Privacy Preservation in Social Network Analysis

Olivera Grljević, Zita Bošnjak
Faculty of Economic Subotica
University of Novi Sad
Segedinski put 9-11, 24000 Subotica, Serbia
{oliverag,bzita}@ef.uns.ac.rs

Renata Mekovec
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
renata.mekovec@foi.hr

Abstract. *Social networks are in great expansion nowadays. Opposite to many benefits, such as fast access to information, simple interaction among users, or being a perspective data source for decision makers and analysts, they raise many problems related to privacy protection. Vast amount of personal information that are voluntarily provided is disclosed in social networks and prone to misuse. Therefore, in this paper we investigated named erosions of privacy from (1) the viewpoint of data mining techniques used in social network analytic, and from (2) the perspective of different privacy-preserving mechanisms and methods applied to social networks.*

Keywords. Social network analytics, privacy preservation

1 Introduction

Social networks have gained significant importance with the increase of proliferation and availability of devices with Internet access (like PCs, Internet tablets, smart phones, etc). This is evidenced by a remarkable success and expansion of social networks sites (applications) such as Facebook or Twitter. A social network can be defined as a network of interactions or relationships, where the nodes consist of actors and the edges consist of the relationships or interactions between these actors [1]. Therefore, these social networks' sites do not only provide information, but they enable interaction between the users (while the information is delivered). A particular user can interact with others by (1) bookmarking web sites and browsing through marked sites of other users, (2) voting for articles and making comments about them, (3) adding friends, (4) making comments about the content of their profiles, (5) joining groups and discussions, (6) exchanging photos and video content, and by (7) adding articles, editing and modifying the existing ones [5]. Taking all that into consideration it can be concluded that every site that invites users to interact can be defined as social media.

Besides their entertainment character, social networks provide rich content-based knowledge which can be exploited for development of effective social business strategies. Under the influence of the expansion of social network usage, not only the way of doing business has changed (there are numerous companies that build their marketing strategy around the available social networks) but the behavior of customers has changed too. Furthermore, by publishing content about a certain product or a brand a particular user can influence (1) behavior and opinion of their social network friends, or (2) their future willingness to purchase a particular product/service. Therefore, one of the main tasks of successful business is identification of "influential" users on social network. A chain reaction led by word-of-mouth marketing can be achieved by targeting these users, allowing a company to reach a great number of social network users with small marketing costs. By analyzing the spread of the impact over social networks one can acquire a better insight into the ways in which information is propagated and innovations are adopted.

Social networks sites enable interactions through content exchange, in form of links, texts, photos, multimedia content, and other. Consequently, they represent a perspective data source for decision makers and analysts. On the other hand, social networks sites have obtained a poor reputation owing to problems related with privacy protection. When considering inherent characteristics of social networking sites, privacy is a critical issue to individuals who use them. A vast amount of personal information such as users full names, photos, or e–mail address are available on many networking sites. Although the disclosure of this kind of information on web is voluntary, many users are unaware of the risks that can occur (e.g. who can use their data and for what purposes). Many users lack awareness that information disclosed on these sites cannot be permanently deleted or that it is difficult to

protect these disclosed information from being divulged or misused [11]. In order to gain customer trust and loyalty companies should use a privacy-friendly method for taking advantage of user-generated content on social networking sites [7].

In the first part of the paper the relevance of social network data analysis is emphasized, while in the second part the light is shed on the question that is often asked nowadays and which poses certain limitations in analytics in general - data privacy preservation. The rest of the article is organized as follows. In the next section the relevance of social network data analytics is discussed. In section 3, the review of key data mining techniques in social network analytics is presented. In section 4, the focus is on privacy preservation in social network analytics. Section 5 presents conclusions and related directions in the field.

2 The relevance of social network analytics

There is a gap between opinions of a company and customers regarding the role of social media in the development, design and improvement of customer experience, as well as regarding the way and goals of social networks usage [3].

