SONET: A SOcial NETwork Model for Privacy Monitoring and Ranking

Raj Kumar Nepali
College of Business and Information Systems
Dakota State University
Madison, SD, 57042
rknepali@pluto.dsu.edu

Yong Wang
College of Business and Information Systems
Dakota State University
Madison, SD, 57042
yong.wang@ dsu.edu

***Abstract*—Many concerns have been raised regarding the privacy issues in social media. The risks, as well as the security and privacy issues of social media in business, public policy, and legislation need to be evaluated and studied. However, there is lack of effective and practical way to quantify, measure, and evaluate privacy. In this paper, we propose a social network model, SONET, for privacy monitoring and ranking. The model provides a novel, effective, and practical way to quantify, measure, and evaluate privacy. Further, the proposed model is also flexible and built on real data from social media. The proposed privacy risk indicator, *PIDX*, can be calculated in real time and the value can be used for privacy monitoring and risk control.**

Keywords: social networks, privacy index, monitoring, ranking

I. INTRODUCTION

Social media often refers to online social networking sites (OSNs), such as those web-based and mobile-based web sites which could be used for interactive communication among organizations, communities, and individuals. Facebook, Twitter, LinkedIn, Google+ all belong to this category. In addition, online game web sites, such as Xbox Live and PlayStation Network, are not only places to play games, but also places for social interactions and staying connected with friends. Social media has become a very popular tool for individuals and enterprises. In March 2012, there were 901 million active users in Facebook according to its web site. In this paper, we consider a broad definition of social media. We consider all public accessible information as social media, such as OSNs, public records of law court proceedings, records of births, marriages, etc.

The following characteristics unique to social media contribute to the challenges involved in our research:

- **Information Sharing**: social media is built around the premise of information sharing. Information is shared among organizations, governments, communities, and individuals.
- **Public Information:** Social media information is available for public access. Social media includes unclassified information and the information is available for public access. Lots of information which was not accessible before is available now.
- **Scattered Data**: There is no central place to contain all the data. Data is scattered in various data sources.
- **Sparse Data**: Data is loosely connected. No single data structure can be used to represent the data.
- **Big Data Availability**: As the scope of Big Data is so vast, efficient data retrieving and searching technologies are desired.
- **Anonymity:** Information on social media sites may be published anonymously. Data sources are not easily identified.
- **Redundancy:** Information may be duplicated.
- **Validity Uncertain**: Social media includes millions of information sources. The validity of data cannot always be determined.
- **Aggregation:** Information is scattered everywhere and data aggregation is desired to reduce amount of data.

Many concerns have been raised due to security and privacy issues within the social media context. Recent incidents indicate that risks, security, and privacy issues of social media to business, public policy, and legislation need to be further evaluated and studied. For example, between April 17 and April 19, 2011, user account information for the Sony PlayStation Network and its Qriocity service was compromised [1]. Sony suffered a massive breach in its video game online network that led to the theft of names, addresses and possibly credit card data belonging to 77 million user accounts. It is one of the largest-ever Internet security breaches. Sony has estimated a total cost of $172 million for the PlayStation Network breach. As technologies advance, more information will be available for public access. Thus, social media security and privacy is critical and will continue to become essential.

In this paper, we propose a social network model, SONET, for social media privacy monitoring and ranking. Our proposed model provides a novel way to quantify, measure, and evaluate privacy. Few works have been conducted in the literature and have limitations when used to measure privacy. Our proposed model removes those limitations and provides an effective and practical way to measure privacy.

The paper is organized as follows: Section II discuss the related work. Section III introduces our proposed model, followed by comparison and analysis in Section IV. Section V summarizes the paper and future works.

II. RELATED WORK

Social media privacy has raised many concerns and may affect individuals, enterprises, legislatures, and government agencies. Different types of attacks have been investigated in [2][3][4][5] and a few privacy preserving schemes are proposed in [6][7]. Previous works on social media privacy focus on privacy preserving [8][9] and privacy policy

978-0-7695-5023-7/13 $25.00 © 2013 IEEE
DOI 10.1109/ICDCSW.2013.49

conflicts [9][10]. Few works [12][13][14] [15] have been conducted on privacy measurement due to the challenges to quantify the privacy risk associated with online social network users.

