zhou and Pei (2011), the authors use an elaborate coding scheme, based on a depth first search of a tree representation of the local graph neighborhood. The cost function consists of three parts: (i) information loss due to generalizing vertex labels, (ii) information loss due to adding edges and (iii) the number of vertices linked to the anonymized neighborhoods in order to achieve k-anonymity. Synthetic datasets were generated for testing but the only information apart from the topology itself was a sensitive label assigned to each vertex, which was a uniformly distributed random number. We use a modified version of this approach later in Section 6.2 as one of our benchmark methods.

Similar to Zhou and Pei (2011) we consider graphs with labeled nodes. However, contrastingly, in our approach the information attached to a node corresponds to an entry on a table, which has quasi-identifiers and sensitive attributes. This difference together with our assumption that the edges of the graphs are weighted meant that the DFS coding from Zhou and Pei (2011) was substituted by an efficient matching function (see Section 4.1.1) in our framework. Using this modified version (designated 'NoS' in Section 6.2), we used it for validating our seed assignment with non-overlapping local neighborhoods, in our comparison for the empirical testing. Also, we used our own synthetic data generator, cost function and matching method, as are described later in Sections 4 and 5 of this paper.

In (Hao, Cao, Hu, Bhattarai, & Misra, 2014) graphs with structural and textual information are considered, their approach to structural anonymity is k-degree anonymization and they consider textual information only on the edges of the graph. Similarly, (Yuan, Chen, Yu, & Yu, 2013) provide k-degree anonymity with ℓ-diversity of the vertex labels. Therefore, our privacy guarantees are stronger as we combine k-anonymity with t-closeness, which is superior to k-anonymity with ℓ-diversity.

In the context of the anonymization of weighted graphs, that is graphs with weight values assigned to the edges, the solutions tend to be quite similar (see, for example: Das, Egecioglu, and El Abbadi (2010); Liu and Yang (2011); Skarkala et al. (2012) using an information loss minimization function to assign weight range generalizations.

In Liu and Yang (2011), the authors calculate max-min values based on the edge weights to categorize preselected edge weights into ranges, and using a cost function. New edges can be added with an associated percentage value which indicates the probability that the corresponding edge exists. Current edges can also have a probability value assigned. In the current work we have adopted the approach described in Liu and Yang (2011) for anonymizing weighted graphs, and have made some simplifications, as follows. Firstly, we do not assign a probability of existence to the edges. Secondly, we convert all weight values into ranges, instead of selected ones.

The anonymization of labeled graphs has been considered by different authors in the literature (see, for example: Bhagat, Cormode, Krishnamurthy, and Srivastava (2009); Campbell, Dagli, and Weinstein (2013); Clifton and Tassa (2013); Heatherly, Kantarcioglu and Thuraisingham (2013); Song, Karras, Xiao, and Bressan (2012).

Some theoretical approaches to k-anonymity for graphs can be found in Chester, Kapron, Srivastava, and Venkatesh (2013), Salas and Torra (2015, 2016) and Hartung, Nichterlein, Niedermeier, and Suchý (2015). In Salas and Torra (2015, 2016) an optimal solution for k-degree anonymity is presented, hence a bound on the minimum distance that a general k-anonymous graph may have to the original graph. In the present work we use the cost function to measure distance. An efficient algorithm for obtaining k-degree anonymous graphs can be found in (Casas-Roma, Herrera-Joancomartí, & Torra, 2013).

In Chester et al. (2013) graphs with labeled edges and vertices are studied, they focus mainly in the complexity (hardness) of different notions of k-anonymity. In particular they prove that t-Closeness is NP-complete if the equivalence classes are required to be k-vertex label sequence anonymous. This means that the best we can do to guarantee t-Closeness is a heuristic (this is the approach of the current work). Another contribution of Chester et al. (2013) is the study of the case when there are individuals that are not to be anonymized (public), motivating the problem of anonymizing only subsets of the all the vertices. Our approach could be slightly modified to provide anonymizations of subsets of the vertices. On the other hand, we do not add fake nodes as in Chester et al. (2013).

We will now specifically consider two of the main approaches for graph anonymization: **(i)** node modification and **(ii)** node clustering.

2.2.1. Node modification approaches

Node modification approaches act by choosing similar nodes and making them identical. This can be done by adding nodes to make their degrees the same and by adding edges to make their immediate neighborhood topology the same. Using this method, $k-$anonymity is achieved by obtaining that every node in the graph has at least k-1 other nodes which are indistinguishable from it. Zhou and Pei (2008) presented a method which selects nodes based on a cost function and then anonymizes them by adding nodes and edges to their neighborhoods. Nettleton, Sáez-Trumper, and Torra (2011) compared two different types of online social network from a data privacy perspective, using 'add link' as the graph modification operator (that is, a node modification approach). Hay, Miklau, Jensen, Weis, and Srivastava (2007) presented a simple but effective graph anonymization method based on random addition and deletion of edges. The disclosure method attempts re-identification using two types of queries: vertex refinement and sub-graph knowledge. Nettleton, Torra, and Dries (2014) benchmark two contrasting graph anonymization methods, node clustering (the method used in the current work) and node

modification, applied to online social network (OSN) graph datasets. The authors incorporate constraints into the anonymization process which implement user defined utility requirements for the community structure of the graph and major hub nodes. In our present work, we implement a node clustering approach, rather than a node modification approach.

2.2.2. Node clustering approaches

This is the approach we use in the system presented in this paper. Node clustering approaches act by choosing similar nodes and physically grouping them. This can be done by a *k*-Means type algorithm or by a similarity/distance metric to choose similar nodes. Using this method, *k*-anonymity is achieved by obtaining that every node in the graph is incorporated into a cluster within which there are at least *k*-1 other nodes. Skarkala et al. (2012) present an approach for node clustering/grouping which takes into consideration the privacy protection of the edge weights. Skarkala employs a similarity function to form clusters each containing at least k nodes. Nettleton (2012) presents a perturbation method based on node aggregation and a similarity metric with fixed weights for choosing node pairs. Different types of clustering, fuzzy (fuzzy *c*-Means) and crisp (*k*-Means) were applied to graph statistical data in order to evaluate the information loss due to perturbation. In (Hay et al. (2008), an approach is presented in which nodes are grouped into partitions based on a utility function incorporating a distance metric in terms of the number of edges. In order to settle the partitions, the entropy was calculated for the entire graph. Hay's method is distinct to our approach given that Hay's partitions are guaranteed as having at least *k* nodes but can have many more (e.g. hundreds, for $k = 16$), whereas our method guarantees between *k* and 2*k*-1 nodes in each cluster.

2.2.3. Numerical and categorical data anonymization

Loukides and Shao (2011) compare a novel heuristic for guaranteeing the diversity of range values with ℓ-diversity and *t*-closeness, claiming a lower information loss. Loukides and Gkoulalas-Divanis (2012) present a novel anonymization algorithm, called Update-Anonymize-Reorder (UAR), which incorporates user definable constraints for utility and privacy. Yang and Qiao (2010) present an anonymization method which acts by randomly breaking links among attribute values in records. The authors claim that the data randomization method maintains statistical relations among data to preserve knowledge for the quasi-identifiers and sensitive attributes, and compare the results with ℓ-diversity. However, both these approaches minimize information loss globally whereas in our approach we optimize information loss locally (at a local neighborhood level) and then the global information loss is the sum of these.

2.3. Synthetic data generator approaches

For convenience, the related work will be divided into two main areas: (i) synthetic topology generation without data and (ii) generating a topology together with data which is associated with the nodes and edges of the topology. We note that the full details of the synthetic data generator are out of the scope of the current paper, whose focus is on the anonymization method and its evaluation. The in-depth details and benchmarking of the data generator will be published in a separate paper.

2.3.1. Synthetic topology generation

Rmat, presented in Chakrabarti, Zhan, and Faloutsos (2004), is a commonly used method which employs a statistical approach and a recursive process to replicate the power law distributions, skew distributions and community structure (which can be hierarchical), while maintaining a small diameter for the graph. We have used Rmat in the present work to generate the graph topology, which is given as inputs the number of nodes and edges required. The works of Boncz et al. (2014) and Pham, Boncz, and Erling (2013) have reported that the generated topologies have communities with a similar size, instead of the long tail distribution found in real OSNs. A model called "Forest Fire" (with reference to the way link creation propagates), is presented by Leskovec, Kleinberg, and Faloutsos (2005). In terms of structure, the "rich-get-richer" (or preferential attachment) phenomenon is cited as the explanation of the heavy tailed in-degree power-law distribution. In the current work we have used the Rmat generator to create the topology, and then populated it with data in a second step. Hence, the issue of the topology generator *per se* is not the focus of the current work.

2.3.2. Synthetic data and topology generation

Different approaches exist to generating the synthetic data and topology, which tend to be customized for specific applications such as business contact networks as studied in Pérez-Rosés & Sebé, 2015 and Pérez-Rosés, Sebé, and Ribó (2016), online gaming such as Ali, Alvari, Hajibagheri, Lakkaraj, and Sukthankar (2014) or census data exemplified by Barrett et al. (2009). Firstly, the modeling approach of Pérez-Rosés and Sebé (2015), and Pérez-Rosés et al. (2016) is to simulate a LinkedIn network by defining a set of skills and for each skill define a directed graph where the nodes correspond to users' profile and the arcs represent endorsement relations. Our approach is more generic and designed to be domain independent, more along the lines of Boncz et al. (2014) and Pham et al. (2013) whose approaches will be explained below. Our data population approach follows similar lines to that of Boncz with respect to the homophily rules of neighbor characteristics. However, whereas Boncz generates the topology and assigns the data simultaneously, our approach first creates a topology then populates the finished topology with data. We propose that our approach will make it possible to realistically populate real topologies, if they are made available, for example, by online social network application providers. Previously, Nettleton (2015) defined an initial version of a synthetic OSN data generator was described for non overlapping communities using the RMat generated topology as defined in Chakrabarti et al. (2004), and simple control parameters. The version of Nettleton (2015) is the basis of the data generator in the current work.

3. Preliminaries

In the following we will give some basic terminology and definitions which will then be referred to in the remainder of the paper.

A graph G is defined as a set of vertices V interconnected by a set of edges E, denoted by $G = (V, E)$.

In this work, for modeling social interactions we assume that the graph is a weighted graph, that is for each edge *e* it has associated a numerical weight value *w(e)* between 1 and 10 as a quasi-identifier.

We consider the weighted graph *G* together with a table *T*, in which each tuple corresponds to a vertex *v* and has $\{q_1, q_2, q_3\}$ as quasi-identifiers and attributes $\{a_1, a_2, a_3\}$ as sensitive attributes.

We denote the closed neighborhood of a vertex $u \in V(G)$ by $N(u)$ and it consists of all the neighbors of *u* in *G* together with *u* and all the edges of *G* that connect them.

When the entire graph is considered and all the local neighborhoods have to be anonymized, if a node belong to many different neighborhoods, after the anonymization process it will belong to many different *k*-groups, hence constraining a large number of nodes in the anonymized graph to have the same characteristics.

