Research Statement

WANG Yazhe

School of Information Systems, Singapore Management University

Tel: (65) 8269-5719, Email: yazhe.wang.2008@phdis.smu.edu.sg

Updated on 17 Feb 2012

Background

The social network websites and applications have been growing rapidly in the recent years. According to a statistical report published by Nielsen in 2010, three of the world's ten most popular brands online are social network related (i.e., Facebook, YouTube, and Wikipedia). The world now spends over 110 billion minutes on social networks and blog sites and this equates to 22 percent of all time online. The rapidly increasing coverage of their user bases and user engagement endow the social network media great potential to study the logic under the users social organization and behavior, and eventually to transfer this knowledge to real business profit. To this end, our current research focuses on two major areas. On the one hand, we want to establish profound understanding of the structures, dynamics and evolutions of these ever-changing networks social network data. On the other hand, develop privacy protection techniques which make the real-world social network data that contains highly sensitive user information available for the analysis tasks.

Research Areas

(1) Social network mining and analysis

Social network is different from physical networks in many aspects and is characterized by rich context information, human activity and interaction. For example, the twitter social network is not only contains the follower-followee relationship information between users, but also involves user profile information, user activity (e.g. tweeting) and user interaction (e.g. retweeting and replying). This richness of information poses great challenges as well as excellent opportunity to our research. We expect that the new knowledge gained from our study would lend valuable insights for user behavior and information dissemination in social network. In particular, we want to study the following key problems.

1.1 Search and navigation in large social networks

As the richness and large scale of social networks, it is hard for users to find particular information based on complex search conditions. This study will focus on identifying users' needs in searching and navigating large-scale social networks, as well as developing efficient data structure and algorithms for various application purpose.

1.2 Social influence study based on user interaction

Many studies on social influence have been largely focused on the network link structure, such as the various define of centrality measurements, PageRank and HITs algorithms. The interactions between users are usually ignored. In this work, we plan to analyze the twitter social network for users retweeting and replying behavior to identify key users who has the largest influence to other twitter users on specific topics or events.

(2) High utility social network privacy protection

One problem that prevents the free use of the social network data for analysis tasks is the concern of the disclosure of private information about individual users in the social network. This problem contains two fundamental aspects, *privacy* and *utility*. The former guarantees the sensitive information will not be inferred from the released data. As protecting privacy meant to modify the data before publishing, the latter is to make sure that the released data has the comparable level of accuracy for the mining and analyzing uses as the original data. My major contribution is addressing the utility issue in protecting *identity privacy* in the social network data.

Identity privacy is a major privacy concern in social network data publishing. It tries to prevent an attacker from discovering the true identities of the entities (i.e., the vertices of the network). It has been demonstrated that even after removing all the identifiable personal information (e.g., names, social security numbers, and identity card numbers), an attacker might still be able to identify an original entity in a published social network with high confidence based on the knowledge of the topological structure around the entity, such as degree, neighborhood and sub-graph. Consequently, further privacy protection strategies are purposed. These strategies are mainly divided into three categories.

1. Provide *k*-anonymity in the social network data via edge insertions and/or deletions;
2. Add noises to the social network to prevent attackers from identifying the targets;
3. Get the summary of the social network, and then publish the social network summary.

Our work is mainly on the first category. In order to achieve *k*-anonymity, we need to modify the social network by inserting and deleting edges. Most of the previous *k*-anonymization techniques employ the total number of modified edges to measure the social network utility loss, and want the utility loss to be as small as possible. However, we identify that this utility measurement is *not* effective as it assumes each edge modification has an equal impact on the original social network properties. In fact, due to the structural complexity of the social network, modifying a certain small number of edges may have large impact on the network properties, e.g. removing the edges between two loosely connected components could make a connected network unconnected. However, finding a proper utility measurement to control the utility loss during the anonymization process is not straightforward. Because, firstly, the social network has many aspects of its utility/property, it is hard to find one measurement which can capture most of, if not all, the social network properties. Secondly, the utility measurement needs to be quantitative and easy to calculate. Finally, the utility measurement needs to fit nicely with the anonymization process as we want to use it to direct modifying the social network to *k*-anonymity.

To address the above challenges, we proposed a new information theory based utility measurement, named *Hierarchical Community Entropy* (HCE). HCE is based on the hierarchical community structure of social networks, because that the community structure is a central organizing principle of social networks and it is a core graph topological feature which has a strong correlation with other important features. We use the existing binary tree structured *Hierarchical Random Graph* to model the social network. The HCE can be easily calculated according to the connection probabilities on each internal node on the binary tree model. Then, when we do anonymization, each edge modification will have a corresponding impact on the connection probability of one internal node, thus change the HCE value. The change of the HCE value is used to measure the utility loss. Based on this new utility loss measurement, we also develop a new *k*-anonyization framework, called *HRG based k-anonymization*. This framework anonymizes a given social network via edge modifications, meanwhile, makes the HCE utility loss as small as possible. The effectiveness and efficiency of our approach are verified by comprehensive experimental tests on real datasets.

Selected Publications

1. WANG Yazhe, ZHENG Baihua, "Context-Aware nearest neighbor query on social networks". *SocInfo*, 2011.
2. WANG Yazhe, XIE Long, ZHENG Baihua & Ken C. K. LEE. Utility-Oriented K-Anonymization on Social Networks. *DASFAA 2011*.