The opinion gap lies on one side in company' s comprehension that consumers connect to their communities because of the content they offer. On the other hand, consumers are interested in a particular brand because they consider themselves loyal customers and have a desire for a unique experience with a given company. Therefore, in this gap lies a new business opportunity for companies, but also new challenges. Specifically, companies must understand that consumers' behavior has dramatically changed under the influence of social networks. Consequently, it is necessary to understand these changes and adapt business and marketing strategy to them. Most online users connects with others over the Internet, through Facebook or other communities, with a goal of facilitating the solution of their problems, in search for new knowledge about products, promotions and rewards, and as much as 80% of users of social network sites strongly believe their online friends and their suggestions and reviews [5]. Therefore, companies should first understand (1) the way users connect over the Internet and (2) their motives for networking. Only then can companies design their business strategy as a reaction to this findings and knowledge (regarding the common actions of their users).

Companies must also be aware that the potential of social networks for business improvement lies in their massive use. Facebook has over 500 million active users, 700 billion minutes are monthly spent on Facebook, and more than 3,5 billion pieces of content is exchanged each week. Twitter increases its user base each day by 300000 new users, while 83 tweets are generated each second that reference some product or a brand [3]. The amount of data exchanged via these social networks sites represents a promising data source. Their quality analysis can enable company's business improvement. Companies can take an advantage of new data sources and attract customers, as well as develop competitive advantages by analyzing social media content.

Tools and techniques developed for analysis and social network mining can be used in a wide range of business processes. In offering products and services to the users, a social network sites can be used to improve their experience [2]. Marketing and sales can benefit from social network mining which will allow company to leverage the power of social media for customer relationship management through trend spotting to anticipate customer needs and future business opportunities, as well as reputation monitoring. Other business process categories can also be highly influenced by social networks sites, such as human capital that can use internal social networking as well as social search for recruiting, furthermore knowledge management (particularly for knowledge-sharing and strategic-knowledge management), external relationships, and so on. Therefore, it is of utmost importance that companies recognize social networks sites as a valuable data source that can improve their business and help build competitive advantages.

3 Social network mining – techniques and application

Social networks sites contain (collects) a tremendous amount of text, image, audio or video content which can be leveraged for a wide range of business purposes. Such content richness led to two basic data types that are analyzed in the context of social networks [1]:

- Links – Linkage-based and structural analysis is dealing with linkage behavior of a network to identify important nodes, communities, links, and so on.
- Content – Content-based analysis is focused on analysis of text, images, tags, and any additional added contents.

Furthermore, different types of analysis can be conducted over social network data. Static analysis refers to analysis of the whole network, that changes slowly over time, over particular snapshots with a goal of revealing evolving communities, interactions between entities and temporal events in the network. Structural analysis of social networks sites helps to understand and model the nature of large networks that will unveil general structural dynamics. Community detection identifies structurally related

groups determining the regions of the network which have similar linkage behavior. Influence analysis clarifies the way information propagates through the network and determines the most influential members of social network. Influential users can be determined using flow models or page rank methods. Link prediction identifies important linkages that will give an idea about future relationships or those that are missing in the social network.

The bases of these analyses constitute different methods and techniques of data, web and text mining. Text mining refers to the analysis of textual data sources with the aim of extracting meaningful information by detecting lexical and linguistic patterns. Web mining involves the use of data mining techniques to content, structure and use Web resources [5]. In particular, techniques for association rules detection, clustering, classification, and analysis and detection of sequences are applied. *Web content data mining* represents a form of text mining that is applied on web sites with a goal of grouping, categorizing, analysing and retrieving the documents. This analysis is especially focused on pattern detection in big collections of documents that often change. It can also be used in detection and tracking of certain topics, for detection of crucial events that will become new topics in other documents, as well as trends that point out the increase or decrease of interest in a certain topic. *Link mining (analysis of a Web sites structure)* usually refers to the analysis of a number of sites to identify the relative importance of pages that seem equally relevant by only analyzing their contents. Content analysis is often performed together with link mining. *Analysis of Web site visits* is focused on the records of requests made by Web site visitors, most often collected as Web server logs. The content and structure of the Web site reflect the intentions of the authors and designers of the site, while the actual behavior of users of these sites may reveal additional structures (for example, analysis may indicate that customers that buy product X show interest for product Y – new knowledge about customer behavior may be used to personalize the content of an online store) [5].