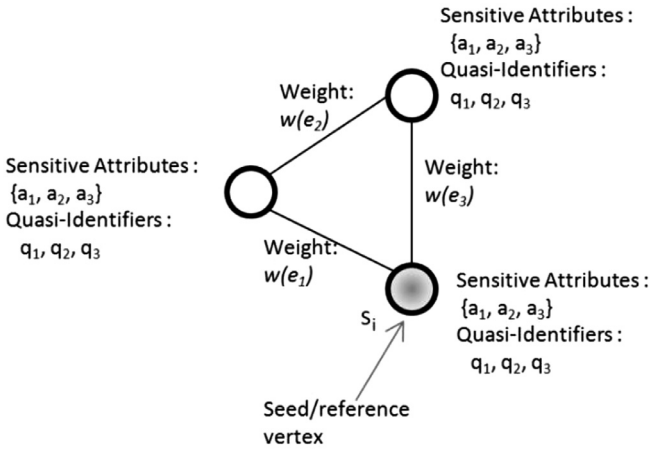

Fig. 3. Example sub-graph and data definition.

This intuition is formalized on the following remark that motivates our approach to the problem.

Remark 1. For any vertex z in a given graph G, let $\{z_1,\ldots, z_r\}$ be the set of neighbors of z such that their closed neighborhoods $N(z_1),\ldots, N(z_r)$ all belong to different k-groups after anonymization. There are at least kr vertices with the same characteristics as z in the anonymized graph G'.

Note that $z \in N(z_1) \cap\ldots\cap N(z_r) \neq \emptyset$. Denote the sets in the k-group of $N(z_1)$ as $N^1(z_1),\ldots, N^k(z_1)$, and similarly for all $N(z_2),\ldots, N(z_r)$.

Since the graph G' is k-anonymous by neighborhoods, there are vertices $z_{11} \in N^1(z_1)$, $z_{12} \in N^2(z_1),\ldots, z_{1k} \in N^k(z_1)$ such that all its characteristics (in the anonymized graph G') are equal to those of z, and so on for all the other k-groups of $N(z_2),\ldots, N(z_r)$. Therefore in a k-anonymized version of the graph, there have to be $z_{11},\ldots,z_{1k}$, $z_{21},\ldots,z_{2k},\ldots$, $z_{r1},\ldots,z_{rk}$ vertices with the same characteristics as those of z.

Hence z would belong to a group of kr elements with the same characteristics.

In order to tackle this problem we suggest a different approach for the k-anonymization for the information rich OSN graphs.

We use a set of seed nodes that are going to be chosen with the following properties:

- Each seed has to have distance at least 3 to all the other seeds,
- Each node of the original graph G is at distance at most 2 to some seed node.

It is a natural assumption that the OSN graphs have to be similar between close acquaintances, hence, the condition of having a seed vertex at distance at most 2 guarantees that the vertices that are out from the set of seeds are at distance at most one from some seed's neighbor and therefore will intuitively be well represented.

Denote the set of seed vertices as $S=\{s_1, s_2, \ldots, s_-\}$.

For a seed vertex s_i, our anonymization method chooses the k-1 seed vertices $s_2, \ldots, s_k$ whose sub-graphs $N(s_2), \ldots, N(s_k)$ are most similar to $N(s_i)$. These sub-graphs are the ones to be anonymized. The matching is based on a distance function $D(s_i, s_j)$, that considers the difference between the characteristics of graphs $N(s_i)$ and $N(s_j)$.

In Fig. 3 we see an example sub-graph which consists of one seed (shaded) and two neighbors. Each vertex has a corresponding set of quasi-identifiers and sensitive attributes, and each edge has a numerical weight assigned, which is also considered a quasi-identifier.

4. Metrics for information loss and privacy level

In this Section, the definitions are given for information loss and privacy level. We also define the measure used for sub-graph matching. Information loss is defined as a distance between the original sub-graphs and the medoid sub-graph of each k-group.

4.1. Information loss

Three metrics are used in order to evaluate information loss. The first is based on graph topological characteristics (degree, clustering coefficient, degrees of neighbors) for each node and its immediate neighborhood, and the other two are related to the attribute values (quasi-identifiers and sensitive values).

- **inf loss$_1$** 'TOP' (degree, clustering coefficient, degrees of neighbors)
- **inf loss$_2$** 'QUASI' (quasi-identifiers)
- **inf loss$_3$** 'SENSI' (sensitive attributes)
- **inf loss$_4$** 'ALL') weighted sum of first three information loss values.

The names 'TOP', 'QUASI', 'SENSI' and 'ALL' are used later in the empirical Section 6.2 to refer to the information loss metrics.

For a given value of k, the data is partitioned into groups each of which contains at least k sub-graphs H_{ij}. The information loss is calculated in terms of a 'cost distance', which we use to calculate the medoid sub-graph P_i of each k-group, and then to calculate the sum of the distances of each sub-graph H_{ij} in the k-group, to the medoid sub-graph P_i. Hence for a given k-group C_i of cardinality N_i, its cost is then obtained:

$$Cost(Ci) = \sum_{j=1}^{N_i} dist\left(P_i, H_{ij}\right) \tag{1}$$

Then we calculate the average distance for all k-groups (which are M and represent the complete graph dataset). A final average value will be interpreted as the information loss measure for a given value of k:

$$Avg = \frac{\sum_{i=1}^{M} Cost(Ci)}{M} \tag{2}$$

In the following Section we see how to define the distance between two sub-graphs for the neighborhood sub-graph matching operation. In this case the minimum distance, or cost, implies the closest match.

4.1.1. Similarity metric for sub-graph matching

In order to calculate the similarity between two seed neighborhoods, computational efficiency is a key consideration. Hence, a similarity metric has been devised which calculates a distance based on sub-graph characteristics which can be pre-calculated. The sub-graph has three major characteristic groups: (i) topology; (ii) quasi-identifier attribute-values; (iii) sensitive attribute-values. Example data assignments can be seen later in Figs. 5–7.

The matching algorithm is effectively a tradeoff between the different groups and individual characteristics which can be weighted. The weights are calibrated by an optimization process with the distance as the fitness value. The sub-graph matching method used in this work has been presented as a Patent application in Nettleton and Dries (2013).

The idea of the matching function is to find the most similar sub-graphs to assign to a given k-group, because the most similar sub-graphs will require the least changes to make them the same.

Once we have performed the matching, we can calculate information loss by quantifying the difference between the matched

Table 1
Relative distances between residence categories.

Residence category	Distance
Same town	0.00
Same county	0.25
Same state	0.50
Same division	0.75
Same region	0.90
Not in same region	1.00

sub-graphs. In a similar manner to the information loss calculation, the matching function is composed of three components: topological (T), quasi-identifiers (Q) and sensitive attributes (S). If, for a given sub-graph H, T, Q and S represent the topology, quasi-identifiers and sensitive attributes, respectively, and T', Q' and S' represent the corresponding entities of a second sub-graph H' which has been matched with H, then:

$$dist(H, H') = (\Delta(T, T')) + (\Delta(Q, Q')) + (\Delta(S, S')) \quad (3)$$

Where $\Delta(T,T')$ signifies the difference (delta) between T and T', and σ, α and β are user defined weights. After empirical testing, we settled the weights to γ=0.50, α=0.25 and β=0.25, which gave better results for overall matching and specifically for the topological structure.

In order to calculate the distance between the different attributes, we use the following categorizations for the quasi-identifiers and the sensitive attributes. The categorizations have been defined taking into account typical demographic and marketing criteria. Age is categorized as follows: "18–25", "26–35", "36–45", "46–55", "66–75", "76–85". Gender can be "male" or "female". Residence has a hierarchy of five levels, following the geographical definition of the United States: town, county, state, division and region. The edge weight is categorized as "1–2", "3–5" and "6–10". The likes are categorized into five categories: "entertainment", "music artist", "tv show", "soccer club", "drink brand". We note that each vertex has assigned the corresponding user's top 3 likes. The categorization of the likes gives a solution to the problem of calculating distances between heterogenous values, such as 'Lady Gaga' and 'Justin Beiber' (both classed as 'music artist'), or 'Lady Gaga' and 'Heineken' (classed as 'music artist' and 'drink brand', respectively). Inevitably, we lose precision in the values due to this classification, however, for the purposes of the current work, we evaluated that the classification provides an adequate diversity to test the anonymization process. This was concluded after testing with subclasses, such as 'rock artist' (e.g. Rolling Stones) and 'pop artist' (e.g. Lady Gaga), 'soft drink' (e.g. Coca Cola) and 'alcoholic drink (e.g. 'Heineken'), for which the results (*t*-closeness and information loss for different values of *k*) did not differ significantly due to the greater diversity.

Having defined the categorization, the distance between attribute-values is calculated as follows:

'TOP' is calculated by using normalized numerical values for the respective degrees and clustering coefficient.

'QUASI' is calculated on an individual attribute basis: the distance value of 'gender' can be 1 or 0 ({male, male}=0; {male, female}=1). 'Age' is an ordinal category whose distance goes from 0 to 5, given that there are six possible categories, as defined previously. For example, categories "18–25" and "46–55" are at distance 3. The distance between the edge weight categories is calculated in a similar manner. For example, categories "1–2" and "3–5" are at distance one and categories "1–2" and "6–10" are at distance two.

'Residence' uses a hierarchical tree, with five levels (defined previously), and their mutual distances are shown in Table 1. For example, two residences which are in the same town are at distance 0.00, otherwise if they are in the same county their distance is 0.25, otherwise if they are in the same state their distance is 0.50, and so on.

Finally, 'SENSI' has a customized similarity weighting for the categorized 'like' values, as can be seen in Table 2. For example, the like categories 'entertainment' and 'music artist' are at distance 0.25, whereas 'music artist' and 'drink brand' are at distance 0.50. Again, we have chosen categories which give overall coherent results but which could further refined. For example, in the case of a particular music artist who appears in a given drink brand commercial, we could reduce the 'like' distance between these categories.

The insight of using the distance for the information loss value is that if the sub-graphs are more similar (closer together) in a given k-group, then when they are all made the same (*k*-anonymized), the information loss will be lower than if the sub-graphs are less similar (further away).

4.2. Definitions for privacy approaches

In this section we will define three notions of privacy relevant to the present work: *k*-anonymity, ℓ-diversity and *t*-closeness. Other relevant issues described are the approach to the anonymization of weighted edges and the treatment of multiple sensitive attributes.

4.2.1. k-anonymity

The objective of *k*-anonymity is to hide a record among *k-1 other records*. Nodes are grouped into groups of at least *k*, based on similarity using the distance metric described in Section 4.1.1. That is, in terms of the topological characteristics, after the anonymization for each node there will be *k-1* other nodes with the same degree, number of edges in the sub-graph, and same clustering coefficient (that is, the connectivity between neighbors). Hence, the probability of an adversary re-identifying a node will be at most *1/k*, based on these criteria. It is noted that nodes which are already identical will not be modified and there will probably be nodes in the graph which already have more than *k* identical nodes (especially those with a low degree). *In the current work we obtain a new type of model: k-subgraph anonymity for a dense sample. This is primarily due to the seed nodes being hub nodes (high degree and connectivity).*

Table 2
Relative distances between like categories.

	Entertainment	Music artist	TV show	Soccer club	Drink brand
Entertainment	0.00	0.25	0.25	0.25	0.50
Music Artist		0.00	0.25	0.25	0.50
TV show			0.00	0.25	0.50
Soccer club				0.00	0.25
Drink brand					0.00

4.2.2. ℓ-diversity

As a consequence of a deficiency identified of k-anonymity, ℓ-diversity was defined in Machanavajjhala et al. (2006). It addresses the problem that sensitive values in an equivalence class lack diversity and thus may facilitate the identification of an individual. It assumes the attacker has background knowledge. The sensitive attributes must be 'diverse' within each quasi-identifier equivalence class.

