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/

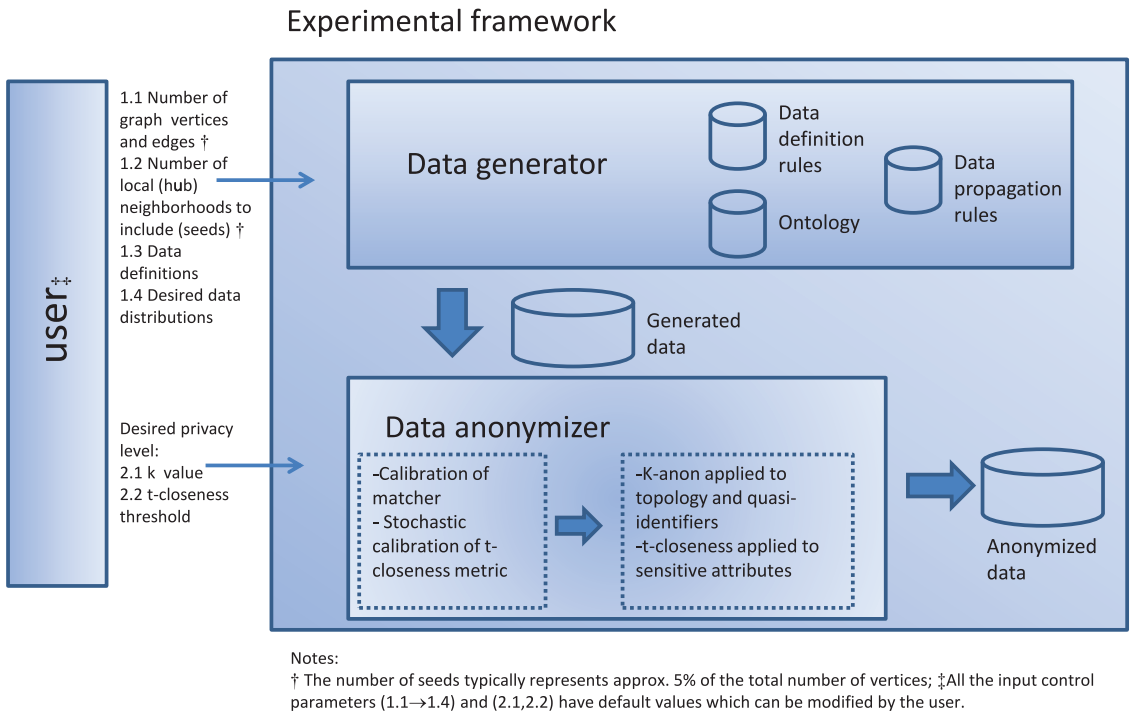

Fig. 1. Overall system architecture.

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