Towards Modelling Privacy Risks in Geo-Social Networks

Alia I. Abdelmoty

School of School of Computer Science & Informatics,
Cardiff University, Wales, UK
Email: AbdelmotyAI@cardiff.ac.uk

***Abstract*—"Social privacy" concerns how individuals manage self-disclosure, availability, and access to information about themselves by other people when using social-driven applications. To manage social privacy, one needs to understand the level of threat implied by his information disclosure and be able to relate it to the scope of visibility granted for this information. This paper argues that the risk to personal privacy comes from the implicit information embedded in the relationships between the different elements of data collected on these networks. A proposal is made to explicitly represent such relationships and use them to model the level of threat to personal privacy on these networks. Exposure of this information will enable users to be aware of their own data and to make informed decisions on their sharing behaviour online.**

Keywords–Location Privacy; Geo-Social Networks; User Profiles.

I. INTRODUCTION

The proliferation and affordablity of location tracking-enabled devices are allowing individuals to accumulate an increasing amount of personal information, such as their mobility tracks, geographically tagged photos and events. Embracing these new location-aware capabilities by social networks has led to the emergence of Geo-Social Networks (GeoSNs) that offer their users the ability to geo-reference their submissions and to share their location with other users. Subsequently, users can use location identifiers to browse and search for resources. GeoSNs include Location-Enabled Social Networks (LESNs), for example, Facebook, Twitter, Instagram and Flickr, where users' locations are supplementary identification of other primary data sets, and Location-Based Social Networks (LBSNs), for example, Foursquare and Yelp, where location is an essential key for providing the service.

In addition to location data that describe the places visited by users, GeoSNs also records other personal information, such as user's friends, reviews and tips, possibly over long periods of time. User's historical location information can be related to contextual and semantic information publicly available online and can be used to infer personal information and to construct a comprehensive user profile [1] [2]. Derived information in such profiles can include user activities, interests and mobility patterns. Users may not be fully aware of what location information are being collected, how the information are used and by whom, and hence can fail to appreciate the possible potential risks of disclosing their location information. Methods of exposing both the explicitly collected and implicitly derived information from user location are needed to enable users' awareness, and to allow users to make informed decisions about sharing their data online.

In this paper, the type of information stored in user profiles on GeoSNs are considered as a folksonomy structure of user, place and tag entities. A layer of privacy risk levels is proposed to label the relationships between these entities in the folksonomy graph, based on the degree of associations between them. A lot of work has been done recently on exploring the content of information shared by users on GeoSNs. On the other hand, a lot of work is ongoing to explore the privacy threats posed by sharing this information online. In this paper we combine both lines of research and propose a new approach to associating the information shared with its possible privacy risks. By representing the implicit content in the user profile data, application can help users appreciate the possible privacy risks associated with their sharing behaviour and thus allow them to make informed decisions on disclosing their information. The work presented here is a first step towards building privacy-aware GeoSNs.

An overview of related work is presented in Section II. In Section III, the geo-folksonomy data model is used to store the information collected by the GeoSNs. The model is extended with the proposed privacy levels information. A framework for a privacy-enabled GeoSN is also presented. In Section IV, example user profiles, defined using the enriched geo-folksonomy model, are described. Conclusions and an overview of future work are given in Section V.

II. RELATED WORK

Significant interest is witnessed recently in studying the value and utility of location information in GeoSNs for the purpose of user and place profiling. Here, we review some of the methods used for extracting the explicit and implicit content of the data generated on these networks and some of the work done on user profiling with this information.

Some works utilised publicly available information from GeoSNs in order to derive or predict users' location. In [3], Twitter users' city-level locations were estimated by exploiting their tweet contents with which it was possible to predict more than half of the sample set within 100 miles of their actual place. Similarly, Pontes et al. [4] examined how much personal information can be inferred from the publicly available information of Foursquare users and found the home cities of more than two-thirds of the sample within 50 kilometres. Sadilek et al. [5] investigated novel approaches for inferring users' location at any given time by knowing the GPS positions of their friends on Twitter. For almost 84% of users the exact locations were derived even when setting their location data as private, where an accuracy of 57% was accomplished by using information of only two friends.