ℓ-Diversity Principle Machanavajjhala et al. (2006)**:** A q$\star$-block is ℓ-diverse if contains at least ℓ "well-represented" values for the sensitive attribute S. A table is ℓ-diverse if every q$\star$-block is ℓ-diverse.

Entropy ℓ-Diversity Machanavajjhala et al. (2006)**:** A table is Entropy ℓ-Diverse if for every q$\star$-block

$$-\sum_{s\in S} p_{(q^*,s)} log\left(p_{(q^*,s)}\right) \geq log(l) \tag{4}$$

where

$$p_{(q^*,s)} = \frac{n_{(q^*,s)}}{\sum_{s'\in S} n_{(q^*,s')}} \tag{5}$$

is the fraction of tuples in the q$\star$-block with sensitive attribute value equal to s.

Multi-Attribute ℓ-Diversity Machanavajjhala et al. (2006)**:** Let T be a table with non-sensitive attributes $Q_1, \ldots, Q_{m1}$ and sensitive attributes $S_1, \ldots, S_{m2}$. We say that T is ℓ-diverse if for all $i=1 \ldots m_2$, the table T is ℓ-diverse when S_i is treated as the sole sensitive attribute and $\{Q_1, \ldots, Q_{m1}, S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_{m2}\}$ is treated as the quasi-identifier.

4.2.3. t-closeness

As a consequence of a deficiency identified in k-anonymity and ℓ-diversity, t-closeness was defined in Li et al. (2007). This is because the distribution is not considered by ℓ-diversity. The basic idea is that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the complete dataset. Also, the distance between two distributions should not be more than a given threshold t.

The t-closeness Principle Li et al. (2007)**:** An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. Empirically, a threshold of approx. 0.2 is often cited in the literature, although the optimal value is data dependent.

Li et al. (2007) outlines a method to measure the distance between two probabilistic distributions. There are a number of ways to define the distance between them. Given two distributions $P=(p_1, p_2, \ldots, p_m)$,$Q=(q_1, q_2, \ldots, q_m)$, the authors use the Earth Mover Distance, with two variants, one for numerical attributes and another for categorical attributes.

4.2.4. Edge anonymization

In Liu and Yang (2011), the authors calculate max–min values based on the edge weights to categorize preselected edge weights into ranges, and using a cost function. New edges can be added with an associated percentage value which indicates the probability that the corresponding edge exists. Current edges can also have a probability value assigned.

In the current work we have adopted the approach described in Liu and Yang (2011) for anonymizing weighted graphs, and have made some simplifications, as follows. Firstly, we do not assign a probability of existence to the edges. Secondly, we convert all weight values into ranges, instead of selected ones.

4.2.5. Multiple sensitive attributes

A greater challenge is presented by a dataset in which more than one of the attributes is considered "sensitive". As usual, we consider a dataset which is composed of several quasi-identifier attributes which are anonymized, for example, by generalization. In the case of the sensitive attributes, ideally, they would be left unchanged. However, we must achieve that an adversary query returns not less than k distinct records from any k-group. We must also guarantee the property of t-closeness for its distribution. Different approaches exist in the literature (see, for example: Das & Bhattacharyya, 2012; Gal, Chen, & Gangopadhyay, 2008; Maheshwarkar, Pathak, & Choudhari, 2012). In (Das & Bhattacharyya, 2012), sensitive attributes are placed in a distinct table from the quasi-identifiers, and are thus disassociated from individual records. Instead, they are associated with a given k-group, and noise is added to obtain ℓ-diversity. In (Gal et al., 2008), the sensitive attributes are again transferred to separate tables, one for each group, and records are added to each table in order to obtain ℓ-diversity.

Our approach: we follow the scheme of defining the sensitive attribute-values in a separate table. Each entry is thus disassociated from the original record. We then process this table to obtain a distribution which complies with t-closeness to a given threshold. The threshold we use is 0.20, which is commonly stated in the literature. We use a temperature optimization process (simulated annealing) to find the minimum perturbation which obtains t=0.20.

5. Description of anonymization approach

In this section the anonymization approach used in the present work will be described. There are three aspects which are anonymized: topology, quasi-identifiers and sensitive attributes. The method is based on selecting the k most similar sub-graphs and then perturbing them to make them identical, by modification. The distance metric used for sub-graph matching has been described in Section 4.1.1.

It is important to note that the graphs in group of k are chosen based on a similarity distance. The k graphs must be as mutually similar as possible so that when we modify them to make them identical, it will cause the minimum perturbation/information loss.

5.1. Anonymization

We use the following steps to anonymize the graph: assign seeds; for each seed find k-1 most similar sub-graphs, whose reference nodes are also seeds, thus defining the k-groups; anonymize topology for each k-group; anonymize quasi-identifiers for each k-group; anonymize sensitive-attributes for each k-group.

5.1.1. Seed assignment

The seed assignment has to comply with two criteria:

(i) Each seed node must be at least at distance 3 from any other seed node, so that each sub-graph (with the seed as center) can be modified without perturbing the adjacent sub-graphs. In order to achieve this, seed nodes are added one by one, and checked for distance using a proximity routine. This routine checks if a given node is an immediate neighbor (distance = 1) or neighbor of a neighbor (distance = 2). If the given node is not a neighbor or a neighbor of neighbor, then it must be at distance 3 or more. The number of seed nodes assigned is initially defined as the total number of vertices in the graph divided by the average degree.

(ii) The distributions of the characteristics of the sub-graphs formed by the seed node neighborhoods must be within γ of the corresponding distributions in the complete graph.

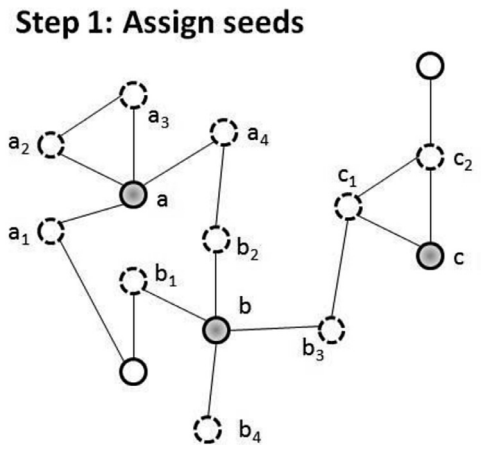

Fig. 4. Seed assignment.

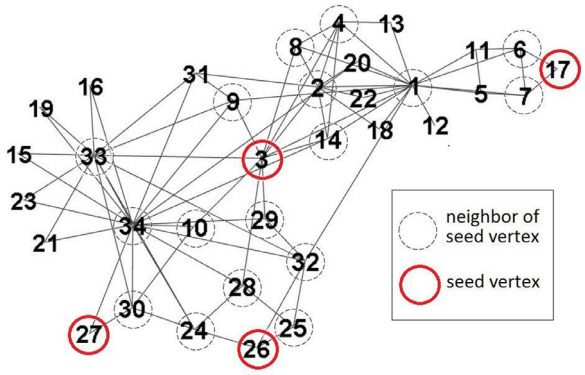

Fig. 5. Example of seed and seed neighbor assignment (Karate graph dataset).

The characteristics are defined as: *degree distribution, clustering coefficient, distributions of quasi-identifiers and sensitive attributes.*

Fig. 4 shows a generic example of the seed assignment process, for which the seeds are labeled as *a, b* and *c*. We note that each seed is at a shortest distance of three from any other seed, hence the immediate neighborhoods do not overlap.

The assignment of the seed nodes is actually an optimization process. It is possible that the random assignment of seeds, especially the first seed, can result in a sub-optimal assignment. For example, if the first seed is assigned to a major hub node, the average path length to it of many nodes in the graph will be short.

Coverage: the assignment of the seed nodes in the manner described has a coverage of between 20% and 50% of the nodes in the graph. This is because isolated nodes tend to remain between sub-graphs which cannot be assigned due to the minimum distance requirement between seeds ($\geq$ 3) and because the seed sub-graphs cannot overlap.

In Fig 5 we see an example of the assignment of four seed vertices (with ids 3, 27, 26 and 17) and the corresponding neighbor vertices of the seeds. The neighbors of each seed are as follows: 3, {33, 9, 2, 8, 1, 14, 4, 29, 28, 10}; 27, {34, 30}; 26, {24, 25, 32}; 17, {7, 6}. In total, 21 vertices are assigned from a total of 34, giving a coverage of 62%). Vertices such as 21, 23, 15, 19 and 16 cannot be assigned as seeds because their neighbors would overlap with an existing seed sub-graph (those of seeds 27, 26 and 3). As mentioned previously, the distributions of the characteristics of the assigned vertices (topology and data) are checked for similarity with that of the complete graph. In this way we guarantee a sample which has a statistical quality sufficient for the type of end user of the published data (a data analyst).

However, from a sampling point of view we guarantee two aspects:

(i) The maximum number of possible seeds has been assigned, resulting in an optimal or close to optimal coverage of the complete graph, given the restriction that seeds must be at a distance of three or more from each other.
(ii) The distributions of the characteristics of the seed sub-graphs (seeds and their neighbors) has a given similarity to the corresponding distributions of the characteristics in the complete graph (all the nodes of the graph).

Hence, having guaranteed the two above points, we effectively extract a large sample which is statistically representative of the complete graph. We note that it is the sample which will be anonymized and published.

Procedure seed assigner

Input: graph G, δ=0.05
Output: seed set S
1. **Assign** *Seeds*
2. **While** S does not comply **do**
3. **While** S is not topologically optimal **do**
4. **Choose** a random vertex w such that:
Each s δ S is at least at distance 3 from w.
Check optimality of S
5. **End do**
6. **Check** that distributions of key metrics and attribute-values of s $\in$ S are within margin $\in$ of corresponding distributions of v δ G
7. Degree distribution, clustering coefficient, distributions of quasi-identifiers and sensitive attributes.
8. **Assign** compliance of S
9. **End do**

5.1.2. Matching to form a k*-group*

We consider the immediate neighborhood of each seed vertex as a sub-graph. The seed vertices are ordered by decreasing degree. The first seed vertex in the ordering (i.e. which has the biggest degree) is selected. Let's call this vertex s_1. A matching function finds the *k-1* seed vertices whose sub-graph is most similar to that of s_1. The matching function takes into account the similarity of the topology, the quasi-identifiers and the sensitive attributes. This process is depicted for a given k-group in Fig. 6.

5.1.3. Sensitive attribute anonymity

The sensitive attribute we consider is 'likes', which has three values per record. As the sensitive attribute is a tuple, we apply the method for multiple sensitive attributes by decomposition into a separate table per group.

We use the *t*-closeness metric in order to guarantee a privacy level for the sensitive attributes. Firstly, the *t*-closeness value is calculated for the sensitive attribute-values of each *k*-group. Next, we identify the *k*-groups whose "*t*" value is greater than the given threshold (0.2) with respect to the complete graph dataset. If a tuple does not comply, the frequency distribution is modified using a temperature value which is incrementally increased until the threshold of 0.2 is reached. A temperature of 0.0 would give as a result a t-closeness of zero, whereas a temperature of 1.0 will leave the values as they currently are.

For these groups, we progressively modify the sensitive attribute values until the "*t*" values is less than or equal to 0.2, while minimizing information loss. We note that 0.2 is a value which in the literature is considered optimum, which minimizes information loss for the given privacy level which is represented by the value of *t* in this case. This process is depicted for a given k-group in Fig. 7.