Recent works attempting to quantify the privacy risks associated with the usage of online social networks can be found in [12][13][14] [15]. In [12], the authors propose to use the amount of information revealed in online social networks to quantify the privacy risks. In [13], the authors present an approach in which privacy score is calculated by computing sensitivity and visibility of attributes. Naïve approach for evaluating sensitivity and visibility of attributes is demonstrated in [13]. The authors extend their works to another approach in [14]. They use Item Response Theory (IRT) to evaluate sensitivity and visibility of attributes when evaluating privacy scores. The authors in [15] develop a tool, Privometer, to measure information leakage. The leakage is indicated by a numerical value. The tool can suggest self sanitization actions based on the numerical value.

III. SONET MODEL

In this paper, we propose a social network model, SONET, for privacy monitoring and ranking. Figure 1 shows an overview of the model. The model includes 6 components which are described below:

- Social network model
- Deep web searching and SONET formation
- Data aggregation to explore and deduce data
- Privacy Index (*PIDX*) and privacy invasion
- Security protection and privacy preserving
- Privacy monitoring and countermeasures

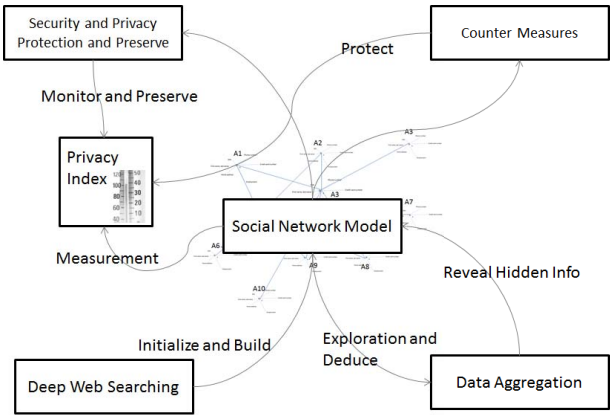

Figure 1. SONET Model

A. Social Network Model and Definition

Social media includes huge amounts of data and the data is scattered everywhere. It can be extremely challenging to search for, organize, and analyze desired data. The proposed social network model is based on two fundamental components, actor model for individuals and community model for groups.

1. Actor Model

Definition *An actor is a social entity (e.g. people, organization, etc) in a social network.*

Definition *Actor has certain characteristics that describe its features known as attributes.*

Definition *Each attribute has a different impact on privacy. This impact is referred as Attribute Privacy Impact Factor. Privacy impact factor is a numerical value.*

An actor has attributes like name, address, social security number (SSN), phone number, education, marital status, etc. In general, SSN has higher impact on privacy than phone number. Let A_i be an actor and $L_i = \{a_{i1}, a_{i2}, \dots, a_{in}\}$ represent its attributes. Then, $A_i(a_{i1}, a_{i2}, \dots, a_{in})$ is a representation of an actor with attributes.

We consider privacy impact factor for full privacy disclosure is 1. An attribute's privacy impact factor is a ratio of its privacy impact to full privacy disclosure. Thus, an attribute's privacy impact has a value between 0 and 1. We use $S_i = \{s_{i1}, s_{i2}, \dots, s_{in}\}$ to represent attributes' privacy impact factors.

Actor and attribute relationship can be represented using a weighted graph $\mathcal{G}_i = (V_i, E_i, S_i)$. In this graph, we have $V_i = (A_i, a_{i1}, a_{i2}, \dots, a_{in})$ where A_i is an actor and a_{ij} is one of the actor's attributes. An edge e_j belongs to E_i if

$e_j: (a_{ij}, A_i) \in E_i$ *if* a_{ij} *is one of the actor's attributes*

There are some attributes which are dependent on each other. For example, if an actor's occupation is known, his salary information can be inferred too. Such information is known as hidden information.