 ISBN: 978-1-61208-617-0

Sharing location information on GeoSNs can be utilised to analyse and predict spatiotemporal user behaviour including their interests, activities, mobility patterns as well as future movement. Location-Centric Profiles (LCPs) are proposed in [6] that contain aggregated statistics extracted from profiles of users who visited a given location on GeoSNs. These were provided for the venue owner as a way for monitoring their business. Vosecky [7] modelled users' interests shared on microblogs in relation to their corresponding disclosed locations. Users' geographic location from Twitter was extracted from the locations directly tagged by them or from those mentioned in their tweets [7]. Users' geographical regions of interests are then derived that represent clusters of personal activity.

Rossi and Musolesi [8] proposed and tested three approaches for identifying users by exploiting their check-in information on LBSNs, particularly spatiotemporal tracks, frequency of visit, and users' social ties. Evaluation results showed that only a small amount of check-in information was adequate to identify users with high accuracy, where almost 80% of users were successfully identified in some datasets. Zhong et. al. [9] were pioneers in exploiting the predictability aspect of location check-ins in order to develop location-to-profile framework that infers demographics of users. In particular, they derived enriched check-ins' semantics based on three main factors, namely, spatiality, temporality and location knowledge such as customer review sites and social networks. A series of experiments were carried out on the dataset that revealed the feasibility of deriving users' demographics from their check-in information, where gender and educational background attributes provided the best outcomes followed by age, sexual orientation, marital status, blood type and zodiac sign. More recently, researchers have exploited GeoSNs to explore the personality aspect by examining the reciprocal relationship between users and spatiotemporal features. In Chorley et. al. [10], a study was conducted to understand human behaviour in terms of examining the relationship between the location types visited by Foursquare users and their personality. A five-factor personality model was proposed and correlations were observed between the personality traits and Foursquare check-in attitude.

The above studies show a significant potential for deriving personal information form GeoSNs and the implication of possible privacy threats to users of these applications. A lot of work considered methods of user profiling with personal location information collected on GeoSNs, but no works have yet considered the privacy implications of building such profiles and how to address the threat for the users of these networks.

III. The Geo-Folksonomy Model

In this work, we use a folksonomy data model to represent user-place relationships and derive tag assignments from users' actions of check-ins and annotation of venues [1]. In particular, tags are assigned to venues in our data model in two scenarios as follows.

1) A user's check-in results in the assignment of place categories associated with the place as tags annotated by this user. Thus, a check-in by user u in place r with the categories (represented as keywords) x, y and z, will be considered as an assertion of the form $(u, r, (x, y, z))$. This in turn will be transformed to a set of triples $\{(u, r, x), (u, r, y), (u, r, z)\}$ in the folksonomy.
2) A user's tip in the place also results in the assignment of place categories as tags, in addition to the set of keywords extracted from the tip. Thus, in the above example, a tip by u in r with the keywords $(t_1, \cdots, t_n)$, will be considered as an assertion of the form $(u, r, (x, y, z, t_1, \cdots, t_n))$, and is in turn transformed to individual triples between the user, place and tags in the folksonomy.

The data collected by the GeoSN can be represented as a geo-folksonomy, which can be defined as a quadruple $\mathbb{F} := (U, T, R, Y)$, where U, T, R are finite sets of instances of users, tags and places respectively, and Y defines a relation, the tag assignment, between these sets, that is, $Y \subseteq U \times T \times R$.