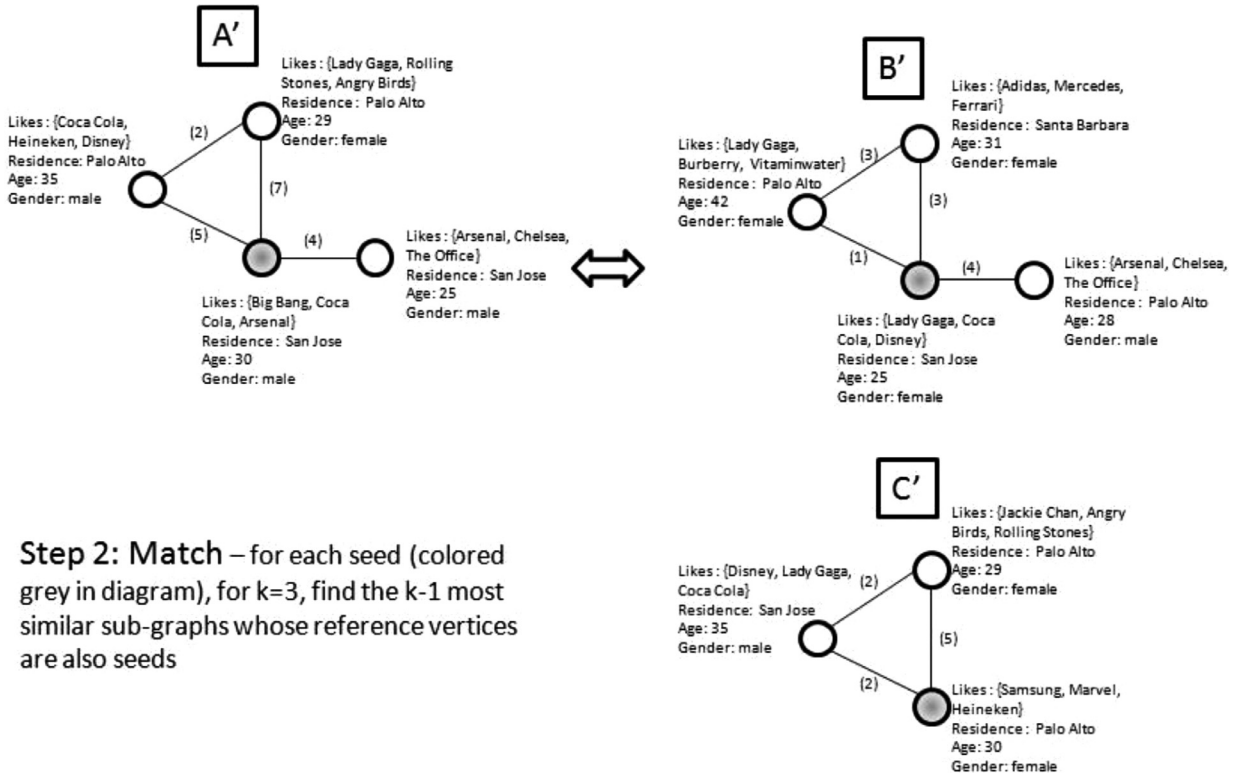

Fig. 6. Match. Seeds are chosen and for each seed the k-1 most similar sub-graphs are identified.

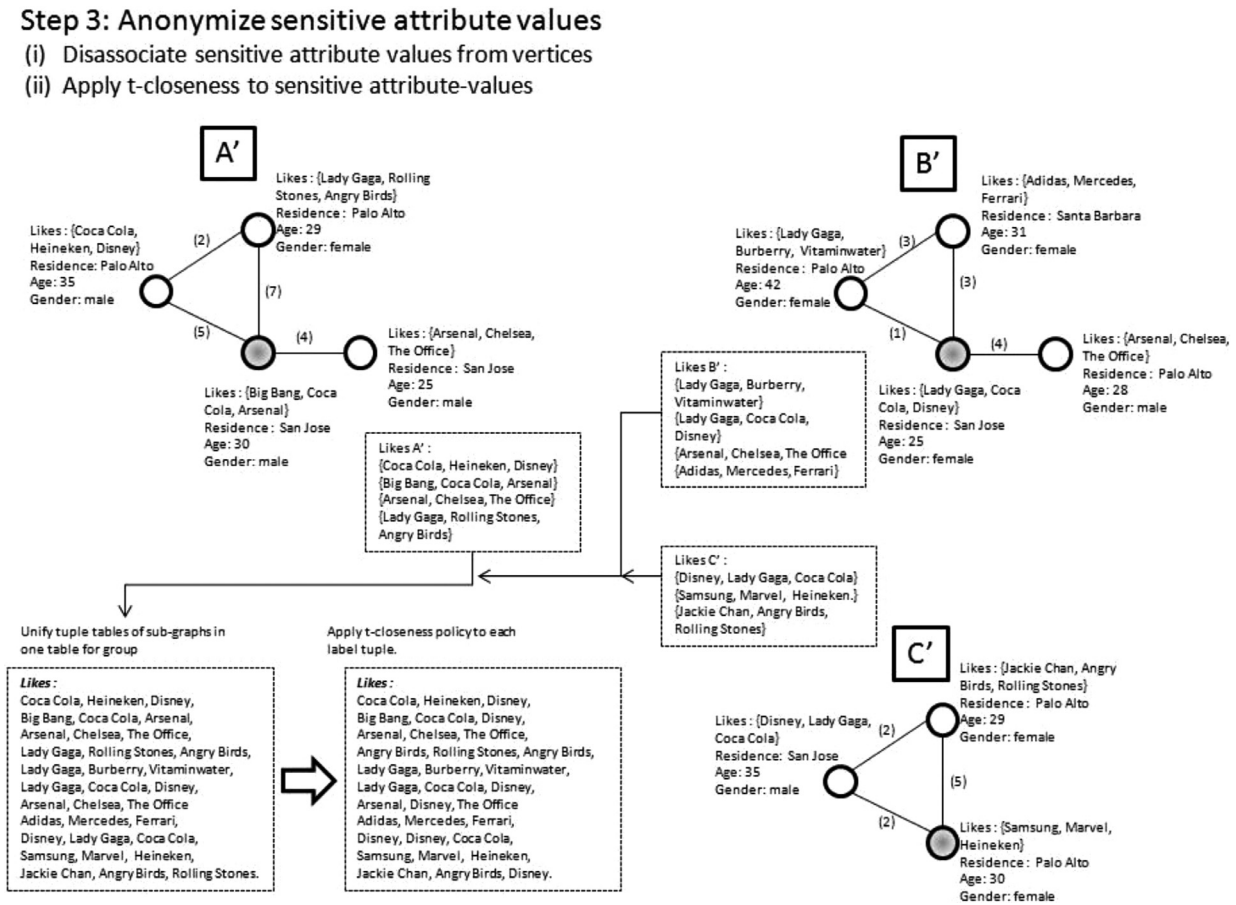

Fig. 7. Sensitive attribute-values are processed: (i) the sensitive attribute-values are disassociated from the vertices by placing them in a separate table; (ii) t-closeness is applied to table of sensitive attribute values.

5.1.4. Topological and quasi-identifier anonymity

Following on from the previous sections, for the example *k*-group shown in Fig. 7, we now identify the sub-graph which is the closest to all other sub-graphs in that group. That is, we choose as prototype P_i the sub-graph in a *k*-group whose sum of distances to all other sub-graphs is minimal. This results in a lower information loss because fewer changes will be necessary to each sub-graph because their closeness to the prototype will be minimal. For the matching we use the same matcher used previously to identify the *k*-1 sub-graphs most similar to the seed sub-graph.

Next, we modify all other sub-graphs in the *k*-group to make the topology and the quasi-identifiers the same as P_i. This is done by simply substituting each sub-graph by the prototype. We note that the external links of each sub-graph to other sub-graphs and vertices which are not in sub-graphs is not considered, because we will publish only the sub-graphs (which corresponds to the original graph sample). See Section 5.1.5 for details of what is actually published as output of the anonymization process.

At this point we will have obtained topological and quasi-identifier k-anonymity. This process is depicted for a given *k*-group in Fig. 8.

At the end of the process we will have obtained the anonymity of the three major characteristics: topological, quasi-identifiers and sensitive attributes.

Procedure anonymizer

Input: graph G
Output: anonymized graph G'
1. Order seeds *s* in descending degree size
2. **For each** $s_i \in S$ **do**
3. **Find** set of k seeds $\{s_{ij}\}$ in S such that the corresponding subgraphs H_{ij} are most similar to H_i, $s_{ij} \neq s_i$
4. **Identify** prototype P_i which is closest to all other sub-graphs H_{ij} with seed $\{s_{ij}\}$
5. **Make** topologies of sub-graphs in H_{ij} equal to P_i
6. **Make** quasi-identifiers of sub-graphs H_{ij} equal to those of P_i
7. **Assign** the sensitive-attributes of the vertices to a table whose records are disassociated from said vertices
8. **Alter** sensitive attributes of sub-graphs in S_i so that the *t*-closeness of their distributions with respect to their corresponding distributions of the complete graph is equal to 0.2.
9. **End do**

5.1.5. Published graph data

Once we have the anonymized graph sample, we publish the data. The published data consists of a set of *k*-groups. Each *k*-group consists of one prototype and a table of sensitive attributes which complies with *t*-closeness. Each prototype will be a sub-graph with vertices and edges. Each vertex will have associated with it three quasi-identifiers: residence, age and gender. Each edge will have associated with it one quasi-identifier: number of interactions. Each table of sensitive attributes will consist of a set of likes which have been disassociated from the original vertices, but which represent the original likes of each vertex, having been altered to comply with t-closeness.

The sample extraction method for the sub-graphs is designed to give a good coverage of the complete graph both topologically and in terms of data distributions. Any given vertex, not included in the sample, will be at a distance of at most 2 from a seed and it is reasonable to assume that this proximity will give the vertex similar attributes to the seed and its neighbors. Hence, effectively, the sample is giving a representative vision of the complete graph, while guaranteeing a given privacy level of *k*.

5.2. Pseudo-code of overall data processing scheme

The main data processing procedure has four main steps: "synthetic data generation" in which the graph is created and the data is assigned; "pre-calculations" in which the sub-graph statistics are calculated which are later used by the distance metric; "Train", which calibrates the distance metric matcher function; and "Anonymize", which anonymized the graph *G* with privacy parameters of *k* and *t*.

Main procedure

Input: original graph G=(V, E), anonymization level *k*, *t-closeness threshold t*
Output: anonymized graph G'
1. *Synthetic graph and data generator*
2. **Call** *SyntheticDataGenerator(|V|,|E|)*
3. *Pre-calculate*
4. **Calculate** topological statistics for each sub-graph G_1 … G_n
5. *Train*
6. **Calibrate** matcher for data
7. *Anonymizer*
8. **Call** Anonymizer(G, k, t)

6. Empirical testing and results

In this Section, we first present the details for the synthetic data generation step. Then we present the results of the data anonymization step in terms of information loss versus privacy level, which is the typical empirical benchmark approach. The information loss is represented as different components of the anonymization 'cost' (see Section 4, Eqs. 1 and 2) and the privacy level is represented by different values of 'k' (see Section 4.2) for different the comparison methods (See Section 6.2.1).

Experimental setup hardware/software. The hardware used was a desktop PC with a 64 bit Intel i5 processor @2.3Ghz, quad core and 8Gb RAM; the operating system was Windows 10; for software development and execution Eclipse Mars release and Java 7 was used. With respect to implementation details, the data generator and the anonymizer are defined in the same Eclipse project and share the same data structures for loading and processing the graph data files, hence they are integrated. The programs are structured such that they could be used in the future as an 'api' library 'back-end' for a user interface 'front end'.