Definition *Hidden information is indirect information which is not available firsthand but could be inferred from existing data.*

We use (a_{ij}, a_{ik}, p_{jk}) to further indicate hidden relationship $a_{ij} \xrightarrow{p_{jk}} a_{ik}$ between two attributes. (a_{ij}, a_{ik}, p_{jk}) indicates that a_{ik} can be inferred from a_{ij} in possibility p_{jk}. For each hidden relationship (a_{ij}, a_{ik}, p_{jk}), we also have:

$$unidireciton\ edge\ \overrightarrow{e_{jk}}: (a_{ij}, a_{ik}) \in E_i \text{ if } a_{ij} \rightarrow a_{ik}$$

Thus, an actor model is defined and represented by a graph $\mathcal{G}_i = (V_i, E_i, S_i)$. An example of actor model can be represented using a star graph and is shown in Figure 2.

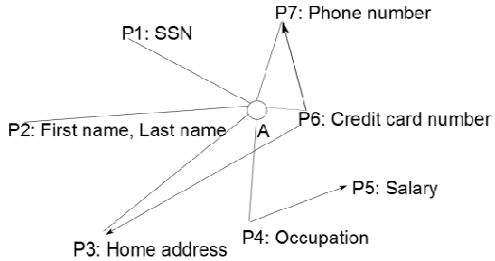

Figure 2. Actor Model Graph

2. Community Model

Definition *A community is a network of actors tied together with some common interests.*

A community consists of a group of actors and the relationships between them, for example, a community of students. Actors depend on each other in a social network. We represent such relationships as undirected relationship as well as directed relationship. In undirected relationship, either the relationship exists or do not exist. In directed relationship, only one user contributes to the relationship. For example: Actor *A* knows *B* while *B* does not know *A*.

Let *A*, *B* be two actors in the community C, the relationship between *A* and *B* can be described as in Table 1:

Relationships	Notation	Edge
A knows B; B does not know A	A→B	Directed
B knows A; A does not know B	B→A	Directed
A, B knows each other	A—B	Undirected

Table 1. Actor Relationships

Let $G_c = (V_c, E_c, S_c)$ be a graph and represent the community C. The community has n actors, i.e., $A_1, A_2, \dots, A_n$. We use $G_i = (V_i, E_i, S_i)$ to represent each actor's actor model. Then, we have

$$\begin{cases} V_c = V_1 \cup V_2 \cup \dots \cup V_n \\ e \in E_c \ if \ e \in E_i \\ S_c = S_1 \cup S_2 \cup \dots \cup S_n \end{cases}$$

We use (A_j, A_k, p_{jk}) to further indicate actor to actor relationship. $A_j \xrightarrow{p_{jk}} A_k$ indicates that A_j knows A_k and A_k's information (attributes) can be inferred from A_j in possibility p_{jk}. Thus, we have

$$\overrightarrow{e_{jk}}: (A_j, A_k) \in E_c \ if \ A_j \xrightarrow{p_{jk}} A_k$$

It may be difficult sometimes to find certain attributes of an actor from a user's profile. However, such information can be inferred easily from its social networks. For example, in a small community has a common interest in soccer, it is highly likely that an actor *A*, associating to the community, likes soccer too. Figure 3 shows a simple community graph which includes 3 actors.

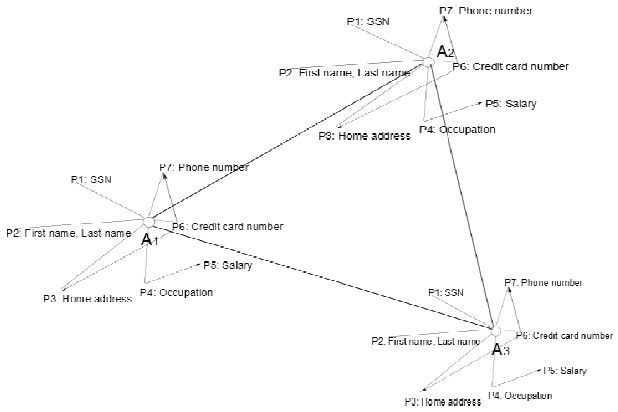

Figure 3. Community Model Graph

3. Social Network Model

Definition *A social network is a network of communities.*

A community is a subset of a social network. Let $G_S = (V_S, E_S, S_S)$ represent a social network and $G_{ci} = (V_{ci}, E_{ci}, S_{ci})$ be one of the communities. We have

