Journal of Applied Logic 12 (2014) 151–174

Contents lists available at ScienceDirect

Journal of Applied Logic

www.elsevier.com/locate/jal

A logical framework for privacy-preserving social network publication ☆

Tsan-sheng Hsu, Churn-Jung Liau *, Da-Wei Wang

Institute of Information Science, Academia Sinica, Taipei 115, Taiwan

ARTICLE INFO

Article history:
Received 4 August 2013
Accepted 11 December 2013
Available online 31 December 2013

Keywords:
Social network
Privacy
Description logic
Positional analysis
Information granule

ABSTRACT

Social network analysis is an important methodology in sociological research. Although social network data are valuable resources for data analysis, releasing the data to the public may cause an invasion of privacy. In this paper, we consider privacy preservation in the context of publishing social network data. To address privacy concerns, information about a social network can be released in two ways. Either the global structure of the network can be released in an anonymized way; or non-sensitive information about the actors in the network can be accessed via a query-answering process. However, an attacker could re-identify the actors in the network by combining information obtained in these two ways. The resulting privacy risk depends on the amount of detail in the released network structure and expressiveness of the admissible queries. In particular, different sets of admissible queries correspond to different types of attacks. In this paper, we propose a logical framework that can represent different attack models uniformly. Specifically, in the framework, individuals that satisfy the same subset of admissible queries are considered indiscernible by the attacker. By partitioning a social network into equivalence classes (i.e., information granules) based on the indiscernibility relation, we can generalize the privacy criteria developed for tabulated data to social network data. To exemplify the usability of the framework, we consider two instances of the framework, where the sets of admissible queries are the $\mathcal{ALCI}$ and $\mathcal{ALCQI}$ concept terms respectively; and we exploit social position analysis techniques to compute their indiscernibility relations. We also show how the framework can be extended to deal with the privacy-preserving publication of weighted social network data. The uniformity of the framework provides us with a common ground to compare existing attack models; while its generality could extend the scope of research to meet privacy concerns in the era of social semantic computing.

1. Introduction

Social network analysis (SNA) is widely used in the social and behavioral sciences, as well as in political science, economics, organizational science, and industrial engineering. The social network perspective,

☆ This is an extended version of [31].

* Corresponding author.
E-mail addresses: tshsu@iis.sinica.edu.tw (T.-s. Hsu), liaucj@iis.sinica.edu.tw (C.-J. Liau), wdw@iis.sinica.edu.tw (D.-W. Wang).

1570-8683/$ – see front matter
http://dx.doi.org/10.1016/j.jal.2013.12.001

which focuses on the relationships between social entities, has been developed over the last sixty years by researchers in psychology, sociology, and anthropology [56,66]. As a result, the paradigm is now gaining recognition and standing in the general social and behavioral science communities as a theoretical basis for examining social structures. The theoretical foundation has been clearly defined by many scholars, and the paradigm has been applied effectively to substantive problems. A social network can be visualized as a graph in which the nodes are the individuals in the network (or actors, in the terminology of SNA), and the links represent the relationships or flows between the individuals. SNA thus facilitates both visual analysis and mathematical analysis of complex human systems. Many computer programs have been designed and implemented to support SNA tasks, and the programs also enable social scientists to derive useful knowledge from social networks.

In the cloud computing era, the cloud servers in data centers are very powerful information sources that users can exploit to access information in different ways to meet their individual needs. Social network data stored in cloud servers can be accessed in at least two ways. On one hand, data analysts may want to know the global structure of the network; and on the other hand, general users may wish to query the data center about actors that satisfy particular properties for the purposes of advertising, seeking friends, or simply sharing information. In the context of privacy awareness, the revealed network structure should not contain any identifiers of the actors, and the answers to users' queries should not contain any sensitive information about the actors. To meet these requirements, the data center can only publish the structure of a social network after removing all personal identifiers of the actors and restricting users' queries to the non-sensitive part of the network. As a result, the publication of the anonymized network is secure because an attacker cannot infer the identity of any anonymized node in the published network structure. Furthermore, because of the restriction, the query-answering process cannot disclose any sensitive information about the actors.

Although the publication of anonymized networks and the answers to restricted queries seem secure individually, previous studies on tabulated data have shown that it is possible to re-identify the actors by linking these two types of information [53,60]. It has long been recognized that several quasi-identifiers (e.g., ZIP codes, age, and sex) can be used to re-identify individual records because the quasi-identifiers may appear with an individual's identifiers in another public database. Therefore, the problem is how to prevent adversaries inferring sensitive information about an individual by linking the released data set to public databases. An analogous situation may arise with social network data when both the network structure and non-sensitive information about the actors are released simultaneously, as the non-sensitive part of a social network may behave like quasi-identifiers in tabulated data. In this paper, we consider the issue of privacy preservation when the data center is required to publish the network structure and provide non-sensitive information about actors at the same time. The privacy risk depends on he amount of detail in the released network structure and the expressiveness of the allowed query language. Therefore, we propose a framework for evaluating the privacy risk of using a query language together with the publication of a particular network structure. To exemplify the usability of the framework, we consider two instances of query languages based on description logic (DL) formalisms [50], which are recognized by the semantic web community as appropriate for representing social network data [34].

Our methodology follows that of privacy-preserving tabulated data publication, where actors with the same combination of quasi-identifier values are grouped in a bin or an information granule. Some qualitative or quantitative safety criteria are then defined based on the distribution of the confidential attribute values of the actors in the same information granule [11,30,61]. The main difference between tabulated data and social network data is that, in a social network, two actors with same quasi-identifier values may still be distinguishable by their relationships with other actors. Thus, to formulate information granules for social networks, we have to consider the attributes of the actors as well as the relationships between the actors.

The formulation of information granules for social networks depends on the indiscernibility of actors according to the available information. In our framework, two actors are deemed indiscernible if they satisfy the same subset of admissible queries. To compute the indiscernibility of actors in a social network, we