6.1. Data generation

In the following section we describe the synthetic data processing for the datasets used later in the anonymization experiments of Section 6.2.

6.1.1. Test datasets

We have used the synthetic OSN graph topologies generated by 'RMat' (Chakrabarti et al., 2004) as starting point and have then added additional node and edge data (the quasi-identifier attribute values and sensitive attribute values), as described in Section 4 of the paper. This allows us to easily evaluate and compare the resulting graph with respect to the original graph, for the information loss measures defined in Section 4.1.

We have generated the following datasets: 1 K vertices, 30 K edges; 10 K vertices, 300 K edges; 100 K vertices, 3 M edges. The datasets are designed to test the scalability of the solution.

6.1.2. Synthetic data generation and OSN data example

We use the following process to create synthetic social network graphs with quasi-identifiers and sensitive attributes. First we use the 'Rmat' algorithm to generate the unlabeled graph topology. We recall that 'Rmat' creates topologies in which the graph metrics approximate the characteristics of an online social network graph. Secondly, we identify some key structures which will be used to

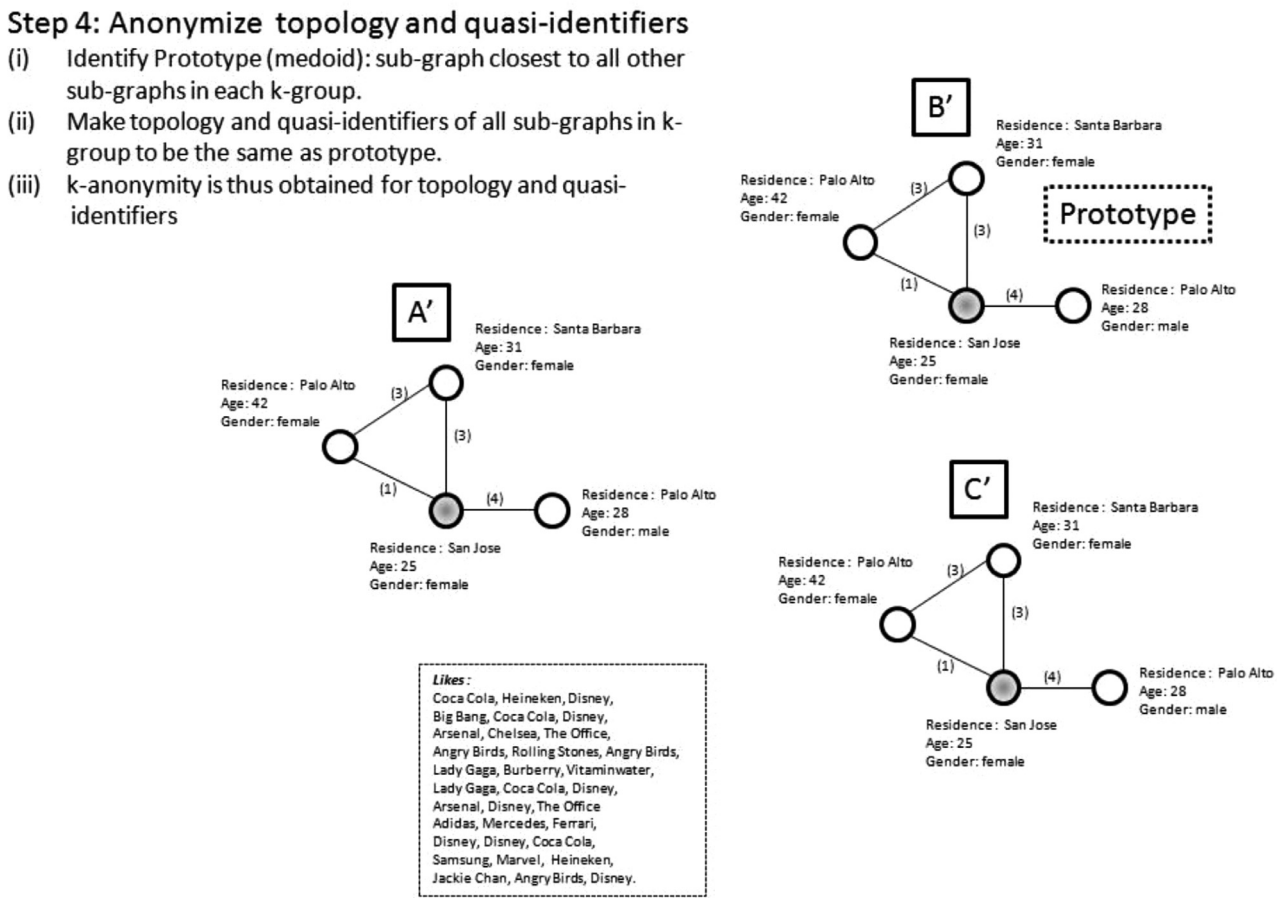

Fig. 8. For each *k*-group of sub-graphs, the topology and quasi-attributes for each sub-graph are made the same as the prototype sub-graph.

assign the data: (a) communities, (b) authority nodes and (c) significant sub-graphs with a high connectivity density. These structures are chosen for two reasons: (i) user profile similarity; (ii) information propagation capacity.

To identify the communities, we run Louvain's community detection algorithm to assign a community label to each vertex. To identify the authority nodes, we run the HITS algorithm, and to identify the significant sub-graphs with high connectivity density, we calculate the clustering coefficient of each neighborhood sub-graph and then find the sub-graph which maximizes the clustering coefficient and the number of vertices (i.e. maximum connectivity for the maximum number of vertices). Once we have the communities, authority vertices and dense sub-graphs, we can assign the data.

In Fig. 9 we see a simplified graphical representation of the resulting OSN structure.

We consider the immediate neighborhood of each authority vertex as a sub-graph. Also, the maximum diameter of the dense sub-graphs must be equal to 3 (i.e. a reference node and its immediate neighbors).

The next step is to assign the quasi-identifier and sensitive attribute data to the vertices and edges to the authority vertices and dense sub-graphs of each community.

This is done based on information propagation theory (which vertices will influence other vertices), specific rules which define which data combinations to assign, and a random factor which introduces a percentage of noise (initially set to 15%).

Also, we use rules to make the quasi-identifier data and sensitive attributes mutually coherent and realistic.

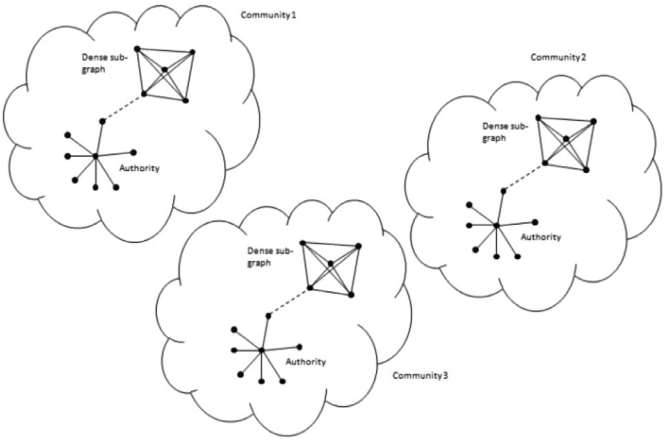

Fig. 9. Example structure identification.

Examples of general rules/guidelines:

Rule 1: neighbors and close knit groups will have a similar geographical residence.
Rule 2: the age of OSN users tends to young
Rule 3: the likes are taken from the current list of the 100 most liked pages in Facebook.
Rule 4: the gender is influenced by the likes. For example, females will have more likes for female pop artists and soft

Table 3
Default attribute-values and their target distribution proportions (based on US demographic Census, 2010).

Attribute	Values
Age	"18–25" (25%), "26–35" (25%), "36–45" (16.67%), "46–55" (8.33%), "56–65" (8.33%), "66–75" (8.33%), "76–85" (8.33%)
Gender	male (50%), female (50%)
Residence	"Palo Alto" (17%), "Santa Barbara" (16%), "Boca Raton" (16%), "Boston" (17%), "Norfolk" (17%), "San Jose" (17%)
{like1, like2, like3}	Taken from top 47 most liked pages in Facebook (see text), then generalized into categories as described in Section 4.1.1. Patterns: {"entertainment", "entertainment", "music artist"} (25%), {"music artist", "music artist", "entertainment"} (25%), {"drink brand", "drink brand", "entertainment"} (25%), {"tv show", "drink brand", "soccer club"} (25%).

drinks, and males will have more likes for a football clubs and beer drinks.

Rule 5: the age is influenced by the likes. For example, younger people will prefer contemporary pop artists and mobile apps.

A weight is randomly assigned to each edge (which represents the number of interactions between the corresponding vertices), three quasi-identifiers (residence, age, gender) to each vertex and one sensitive attribute per vertex. The sensitive attribute is composed of three sub-attributes (top 3 likes). The edge weight is also considered as a quasi-identifier.

The attribute-values are assigned in a pseudo-random fashion to each seed vertex. Then, the attribute-values of the immediate neighbors of a key vertex are assigned so that they have a predetermined similarity (within given limits and with a small random aspect) to the attribute values of their corresponding seed. In this way we obtain an approximation of the nature of a real social network in which a member of an OSN and his/her neighbors tend to bear a similarity between them.

Data values: *Age* is assigned a numerical value between 18 and 85, with a weighted distribution skewed on younger individuals. *Gender* can be "male" or "female", with equal probability. Residence can be "Palo Alto", "Santa Barbara", "Boca Raton", "Boston", "Norfolk" and "San Jose", with equal probability, which are all locations in the United States. The edge weight (interaction activity) is assigned a value between 1 and 10, with equal probability.

The *likes* are taken from the top 47 most liked pages in Facebook, weighted by their normalized number of fans. For example, the most liked page is 'Disney' with frequency weight 100, followed by 'Lady Gaga' with 50, 'Coca Cola' with 43, 'Jackie Chan' with 25, 'Angry Birds' with 20, and so on. Thus the assignment probability of 'Disney' is equal to the sum of all the frequency weights (416) divided by the frequency weight of 'Disney' (100). We note that each vertex is assigned a tuple of three likes, which are influenced by the quasi-attributes of age and gender, and the mix of likes *per se* (from the list of 47 possible likes). The likes are categorized into five categories: "entertainment", "music artist", "tv show", "soccer club", "drink brand", as described previously in Section 4.1.1.

In Table 3, we see the default attribute-values and their distribution proportions assigned for the empirical testing of Section 6.2. As mentioned previously the user can customize the proportions, the attribute-values and the attributes themselves. In the latter two cases it would be necessary to adapt the matching and propagation rules. The target (optimum) values of Table 3 are then used to measure the fitness of the data assignment, which is adjusted to give a best approximation to the target values.

In Figs. 6–8 we can see examples of the data assignments to the quasi-identifiers and sensitive attributes.

6.2. Results: information loss versus privacy level

In the following section we present the results for the data anonymization process, for different approaches, datasets and privacy levels.

6.2.1. Experimental setup – comparison methods

The following experiments compare our method (designated in the following as "NoInt" or "No Intersecting") with two other approaches, in order to demonstrate the relative improvement.

(i) Intersecting local neighborhoods (designated as "Int") which uses seeds which allow overlapping neighborhoods by allowing an inter-seed distance equal to 2.