$$\begin{cases} V_S = V_{c1} \cup V_{c2} \cup \dots \cup V_{cn} \\ e \in E_S \ if \ e \in E_{ci} \\ S_S = S_{c1} \cup S_{c2} \cup \dots \cup S_{cn} \end{cases}$$

Similarly, we use (A_j, A_k, p_{jk}) to indicate actor to actor relationship. For each such relationship $A_j \xrightarrow{p_{jk}} A_k$ in a social network, we have

$$\overrightarrow{e_{jk}}: (A_j, A_k) \in E_c \ if \ A_j \xrightarrow{p_{jk}} A_k$$

4. Privacy Index

Among all the attributes $L = \{a_1, a_2, \dots, a_n\}$ in an actor model, there is a subset of attributes which might be known. If these attributes are known, privacy might be disclosed. Let $L_k = \{a'_1, a'_2, \dots, a'_m\}$ be a known attribute list and $S_k = \{s'_1, s'_2, \dots, s'_m\}$ be the corresponding privacy impact factors of these attributes. Let w_{L_k}

$$w_{L_k} = s'_1 + s'_2 + \dots + s'_m = \sum_{k=1}^{m} s'_k$$

be the total weight of L_k. **Privacy Index** is defined as below.

Definition *Privacy Index (PIDX) is used to describe an entity's privacy exposure factor based on the known attributes. Higher PIDX indicates higher exposure of privacy. Privacy Index PIDX is between 0 and 100.*

$$PIDX = \frac{w_{L_k}}{w_L} \times 100$$

Let T represent a threshold which is critical to entity's security and privacy. We define **privacy invasion** as

$$PIDX \geq T$$

If $PIDX$ is lower than T, privacy is considered to be preserved.

a. Senstivity

Each attribute is given a privacy impact factor. The privacy impact factor reflects the sensitivity of the attribute. Large number indicates more sensitive information. It also indicates a higher privacy risk if the information is disclosed.

b. Visibility

SONET model defines three relationships, i.e., attribute to actor, attribute to attribute, and actor to actor. As discussed above, the attribute to attribute, actor to actor relationships are described by possibilities. The possibility is a reflection of information visibility. The visibility for an actor model can be represented using $V_i = \{p_{i1}, p_{i2}, \dots, p_{in}\}$.

In general, let $L_k = \{a'_1, a'_2, \dots, a'_m\}$ be a known subset of *A*'s attributes, $S_k = \{s'_1, s'_2, \dots, s'_m\}$ be the corresponding

weights, and $V_i = \{p'_1, p'_2, \ldots, p'_m\}$ be the visibility of each attribute. We have

$$w_{L_k} = p'_1 s'_1 + p'_2 s'_2 + \ldots + p'_m s'_m = \sum_{k=1}^{m} p'_k s'_k$$

Thus, in consideration of the attributes' visibility, privacy index can be defined as

$$PIDX = \frac{w_{L_k}}{w_L} = \frac{\sum_{k=1}^{m} p'_k s'_k}{\sum_{k=1}^{n} s_k} \times 100$$

Visibility is measured by the possibilities of attribute to attribute, actor to actor relationships. To measure A's privacy index in a social network graph $G_S = (V_S, E_S, S_S)$, assume traverse depth is d, depth-first or breath-first traversals could be used to calculate the visibility of each known attributes.

Privacy index provides a practical way to measure and evaluate privacy. We consider both sensitivity and visibility of attributes in the model and privacy index can be calculated accordingly.

B. Deep Web Searching

A social network model can be built and formed using the presented information. The more information we can collect, the better we can build the model. The social network model can be built using a two-step approach, i.e., SONET formation based on regular search engines, SONET formation based on deep web searching engines.