Sentiment analysis is another important aspect of text analysis in social networks, which aims to extract opinions, emotions and the general sentiment of the text. It allows tracking attitudes and feelings on the Web through analysis of blogs, comments, criticisms, and tweets written on different topics. Additionally, it is possible to follow products, brands and people, and to assess global opinion of other users, whether positive or negative opinion is spread among users on social network. Therefore, sentiment analysis helps companies to follow bad reviews about their products, services or brand in general, perceptions of new products, and through this allows a company to manage its reputation.

4 Privacy preservation in social network analytics

An increasing part of individuals' social, communicative, and private actions take place in digital world, where they persist in digital form. Therefore, information technology (new possibilities and capacities that it brings) erodes privacy in various ways. Privacy is/was described using diverse conceptions, e.g. the right to be left alone, limited access to oneself, secrecy, control over personal information, personhood, and intimacy [10]. The privacy debate is quite popular, but it is only focused on information privacy (especially in online environment). The main idea in securing one's privacy is to enable control over one's personal information. Therefore, privacy is generally described as the ability or the right of an individual to control conditions under which personal information are collected and how they will be used (in future) [8]. Information privacy taxonomy describes various parts of privacy field and defines relationships among them. There are three main dimensions that are encompassed in information privacy taxonomy for collaborative environments: (1) computation view, (2) content view, and (3) structural view. The *computation view* refers to time dimension of privacy and is related to "the amount of time and resources required to compromise the stated level of privacy perception". The *content view* refers to different types of data that should be protected where data privacy, identity privacy and meta privacy can be distinguished. The *structural view* refers to privacy of entities included in collaboration (individual privacy, group privacy, and organizational privacy) [9].

Social network data comes from a variety of data sources and as such poses certain issues regarding their preparation and handling. Social network structure usually consists of implicit and explicit connections. Explicit connections are those that the user has explicitly declared, such as friends, groups or pages that are followed. Such explicit connections are often incomplete and do not describe entirely all of the relationships in the network. Implicit connections complement them and they can be discovered through user's activities by analyzing interactions between users. They can also be discovered from users' similarity – users who use same tags to describe themselves. These implicit connections are relevant to business applications.

The second issue that must be addressed in social network analysis refers to computational complexity of a large network that has millions of nodes. Problems regarding social network data refer also to the following [2]:

- Duplicate nodes – one social network user has for example two addresses.

- Inactive nodes – users who did not explicitly remove their profile from a social network but don't access it.

Social networks contain information about the individual in terms of their interests, demographic information, friendship link information, and other attributes. This can lead to disclosure of different kinds of information in the social networks, such as identity, attribute, and linkage information disclosure. Given the sensitivity of information in social relationships additional privacy issues arise - even revealing a list of friends might not be sensitive to the person who revealed the information, but it might be to their friends.

Ones' social network graph can represent a valuable source of information even if the personally identifiable information (names, identification number etc.) is removed from the data. As stated in Ref [2], the mere structure of graph can reveal the identity of the individual behind some of the nodes. Therefore, thorough data preparation must be performed in order to preserve data privacy and before the social graph is released.

Traditional data mining techniques usually operate on the original data set so the leakage of privacy data can occur. In addition, large amounts of data can implicate easy identifying of sensitive information. Since those problems challenge the traditional data mining a new trend has been introduced – privacy-preserving data mining. The privacy-preserving data mining is focused on development (and usage) of algorithms that will modify original data so the private data as well as private knowledge remain private after the mining process. Privacy preservation when using data mining is referring to two main issues [4]: (1) individual privacy preservation and (2) collective privacy preservation. Individual privacy preservation refers to the protection of personally identifiable information. Sometimes it is not enough to protect only personally identifiable information since someone's privacy can be violated through learning sensitive knowledge. The main goal of collective privacy preservation is to protect some patterns or trends indentified in datasets.

There are two privacy breaches that are usually studied: (1) identity disclosure and (2) attribute disclosure. But, for network data two new types of privacy breach can occur: (1) social link disclosure and (2) affiliation link disclosure. Social link disclosure happens when someone is able to identify the existence of a sensitive relationship (that is preferred to be private, hidden from the public) between two users. In addition, affiliation link disclosure happens when someone is able to identify whether a particular person belongs to a specific affiliation group [13].