(ii) No seeds (designated as "NoS"). This second comparison method does not use seeds at all and performs global topological modification. It is based on that described in Zhou and Pei (2011) and Zhou and Pei (2008) (see Section 2.2 of the current paper). We have adapted the method to use t-closeness instead of ℓ-diversity to process the sensitive attributes. Also, the matching is performed by our method described in the present paper, instead of the coding method described by Zhou and Pei. This is justified because in the current paper we are interested in validating the seed assignment method which guarantees non-overlapping local neighborhoods. We do allow the method, following Zhou's original definition, to perform the sub-graph modification in order to obtain isomorphisms, by adding nodes selected in increasing degree order from the rest of the graph.

We recall that a key aspect and novelty of our approach is to anonymize local neighborhoods, rather than individual nodes (the approach of "no seeds") during anonymization, thus taking into account the "social information/context" related to a node (i.e. its immediate neighbors), which it is proposed is an important aspect of online social network data. We also recall that our method employs seeds with non-overlapping neighborhoods (inter-seed distance ≥ 3).

Hence, the experiments are designed to demonstrate that an approach which tries to preserve the local neighborhood of a node has a lower information loss than methods which do not attempt this (such as "NoS"), or do so to a partial extent, as exemplified by "Int".

The information loss is measured using a cost function. Four different cost values are shown: 'ALL' signifies the weighted overall cost (incorporating the three information loss characteristics, TOP, QUASI and SENSI); 'TOP' represents the cost (in terms of information loss) of anonymizing the topology; 'QUASI' represents the cost of anonymizing the quasi-identifiers; 'SENSI' represents the cost of

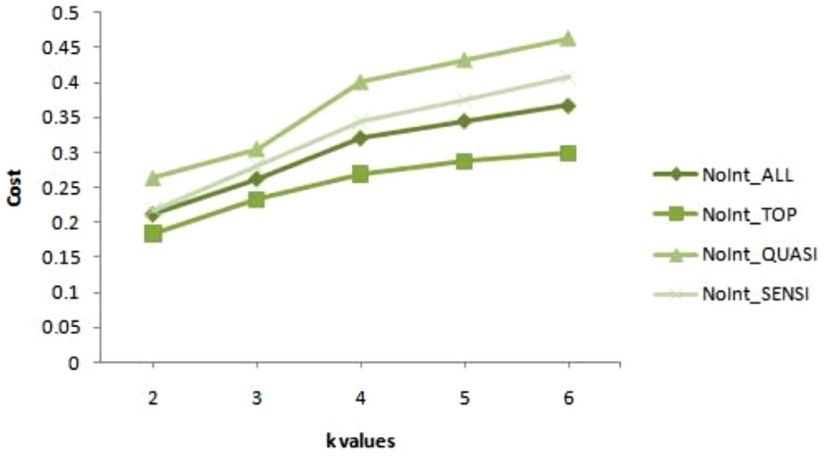

Fig. 10. Graph {1 K, 30 K}: Cost vs. privacy level k for the NoInt approach.

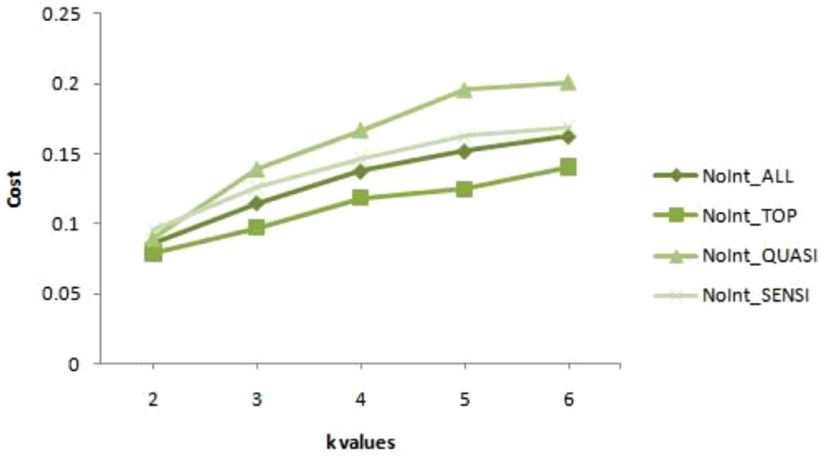

Fig. 11. Graph {10 K, 300 K}: Cost vs. privacy level k for the NoInt approach.

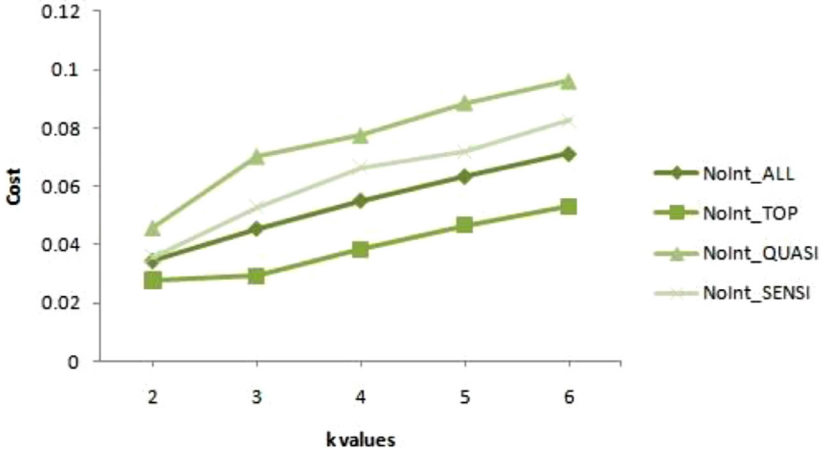

Fig. 12. Graph {100 K, 3 M}: Cost vs. privacy level k for the NoInt approach.

anonymizing the sensitive attributes. These components and the equations have been explained previously in Section 4.1.

The privacy level has been tested for the following levels of anonymity: $k = 2$, 3, 4, 5 and 6 for all datasets and methods. Three graph dataset sizes have been used: 1 K vertices, 30 K edges; 10 K vertices, 300 K edges; 100 K vertices, 3 M edges, which are designed to test the scalability of the solution. The datasets used for testing have been described in Section 6.1.

6.2.2. Results

The results of our method (NoInt) are shown in Figs. 10–12; the results of the second method (Int) are shown in Fig. 13; the results of the third method (NoS) are shown in Fig. 14; finally, in Fig. 15, the overall information loss (cost) 'ALL' is superimposed for each of the three methods to facilitate comparison. We observe that in each plot we use the same y-axis scale to compare the four metrics TOP, QUASI, SENSI and ALL. This is possible because the first three

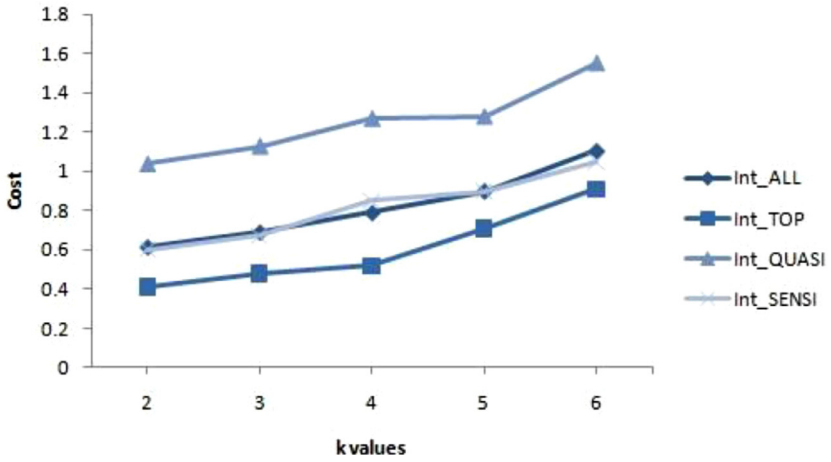

Fig. 13. Graph {1 K, 30 K}: Cost vs. privacy level k for the Int approach.

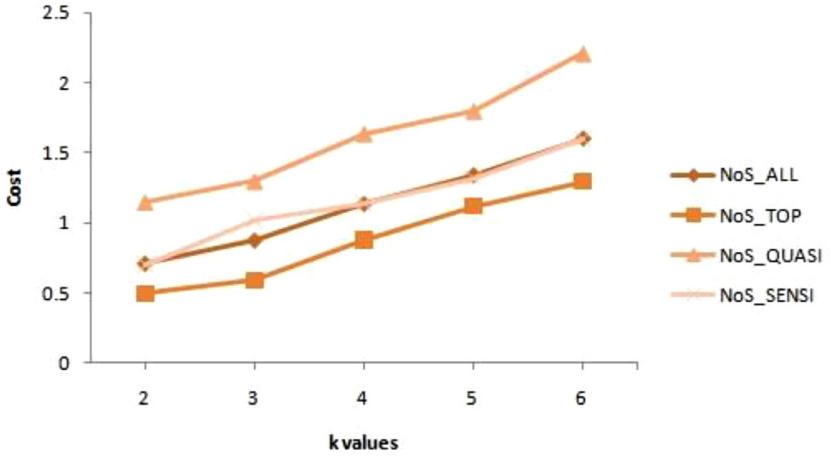

Fig. 14. Graph {1 K, 30 K}: Cost vs. privacy level k for the NoS approach.

metrics are normalized onto the same scale. As detailed in Section 4.1.1, 'ALL' represents the weighted sum of the other three metrics.

In Fig. 10 we see the results for the smallest graph (1 K vertices, 30 K edges) processed using the NoInt approach. On the x-axis are the different k (or privacy values) and on the y-axis is the cost as calculated by the equations described in Section 4.1. We observe in Fig. 10 that in general a similar trend is followed by all measures for this graph dataset. The lowest cost is registered by 'TOP', followed by 'SENSI'. 'QUASI' shows the relatively highest cost of the three measures. Overall we see a linear increase of cost with respect to the privacy level (k). For $k = 2$, we can read off the graph that ALL = 0.21, TOP = 0.18, QUASI = 0.26 and SENSI = 0.22. This agrees with Eq. (3), in which ALL was defined as TOP□0.50 + QUASI□0.25 + SENSI□0.25.

In Fig. 11 we see the results for next graph in order of size (10 K vertices, 300 K edges) processed using the NoInt approach. The axes and legend have the same format as for Fig. 10. We observe a lower cost for all three measures: 'TOP', 'QUASI' and 'SENSI'. With reference to the lower cost for the larger graph size, we note that, due to the larger graph and greater number of nodes, we are able to place a greater number of seed nodes. Hence, the matching algorithm will have a greater probability of finding good partitions for the k-groups. We note that 'TOP', the topological cost is again relatively lower than that of the attribute-value cost of 'SENSI' and 'QUASI'.

In Fig. 12 we see the results for next graph in order of size (100 K vertices, 3 M edges) processed using the NoInt approach. We note a similar scenario to that of Fig. 11: all measures have shown a further reduction in cost, with respect to the previous two graph sizes, and 'TOP' shows a lower cost than the attribute-value measures ('QUASI' and 'SENSI'). Again, we propose that the lower information loss of the larger graph is a consequence of higher quality partitioning made possible by a greater diversity and number of sub-graphs, topologies and attribute-values.

In Fig. 13 we see the results for the smallest graph (1 K vertices, 30 K edges) and for the second approach (Int) with inter-seed distance equal to 2. We recall that in order to avoid overlap, our method (NoInt) sets the inter-seed distance to three (whose results are shown in Figs. 10–12). If the inter-seed distance is set to two, neighbors can be in multiple sub-graphs and therefore k-groups. If we compare Figs. 10–13, we clearly see the greater cost (note y-axis scale) for all three measures when the inter-seed distance is set to 2. The relative performance of the cost measures is maintained: 'QUASI' suffering the greatest cost increase and 'TOP' the least.