A geo-folksonomy can be transformed into a tripartite undirected graph, which is denoted as folksonomy graph $\mathbb{G}_\mathbb{F}$. A geo-Folksonomy Graph $\mathbb{G}_\mathbb{F} = (V_\mathbb{F}, E_\mathbb{F})$ is an undirected weighted tripartite graph that models a given folksonomy $\mathbb{F}$, where: $V_\mathbb{F} = U \cup T \cup R$ is the set of nodes, $E_\mathbb{F} = \{\{u, t\}, \{t, r\}, \{u, r\} | (u, t, r) \in Y\}\}$ is the set of edges, and a weight w is associated with each edge $e \in E_\mathbb{F}$.

The weight associated with an edge $\{u, t\}, \{t, r\}$ and $\{u, r\}$ corresponds to the co-occurrence frequency of the corresponding nodes within the set of tag assignments Y. For example, $w(t, r) = |\{u \in U : (u, t, r) \in Y\}|$ corresponds to the number of users that assigned tag t to place r.

A. Privacy-oriented Geo-Folksonomy Model

Here, a possible model is proposed of the levels of privacy threats with respect to the user geo-profile. Two variables contribute to the level of threat to user's privacy on social networks, namely, the amount and content of the disclosed information, and the visibility scope of this information. Here, we focus on the data content and isolate the visibility variable, i.e. we assume that all data in a user profile is available to potential adversaries. This is not an unreasonable assumption given that the application owns all the user data sets it collects. The scope of visibility can be used to control access to user's data in a privacy-oriented system design, which is the subject of future work.

With respect to data content, the level of risk to personal privacy can be quantitatively assessed using the amount of data disclosed by the user; the level of risk is directly proportional to the amount of data disclosed. The more data stored about the user's spatio-temporal history, the more inferences that can be made in the profile. Data have three explicit dimensions: spatial, social and temporal. Reasoning with the relationships between these dimensions can result in the inference of implicit personal information that the user may not have wished to disclose. For example, reasoning along the spatio-temporal dimension can reveal patterns of presence or absence from places and the degree of attachments to place, etc. An abstract "traffic-light" model is proposed here to communicate the level of risk to user's privacy on GeoSNs. Three levels are defined as follows.

- Green: safe to disclose the information,
- Amber: caution; disclosing the information can result in moderate privacy implications, and,

 ISBN: 978-1-61208-617-0

- Red: danger; disclosing the information can result in risky privacy implications.

The levels are mapped to the degree of association computed between the entities in the geo-folksonomy, namely, between different entities (user and place, user and tag, place and tag), as well as between similar entities (a user and other users, place and other places, and tags). The familiar "traffic light" metaphor is used to enable a quick and accurate interpretation of the communicated information to users.

Every edge e in the geo-folksonomy graph is given a privacy label $vc = Green|Amber|Red$, that is a function of the pre-assigned weight on the edge. Thus, for example, $vc(t,r) = f(w(t,r))$ and $vc(u,r) = f(w(u,r))$, etc. Note that as the weights on the edges are dynamic, the labels used can also change over time. For example, a label may initially be green, and then can change to amber or red as the frequency of the user visits to the place increases. Note also that the function used for assigning the privacy labels can be more realistically defined by considering the pattern of association in addition to the frequency. For example, a periodic tag assignment by the user is more revealing of the user's behaviour than a random assignment for the same frequency.

Figure 1 depicts the overall process of user profile creation. The process starts with data collection of check-ins and tip data from the GeoSN, that are then processed to extract users, places and tags and their associated properties. The modelling stage includes the definition of relationships between the three entities and the computation of weights on the edges of the folksonomy graph using co-occurence methods. The privacy-level detection module takes the folksonomy graph as input and computes the privacy levels for all the edges in the graph. The enriched folksonomy graph is then used to create the different user profiles. The user similarity module uses the generated profile to compute similarity vectors for users in the data set. The privacy notification and feedback module uses the generated privacy levels to present the data to the user through the user interface.