The most often applied privacy-preserving methods are randomization, k-anonymity model, and l-diversity model [6].

Randomization Models introduce some noise into data in order to hide all identification attributes in the dataset. The amount of noise has to be sufficient to disable identification of original data values, while aggregated distributions can be made in order to conduct data analysis. Besides the randomization, methods of multiplicative perturbation can be used, where random data projection or random data rotation techniques perturb the original data. This data perturbation results in hidden individual values while useful information can be recovered, such as distribution of the data values or rules and patterns in the data. The above mentioned methods represent the two most favored randomization approaches. Closely connected with them, and also frequently used with k-anonymity method is data swapping. Data swapping preserves data privacy by interchanging the values of different records. Stated methods are applied in social network analysis, achieving a meaningful level of anonymity for the nodes and remove too many edges in the social network graph. They are usually applied in order to prevent adversaries from identifying their target in the network, or from inferring the existence of links between nodes.

Since indirect identification of an entity is possible in certain data sets even if the personally identifiable fields are removed, the *k-anonymity model* has been developed. Namely, the combination of pseudo-identifiers, such as age, ZIP code or similar attributes, can reveal the identity of a person, or at least narrow down the potential candidates. The k-anonymity model reduces the data granularity by generalization and data suppression techniques. Generalization transforms the values of an attribute to a given interval (like transforming for e.g. a date of birth to age) or groups the categorical values into different value sets. Data suppression completely disregards the attribute. With k-anonymity an original data set can be transformed where it is difficult for an intruder to determine the identity of the individuals in that data set. The transformed data set has the property that each record is similar to at least another k-1 other records (regarding the identifying variables), and is indistinguishable from them.

L-diversity model is an extension of k-anonymity model. It is developed in order to overcome the drawback of the latter to effectively prevent conclusion of delicate attribute values. This is achieved by distribution of quasiidentifier values within a group. More on application and effects of presented privacy preserving techniques can be found in [6].

Since one of the leading applications of social network analysis is marketing, a specific situation emerged as a result that has to be addressed with care: many online social network platforms share their data with third parties for advertising purposes threatening the privacy of their users. As part of their business model, many social network platforms provide open APIs that allow third parties to access user profiles or

profiles of friends with the possibility of user's privacy violation, [2]. It is important that these companies properly anonymize social network data and one possible way is to use privacy preserving techniques at the moment of collecting information from social network sites.

Privacy mechanisms for social networks mostly consider anonymization techniques for anonymizing network structure and user attributes. The anonymization techniques for network structure fall in four main categories: (1) edge modification, (2) randomization, (3) network generalization, and (4) differentially private mechanisms. However, providing the anonymizited structure of social networks is often not sufficient for the purposes of the researchers. On the other hand, the assumption is that anonymized data will have utility only if it contains both structural properties and node attributes [13].

The main reason for preserving anonymization of data from social networks is to preserve the privacy of individuals whose data are collected. Collection and aggregation of private personal data obtained from many users forms a set of data that must be managed with extraordinary attention.

5 Conclusion

Wide availability of devices with Internet access led to the increased importance of social networks. Besides fast information retrieval they offer simple, yet diversified ways of interaction between users (through book-marking/browsing marked sites of other users, voting for articles and making comments about them, adding friends, making comments about the content of their profiles, joining groups and discussions, exchanging photos and video content, and by adding articles, or editing and modifying the existing ones).

There are lots of advantages that social networks sites usage brings to businesses, providing rich content-based knowledge which can be exploited for development of an effective social business strategy. The social networks sites enabled interactions of users through content exchange (in form of links, texts, photos, multimedia content, and other). They also represent perspective data source for decision makers and analysts and can be advantageous for a wide range of business purposes: (1) static analysis reveals evolving communities, interactions between entities and temporal events in the network; (2) structural analysis helps to understand and model a general structural dynamics of social networks; (3) community detection determines network regions with a similar linkage behavior; influence analysis clarifies the way information propagates through the network and determines the most influential members; (4) link prediction gives an idea about future and missing relationships in the social network.