In Fig. 14 we see the results for the smallest graph (1k vertices, 30 k edges) and for the third method (NoS) which performs graph anonymization without seeds. As mentioned previously, this method selects nodes in descending degree order, finding the k-1 most similar nodes for each (without restriction) and then adds nodes (taken from the entire graph) in ascending degree order to obtain k-anonymity. If we compare Fig. 14 with Figs. 10 and 13 we see a greater cost for all metrics.

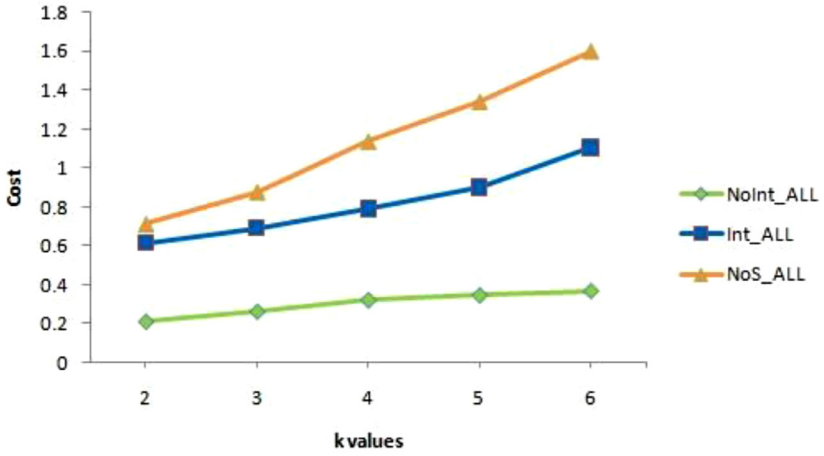

Fig. 15. Graph {1 K, 30 K}: Overall cost "ALL" for three approaches NoInt, Int and NoS.

In Fig. 15 we see the superimposed results for the three different methods and for the smallest graph (1k vertices, 30k edges). In each case the 'ALL' value is used which is a weighted composite of the three information loss metrics, 'SENSI', 'QUASI' and 'TOP'. In Fig. 15 we can clearly see that our method 'NoInt' gives significantly lower cost, in the terms described, than methods 'Int' and 'NoS' which do not protect local neighborhood structure when anonymizing.

As a '**managerial insight'**, it is evident from the results that the use of seeds and non-overlapping neighborhoods ('NoInt', Fig. 10) gives significantly better results than overlapping seed neighborhoods ('Int', Fig. 13) or the method which does not consider the local neighborhood when anonymizing ('NoS', Fig. 14).

In terms of scalability, we see from Figs. 10–12 that the information loss actually reduces for larger graphs, due to superior partitioning of the k-groups and the greater pool of candidate subgraph topologies (and their attribute-values) for matching, all of which reduces overall information loss.

7. Summary and conclusions

A system has been implemented which combines a synthetic graph social network data generator with a strict anonymization method which uses an efficient local sub-graph matcher to mitigate the information loss (anonymization cost) and optimize the data utility for local neighborhood social network structures. By designing a system which works well with default control parameters, an end user (data analyst) without expert knowledge (about data anonymization) or an expert user who wishes to automate part of the calibration and analysis, will be able to use it to generate a rich social network dataset, which can be customized for specific requirements, and then be anonymized with different privacy guarantees, again depending on specific end-user requirements. To the best of our knowledge, there are no systems currently available which provide this combination of functionality.

A **key 'managerial insight'** demonstrated in Section 6.2 (Fig. 15) is that our method 'NoInt', based on non-overlapping local neighborhoods, gives a lower anonymization cost with respect to the methods 'Int' and 'NoS' (exemplified by Zhou and Pei (2008, 2011)) which do not take the full neighborhood information into account. This was stated initially as one of the objectives of the current work. In Fig. 15 it is clearly seen that with respect to increasing privacy levels (as represented by 'k') 'NoInt_ALL' has a much lower anonymization cost (which represents the perturbation between the original social network and the anonymized one), whereas 'Int_ALL' has a cost three times greater and 'NoS_ALL' having an even higher cost. In Figs. 10–12 we have demonstrated the scalability of our approach 'NoInt' for which the cost (n.b. y-axis scale) actually goes down for successively larger graphs. This is explained by the greater diversity of candidate partners for matching when forming the k-groups, resulting in a closer match and therefore less perturbation when the k-group members are substituted by the medoid.

The synthetic data generator has been found to provide realistic data for experimental purposes whose statistical properties approximate a set of user defined profiles and attribute distributions. A strong overall privacy guarantee is obtained by applying k-anonymity to the quasi-attribute and topological, and the t-closeness calculation applied to the sensitive attribute distributions within the k-groups.

The research contributions in terms of expert systems and intelligent systems, as we outlined in the introduction are the creation of an integrated systems for a data analyst which allows this user to experiment in anonymizing realistic synthetic social network data. The user can then either use this data directly or apply the calibrated system to a real dataset. The system incorporated different intelligent components, such as the propagation algorithm, rule set definitions for the data profiles and the optimization of the anonymization of the sensitive attributes using a temperature optimization process (simulated annealing) to find the minimum perturbation which obtains the threshold $t = 0.20$.

A key contribution of our 'NoInt' method is that it anonymizes local neighborhoods as a complete "unit" in the online social network data, thus maintaining the social and contextual information it contains. This contrasts with the state of the art in which, although matching and adversary information may be considered in terms of local neighborhoods, the anonymization process itself tends to consider nodes in an individual manner.

The synthetic data set is comprised of many attributes and values, with a diversity of inter-relations. Typically, researchers in the data anonymization field tend to use simple quasi-identifiers and sensitive attributes which are created as theoretical examples. We generate a dataset which more closely approximates what is found in real online social networks; and we will make source code and datasets available online to the research community, providing a level of dataset complexity which is currently unavailable for testing.

In terms of data anonymization, we utilize a combination of techniques which represent the current frontier for anonymization in order to obtain a stricter and stronger privacy

level: k-anonymity for the topology and quasi-identifiers and t-closeness for the sensitive attributes. This represents a major challenge for the complex data set which is used for testing. In the privacy preserving data anonymization field, t-closeness is an approach which represents one of the strictest privacy guarantees with respect to k-anonymity and even ℓ-diversity. However, there are few empirical implementations and to the best of our knowledge none for a rich graph structured dataset.

With respect to expert and intelligent systems, an expert system embodies a set of rules and heuristics which represent human and machine expert knowledge. In the current work different rule sets are defined to control (i) the synthetic data generation and (ii) the anonymization process. In the first case, the rules can be customized by the end user (data analyst) depending on the data s/he wishes to generate (see Section 6.1.2) and how the graph is populated and the data disseminated. Heuristics are used for creating the initial topology, populating it with seed vertices, assigning profiles to seed vertices and disseminating the profiles (prototypes) to neighbor nodes using a user defined distance (matching) function heuristic. In the second case (anonymization process), the user defines the quasi-identifiers and sensitive attributes, together with the parameters for the k-anonymity and t-closeness heuristics which perform the anonymization. Hence the expert knowledge is embodied throughout the system using rule sets and heuristic algorithms. Also, the distance metric has three user defined weights which can give different levels of importance to the topology, the quasi-identifiers and the sensitive attributes, respectively. They can be assigned by trial and error, or by a stochastic method (such as simulated annealing) to find the optimum value for the information loss cost function (see Section 4.1).

Strengths and weaknesses, advantages and limitations: Given that the anonymization follows a clustering approach (see Section 2), in order to navigate the whole graph it is necessary to publish the links associated with each k-group, connecting it to every other k-group. This is not difficult to implement as during processing the link look-up table from the original graph is maintained. However, it is out of the scope of the current work and is only necessary when the user wishes to study, for example, 'friend of friend' type relations, spanning outside the local neighborhoods. In the present work we have assumed a user who is interested in studying the local neighborhoods of users and comparing them, which is a common focus in online social network analysis. Approaches which are based on node modification (see Section 2) are in principal easier to interpret in terms of overall graph connectivity, however they have other drawbacks such as greater information loss at a local level (see Section 2). The making available of information to the user of links between the k-groups is commented below as future work.

Also, the proposed method anonymizes a representative part of the network (based on the seeds), reducing the modifications needed to anonymize the network to be published, hence it yields a better approximation to the original. All this at the expense of possibly excluding some individuals and relations. In this sense, approaches such as Zhou and Pei (2011) offer a possible advantage when the user of the system (the data analyst) is interested in studying overall graph properties, rather than local neighborhood properties. The advantage of not taking all the nodes in the network is that there are less constraints to fulfill, hence decreasing the modifications needed on the remaining part of the network to achieve privacy definitions (such as ℓ-diversity or k-anonymity). This may be a limitation since not all individuals will be represented explicitly in the network, on the other hand, by the definition of our method a close acquaintance of any node is represented in the published information.

As **future work and research directions** we can consider the definition overlapping communities in the social networks, which represents an additional challenge for data generation and anonymization. Also, we can include link information between the k-groups, as commented above when we discussed the strengths and weaknesses. Another challenge is to automatically optimize the synthetic data against different real datasets. Currently the user defines data profiles which may be taken from the statistics of an existing dataset. Also, we can develop a graphic user interface (GUI) to facilitate user interaction with the system, and evaluate the privacy risk using a set of adversary queries. The GUI could be similar in design to the spreadsheet approach of Roby et al. (2012).We note that the scope of the current work is limited to evaluation of information loss (cost), and the privacy guarantee is assumed from the k-anonymity and t-closeness parameters.

Finally, in the short term we plan to place online examples of the synthetic datasets, and the source code for the system, thus making it available to the data analysis and data privacy research communities.

Acknowledgments

This research is partially supported by the Spanish MEC (projects ARES CONSOLIDER INGENIO 2010 CSD2007-00004 – eAEGIS TSI2007-65406-C03-02 – HIPERGRAPH TIN2009-14560-C03-01 and ICWT (TIN2012-32757)).

The authors are grateful to Dr. David Megías of the Universitat Oberta de Catalunya for his support.

Appendix 1. –Pseudo code of synthetic data generator

Procedure synthetic data generator

Input: Number of vertices and edges, p = degree of data diversity
Output: graph G

1. *RMat*
2. **For** |V| vertices and |E| edges **generate** an OSN-like topology.
3. *Communities*
4. **Calculate** communities using Leuven method and assign community tag to each vertex.
5. *Authorities and dense sub-graphs*
6. **Calculate** top authorities A_c and dense sub-graphs D_c, A_c, $D_c - AD_c$ in each community c using HITS algorithm and clustering coefficient metric
7. Each vertex ad_c must be at a distance ≥ 3 from any other authority or dense sub-graph in community c
8. *Assign data to AD's in each community*
9. **For each** community c do
10. **For each** vertex $ad_c \in AD_c$ **do**
11. **For each** edge e connected to ad_c, assign a random weight between 0 and 10.
12. **Assign** other quasi-identifiers to ad_c
13. **Assign** sensitive attribute-values to ad_c
14. *Assign quasi-identifiers and sensitive attribute-values to neighbors of* ad_c
15. Let Nad_c be the set of neighbors of ad_c
16. **For each** $n \in Nad_c$ **do**
17. **For each** attribute a of n **do**
18. **For each** value v of attribute n **do**
19. **Probability of assignment of {n, a, v} = p**
20. Assign {a, v} of ad_c to neighbor n with probability 1- p
21. **End do**
22. **End do**
23. **End do**
24. **Let** NA_c be the set of vertices in c with data assigned
25. *Assign data to remaining nodes in each community*
26. **Call**
Assign Unassigned Vertices in Community(NA_c, c)
27. **End do** // for each community

Procedure assign unassigned vertices in community

Input: NA_c, the set of vertices in c with data assigned; c, the current community id
Output: assigned vertices in community c

```
For each n in c ∈ NA_c do
    For each attribute a of n do
        For each value v of attribute n do
            Calculate average or modal value of
            corresponding attribute-value of neighbors
            of n as {n′, a′, v′}
        Probability of assignment of {n', a', v'}=p
        Assign {a′, v′} to n with probability 1- p
            Otherwise
                Assign random neighbor {a′′, v′′} to n if at
                least one non-null value
            Otherwise
                Assign random values {a′′′, v′′′}
        End do
    End do
```

References

Ali, A. M., Alvari, H., Hajibagheri, A., Lakkaraj, K., & Sukthankar, G. (2014). Synthetic Generators for Cloning Social Network Data. In *Proceedings of Social Informatics (SocInfo) Workshop*, Barcelona, 2014.