The first step is to build a SONET model based on search results from an existing search engine, such as Google and Bing. Search results will be retrieved and parsed. Interesting information, such as attributes (SSN, address, phone number), is extracted and used to build the model. The model can then be stored and updated when new information is available. The approach can be further integrated with data from Facebook, Twitter, LinkedIn, and other social networking sites.

The second step is to extend the first approach to support deep web searching technologies. Information is scattered everywhere. However, it usually resides as hidden or unknown data in deep web sites, beyond the reach of traditional search engines. According to [16], there were 43,000-96,000 "deep web sites" which could not effectively searched by traditional search engines and there was an information estimate of 7,500 terabytes data hidden inside the "deep web sites". In this approach, the model can be further populated from multiple data sources using deep web searching, such as public accessible databases and public records.

C. Data Aggregation to Explore and Deduce Data

With a SONET model built from deep web searching technologies, data aggregation techniques are further used to explore and deduce more useful data. Data aggregation techniques include redundant information removal, false information filter and hidden information explorer.

The goal of redundant information removal is to remove redundant data from the model. Duplicate information may exist and needs to be removed to improve efficiency and simplify the model.

Data validity is uncertain in social media. One way to counteract this is by implementing the false information filter which could be used to eliminate false information. It will greatly increase the accuracy of the proposed model. Cross-referencing can be used to filter out false information. Different approaches can be used. Examples include:

- **Repetition:** based on the number of repeat times of the information. The information which appeared the most number of times is valid. Any information against it is false information.
- **Source**: based on the authority of the data source. If the information comes from an authority source or a source which is known to be trustable, the information is valid. Any information against it will be identified as false information.
- **Latest first**: latest information is identified as valid. Any other information against it is false information.
- **Earliest first**: the information which appeared earliest is valid. Any other information against it is false information.
- **Rumor first**: any information from unknown data source is valid. Any other information against it is false information.
- **Veto first**: any information against a fact is valid. Any other information against it is false information.

Hidden information (indirect information) explorer is used to further deduce data and reveal more useful data.

D. Privacy Preserving and Countermeasures to Protect Security and Privacy

The goal of social media security and privacy preservation is to ensure consumer, government, and businesses privacy when using social media. It is a challenging issue and lacks measurement metrics to define and describe the problem. However, the SONET model and Privacy Index (*PIDX*) provide an innovative approach to explore the challenge.

The goal of privacy preserving with social media is to minimize the privacy index *PIDX*. Clustering and localization approaches could be used to identify interesting attribute groups. Using the SONET model, we can

- easily calculate *PIDX* for an entity
- evaluate the security and privacy impact if certain information is exposed
- easily evaluate the impact and the severity of the incident should one occur

The SONET model also provides a practical way to monitor privacy exposure in real time. The model can be used as a countermeasure approach to protect privacy. Using the SONET, a privacy monitoring system can be further explored. The system automatically retrieves data and

acquires information from various data sources. Based on the updated data, *PIDX* is calculated in real time. Should *PIDX* exceed a threshold, an alert message is sent to the corresponding parties to prevent data breaches and privacy invasion.

IV. Comparison and Analysis

The proposed SONET model provides a practical way to quantify, measure, and evaluate privacy. The proposed model has the following features:

- *Accurate* SONET model provides a way to quantify and measure individual's privacy.
- *Dynamic* The model utilizes data from real social media, instead of data from surveys. Privacy index can be calculated dynamically.
- *Real time* Privacy index can be measured constantly and thus a real time privacy monitoring system is practical.
- *Flexible* SONET model can be formed based on regular and deep web search engines. It can be extended to the whole Internet.

The proposed model differs significantly from the privacy scores in [13][14]. The proposed approach in [14] is based on IRT. However, IRT is not designed for a complex behavioral network like social networks. IRT has three basic assumptions: items are independent, users are independent, and items and users are independent. However, these assumptions do not apply to social networks. Further, the work in [14] assumes attributes are independent and does not consider relationships among attributes. However, as found in [17], revelation of a combination of a few attributes can jeopardize the privacy of a user since it can lead to easy access of other attributes. The work in [17] shows that 87% of Americans can be uniquely identified by five digit zip code, gender, and date of birth. However, none of them alone can be significantly affects privacy [17].