At the same time, social networks deliver many privacy preserving problems. As stated above, privacy is generally described as the ability or the right of an individual to control conditions under which personal information are collected and how they will be used. Diversified data mining methods and techniques are developed for the analysis of content and structural information (i.e. linkage behavior) in social networks. Analysis of data sources includes text mining (extraction of meaningful information from text by detecting lexical and linguistic patterns), Web content mining (pattern detection in big collections of documents that often change on web pages), and sentiment analysis (racking attitudes and feelings on the Web through analysis of blogs, comments, criticisms, and tweets written on different topics). Analysis of Web resources usage and structure (so called Web mining) includes link mining (identification of relative importance of pages with equal contents relevance), and Web site visits mining (profiling users' intentions, interests and preferences). Data subject in these analyses always contains numerous personal information of social networks users (such as full names, photos, e–mail addresses, demographic information, etc.).

Although, voluntarily provided, personal information is difficult to protect and consequently, a new approach to data mining, so called privacy-preserving data mining has been developed. It is focused on the development and usage of algorithms that modify the original data so the private data, as well as the private knowledge, remain private after the mining process. The most often applied privacy-preserving methods: randomization, k-anonymity model, and l-diversity model, are described in short in the paper. They enable identity, attribute, social links and affiliation links privacy preservation and eliminate the threats of private users' data disclosure, their misuses and other risks. Therefore, in order to gain customer trust and loyalty, companies should use privacy-friendly methods for taking advantage of user-generated content on social networking sites.

References

[1] Aggarwal CC. (Ed.) *Social network data analytics*, Springer, 2011.

[2] Bonchi F; Castillo C; Gionis A; Jaimes A. Social network analysis and mining for business applications, *ACM Transactions on Intelligent Systems and Technology*, 2(3):1-37, 2011.

[3] CMO Council, Variance in the social brand experience: Timely opportunities for social business advantage, http://www.cmocouncil.org/variance-in-the-social-brand-experience.php, downloaded May 2nd 2012.

[4] Ge, X; Zhu, J. Privacy preserving data mining, in *New Fundamental Technologies in Data Mining*, Funatsu, K. (Ed.), 2011, http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/privacy-preserving-data-mining.htm, downloaded April 5th 2012.

[5] Grljević O; Bošnjak Z. Značaj analize sadržaja socijalnih medija, *YU-INFO 2012*, Kopaonik, Serbia, 2012.

[6] Grljević, O; Bošnjak, Z; Mekovec R; Privacy preserving in data mining – experimental research on SMEs data, *Proceeding of the SISY 2011, 9th IEEE International Symposium on Intelligent Systems and Informatics*, Subotica, Srbija, 2011.

[7] Provost, F; Dalessandro, B; Hook, R; Zhang, X; Murray, A. Audience selection for on-line brand advertising: privacy-friendly social network targeting, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09*, Paris, France, 2009.

[8] Schwaig, KS; Kane, GC; Storey, VC. Compliance to the fair information practices: how are the Fortune 500 handling online privacy disclosures? *Information & Management*, 43(7): 805-820, 2006.

[9] Skinner, G; Han, S; Chang, E. An information privacy taxonomy for collaborative environments, *Information Management & Computer Security*, 14(4):382-394, 2006.

[10] Solove, DJ. *Understanding privacy*, Harvard University Press, 2008.

[11] Tootoonchian, A; Saroiu, S; Ganjali, Y; Wolman, A. Lockr: better privacy for social networks, *Proceedings of the 5th international conference on emerging networking experiments and technologies CoNEXT '09*, Rome, Italy, 2009.

[12] Watanabe, C; Amagasa, T; Liu, L. Privacy risks and countermeasures in publishing and mining social network data, *Proceedings of the 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom 2011*, Florida, USA, 2011.

[13] Zheleva, E. Prediction, evolution and privacy in social and affiliation networks, PhD dissertation, July, 2011, http://linqs.cs.umd.edu/basilic/web/Publications/2011/zheleva:phdthesis11/zheleva-phdthesis11.pdf, downloaded April 5th 2012.