Barrett, C. L., Beckman, R. J., Khan, M., Kumar, V. S. A., Marathe, M. V., Stretz, P. E., et al. (2009). Generation and analysis of large synthetic social contact networks. In *Proceedings of the 2009 Winter Simulation Conference, 13-16 Dec. 2009* (pp. 1003–1014).

Bhagat, S., Cormode, G., Krishnamurthy, B., & Srivastava, D. (2009). Class-based graph anonymization for social network data, VLDB '09, August 24-28, 2009. *Proceedings of the VLDB Endowment, 2*(1), 766–777.

Boncz, P., Perez, M., Gavalda, R., Angles, R., Erling, O., Gubichev, A., Spasić, M., Pham, M. D., & Martínez, N. (2014). Benchmark design for navigational pattern matching benchmarking. In LDBC Cooperative Project FP7 – 317548. Coordinators: Arnau Prat, Alex Averbuch. Issue 3 28/09/2014.

Campbell, W. M., Dagli, C. K., & Weinstein, C. J. (2013). Social network analysis with content and graphs. *Lincoln Laboratory Journal, 20*(1), 62–81.

Cao, J., & Karras, P. (2012). Publishing microdata with a robust privacy guarantee. In *Proceedings of the VLDB Endowment: 5* (pp. 1388–1399).

Casas-Roma, J., Herrera-Joancomartí, J., & Torra, V. (2013). An algorithm for k-degree anonymity on large networks. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

Chakrabarti, D., Zhan, Y., & Faloutsos, C. (2004). R-mat: A recursive model for graph mining. In *Proceedings SDM (Secure Data Management) 2004* (pp. 442–446).

Chester, S., Kapron, B., Srivastava, G., & Venkatesh, S. (2013). Complexity of social network anonymization. *Social Network Analisis and Mining, 3*(2), 151–166.

Chester, S., Kapron, B., Ramesh, G., Srivastava, G., Thomo, A., & Venkatesh, S. (2013). Why Waldo befriended the dummy? k-anonymization of social networks with pseudo-nodes. *Social Network Analisis and Mining, 3*(3), 381–399.

Clifton, C., & Tassa, T. (2013). On syntactic anonymity and differential privacy. *Transactions on Data Privacy, 6*, 161–183.

Das, S., Egecioglu, O., & El Abbadi, A. (2010). Anonymizing weighted social network graphs. In *Proceedings of 26th IEEE International Conference on Data Engineering, ICDE Conference* (pp. 904–907).

Das, D., & Bhattacharyya, D. K. (2012). Decomposition+: Improving '-diversity for multiple sensitive attributes. *Advances in computer science and information technology. Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, Vol. 85*(2012), 403–412.

De Capitani di Vimercati, S., Foresti, S., Livraga, G., & Samarati, P. (2012). Data privacy: Definitions and techniques. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20*(6), 793 Dec. 2012.

Gal, T. S., Chen, Z., & Gangopadhyay, A. (2008). A privacy protection model for patient data with multiple sensitive attributes. *International Journal of Information Security and Privacy, 2*(3), 28–44.

Hao, Y., Cao, H., Hu, C., Bhattarai, K., & Misra, S. (2014). K-anonymity for social networks containing rich structural and textual information. *Social Network Analysis and Mining (SNAM), 4*(223), 1–40.

Hay, M., Miklau, G., Jensen, D., Weis, P., & Srivastava, S. (2007). Anonymizing social networks, SCIENCE Technical Report 07-19, pp. 107–103, Vol. 245.

Hay, M., Miklau, G., Jensen, D., Towsley, D., & Weis, P. (2008). Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment (SESSION: Privacy and authentication), 1*(1), 102–114.

Hartung, S., Nichterlein, A., Niedermeier, R., & Suchý, O. (2015). A refined complexity analysis of degree anonymization in graphs. *Information and Computation, 243*, 249–262.

Heatherly, R., Kantarcioglu, M., & Thuraisingham, B. (2013). Preventing private information inference attacks on social networks. *IEEE Transactions on Knowledge and Data Engineering, 25*(8), 1849–1862 Aug. 2013.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of KDD '05, 11th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining: 2005* (pp. 177–187).

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and l-diversity. ICDE 2007. In *Proceedings of IEEE 23rd International Conference on Data Engineering 15-20 April 2007* (pp. 106–115).

Liu, X., & Yang, X. (2011). A generalization based approach for anonymizing weighted social network graphs. Web-age information management, LNCS 6897, pp. 118–130.

Loukides, G., & Shao, J. (2011). Preventing range disclosure in k-anonymised data. *Expert Systems with Applications, 38*(4), 4559–4574 April 2011.

Loukides, G., & Gkoulalas-Divanis, A. (2012). Utility-preserving transaction data anonymization with low information loss. *Expert Systems with Applications, 39*(10), 9764–9777 August 2012.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *Proceedings of IEEE 22nd Internatinal Conference on Data Engineering (ICDE)* (p. 24).

Maheshwarkar, N., Pathak, K., & Choudhari, N. S. (2012). K-anonymity model for multiple sensitive attributes. *IJCA Special Issue on Optimization and On-chip Communication "OOC", 51-56*(1), 51–56 February 2012.

Martin, A. J. (2016). Yahoo dumps 13.5 TB of users news interaction data for machine eating. The Register, 14 Jan 2016. Available at: http://www.theregister.co.uk/2016/01/14/yahoo_dumps_135tb_of_users_news_interaction_data_for_machine_eating/ (Last accessed 18/02/2016).

Nettleton, D. F., Sáez-Trumper, D., & Torra, V. (2011). A comparison of two different types of online social network from a data privacy perspective, Proc. MDAI. LNAI, Vol. 6820, pp. 223–234.

Nettleton, D. F. (2012). Information loss evaluation based on fuzzy and crisp clustering of graph statistics. In *Proceedings of WCCI 2012, World Congress on Computational Intelligence 2012* (pp. 1–8). FUZZ-IEEE.

Nettleton, D. F., & Dries, A. (2013). Local neighbourhood sub-graph matching method, European Patent application number: 13382308.8. (Priority 30/7/2013). PCT application number: PCT/ES2014/065505. (Priority 18/7/2014).

Nettleton, D. F., Torra, V., & Dries, A. (2014). A comparison of clustering and modification based graph anonymization methods with constraints. *International Journal of Computer Applications, 95*(20), 31–38.

Nettleton, D. F. (2015). Generating synthetic online social network graph data and topologies. In *Proceedings of the 3rd Workshop on Graph-based Technologies and Applications (Graph-TA)* March 18th 2015.

Pérez-Rosés, H., & Sebé, F. (2015). Synthetic generation of social network data with endorsements. *Journal of Simulation, 9*(4), 279–286.

Pérez-Rosés, H., Sebé, F., & Ribó, J. M. (2016). Endorsement Deduction and Ranking in Social Networks, Computer Communications, Vol. 73, Part B, 1 January 2016, pp. 200–210, Elsevier.

Pham, M. D., Boncz, P., & Erling, O. (2013). S3G2: A scalable structure-correlated social graph generator. *Selected Topics in Performance Evaluation and Benchmarking, LNCS, 7755*, 156–172.

Prasser, F., & Kohlmayer, F. (2015). Putting Statistical disclosure control into practice: The ARX data anonymizationtool. In Aris Gkoulalas-Divanis, & Grigorios Loukides (Eds.), *Medical data privacy handbook* (pp. 111–148). Springer November 2015. ISBN: 978-3-319-23632-2.

Roby, R. K., Phillips, N. R., Thomas, J. L., & Sprouse, M. L. (2012). Development of an expert system for automated forensic mitochondrial DNA data analysis. The University of North Texas Health Science Center (UNTHSC), Department of Forensic & Investigative Genetics, Document No.: 239675, Sept. 2012. Available at: https://www.ncjrs.gov/pdffiles1/nij/grants/239675.pdf (Last accessed, 29/01/2016).

Rubio, L., De la Sen, M., Longstaff, A. P., & Fletcher, S. (2013). Model-based expert system to automatically adapt linning forces in Pareto optimal multi-objective working points. *Expert systems with Applications, 40*(2013), 2312–2322.

Salas, J., & Torra, V. (2015). Graphic sequences, distances and k-degree anonymity. *Discrete Applied Mathematics, 188*, 25–31.

Salas, J., & Torra, V. (2016). Improving the characterization of P-stability for applications in network privacy. *Discrete Applied Mathematics*.

Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 1998), June 1-3, 1998* (pp. 188–203). WA, USA: Seattle.

Samarati, P. (2001). Protecting respondents' identities in microdata release. In *Proceedings of the IEEE Transactions on Knowledge and Data Engineering: vol. 13(6), November/December 2001* (pp. 1010–1027).

Skarkala, M. E., Maragoudakis, M., Gritzalis, S., Mitrou, L., Toivonen, H., & Moen, P. (2012). Privacy preservation by k-anonymization of weighted social networks. In *Proceedings of ASONAM* (pp. 423–428). IEEE Computer Society.

Song, Y., Karras, P., Xiao, Q., & Bressan, S. (2012). Sensitive label privacy protection on social network data. In *Proceedings of the 24th International Conference on Scientific and Statistical Database Management, SSDBM'12* (pp. 562–571).

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS), 10*(5), 557–570.

Truta, T., Campan, A., & Meyer, P. (2007). Generating microdata with p-sensitive k-anonymity property. In *Proceedings of SDM (Secure Data Management)* (pp. 124–141).

Yang, W., & Qiao, S. (2010). A novel anonymization algorithm: Privacy protection and knowledge preservation. *Expert Systems with Applications, 37*(1), 756–766 January 2010.

Yuan, M., Chen, L., Yu, P. S., & Yu, T. (2013). Protecting sensitive labels in social network data anonymization. *IEEE Transactions on Knowledge and Data Engineering, 25*(3), 633–647 March 2013.

Zhou, B., Pei, J., & Luk, W. S. (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter, 10*(2), 12–22 December 2008.

Zhou, B., & Pei, J. (2008). Preserving privacy in social networks against neighborhood attacks. In *Proceedings of 24th International Conference on Data Engineering (ICDE), 2008, IEEE* (pp. 506–515).

Zhou, B., & Pei, J. (2011). The k−anonymity and l−diversity approaches for privacy preservation in social networks against neighborhood Attacks. *Knowledge and Information Systems, 28*(1), 47–77.