SONET model does not have these limitations and it considers all the relationships between actors and attributes. The sensitivity and visibility of attributes are further characterized by the possibilities in attribute to attribute and actor to actor relationships. SONET model provides a novel, practical, and effective way to quantify, measure, and evaluate privacy.

V. Conclude and Future Works

Many concerns have been raised regarding the security and privacy issues of social media. The risks, as well as the security and privacy issues of social media in business, public policy, and legislation need to be evaluated and studied. In this paper, we propose a social network model, SONET, for privacy monitoring and ranking. The model provides a novel and practical way to quantify, measure, and evaluate privacy. The model can be formed based on real data from social media. The privacy index can be calculated and monitored in real time. Future works are continued to further evaluate the model using the real data from the Internet. Different approaches to assign attribute privacy factors will be studied.

References

[1] C. Morris, SONY: PlayStation Breach Involves 70 Million Subscribers, April 26, 2011, CNBC.com.

[2] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogeneous networks," in WSDM'12, 2012, pp. 743–752.

[3] E. Zheleva and L. Getoor, "To join or not to join?: The illusion of privacy in social networks with mixed public and private user profiles," in WWW 2009, 2009, pp. 531–540.

[4] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, "A Practical Attack to De-anonymize Social Network Users," 2010 IEEE Symposium on Security and Privacy, pp. 223–238, 2010.

[5] A. Narayanan and V. Shmatikov, "De-anonymizing Social Networks," in 2009 30th IEEE Symposium on Security and Privacy, 2009, pp. 173–187.

[6] L. Sweeney, "K-anonymity: a model for protecting privacy," International Journal on uncertainity, Fuzziness and knowledge-based system, vol. 10, no. 5, pp. 557–570, 2002.

[7] A. Machanavajjhala, J. Gehrke, D. Kifler, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in Proceedings of the 22nd IEEE International Conference on Data Engineering, 2006.

[8] J. Anderson, C. Diaz, F. Stajano, K. U. Leuven, and J. Bonneau, "Privacy-Enabling Social Networking over untrusted networks," in WONS, 2009, pp. 2–7.

[9] D. Starin, R. Baden, A. Bender, N. Spring, and B. Bhattacharjee, "Persona?: An Online Social Network with User-Defined Privacy Categories and Subject Descriptors," in SIGCOMM'09, 2009, pp. 135–146.

[10] A. Yamada, T. H. Kim, and A. Perrig, "Exploiting Privacy Policy Conflicts in Online Social Exploiting privacy policy conflicts in online social networks," 2012.

[11] Y. Liu, K. P. Gummadi, and A. Mislove, "Analyzing Facebook Privacy Settings?: User Expectations vs . Reality," in IMC' 11, 2011.

[12] J. Becker and H. Chen, "Measuring Privacy Risk in Online Social Networks," in Web 2.0 security and privacy Workshop, 2009.

[13] E. M. Maximilien, T. Grandison, T. Sun, D. Richardson, S. Guo, and K. Liu, "Privacy-as-a-Service?: Models , algorithms , and results on the facebook platform," in Web 2.0 Security and privacy workshop, 2009.

[14] K. U. N. Liu, "A Framework for Computing the Privacy Scores of Users in Online Social Networks," Knowl. Discov. Data, vol. 5, no. 1, pp. 1–30, 2010.

[15] N. Talukder, M. Ouzzani, A. K. Elmagarmid, H. Elmeleegy, and M. Yakout, "Privometer: Privacy protection in social networks," 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010), pp. 266–269, 2010.

[16] M. Bergman, The Deep Web: Surfacing Hidden Value. July 2000. BrightPlanet LLC.

[17] L. Sweeney, "Uniqueness of simple demographics in the U. S. population," in Data privacy Lab white paper series LIDAP-WP4, 2000.