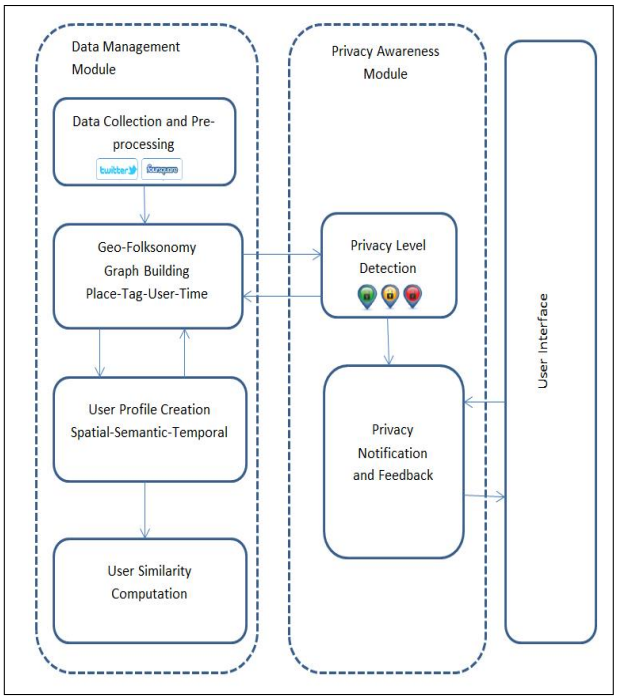

Figure 1. Framework of the privacy-enabled GeoSNs.

IV. Privacy-aware User Geo-Profiles

The geo-folksonomy can be used to represent a user's spatial and semantic association with place. A spatial user profile represents the user's interest in places, while a tag-based profile describes his association with concepts associated with places in the folksonomy model. Similarity between users can be measured on the basis of their spatial or semantic profiles. Spatial profiles gives a measure of user preferences in places, while semantic profiles, on the other hand, is a conceptual measure of user interests.

A. Basic User Profiles

Spatial User Profile

A spatial user profile $P_R(u)$ of a user u is deduced from the set of places that u visited or annotated directly.

$$P_R(u) = \{(r, w(u,r)) | (u,t,r) \in Y, \\ w(u,r) = |\{t \in T : (u,t,r) \in Y\}|\}$$

$w(u,r)$ is the number of tag assignments, where user u assigned some tag t to place r through the action of checking-in or annotation. Hence, the weight assigned to a place simply corresponds to the frequency of the user reference to the place either by checking in or by leaving a tip. We further normalise the weights so that the sum of the weights assigned to the places in the spatial profile is equal to 1. We use $\overline{P_R}$ to explicitly refer to the spatial profile where the sum of all weights is equal to 1, with $\overline{w}(u,r) = \frac{|\{t \in T:(u,t,r) \in Y\}|}{\sum_{i=1}^{n}\sum_{j=1}^{m} |\{t_i \in T:(u,t_i,r_j) \in Y\}|}$, where n and m are the total number of tags and resources, respectively. More simply, $\overline{w}(u,r) = \frac{N(u,r)}{N_T(u)}$, where $N(u,r)$ is the number of tags used by u for resource r, while $N_T(u)$ is the total number of tags used by u for all places.

Correspondingly, we define the tag-based profile of a user; $P_T(u)$ as follows.

Semantic User Profile

A semantic user profile $P_T(u)$ of a user u is deduced from the set of tag assignments linked with u.

$$P_T(u) = \{(t, w(u,t)) | (u,t,r) \in Y, \\ w(u,t) = |\{r \in R : (u,t,r) \in Y\}|\}$$

$w(u,t)$ is the number of tag assignments where user u assigned tag t to some place through the action of checking-in or annotation.

$\overline{P_T}$ refers to the semantic profile where the sum of all weights is equal to 1, with $\overline{w}(u,t) = \frac{N(u,t)}{N_R(u)}$, where $N(u,t)$ is the number of resources annotated by u with t and $N_R(u)$ is the total number of resources annotated by u.

Temporal versions of the profiles can be recorded by considering snapshots of the geo-folksonomy at different points in time. For example, a basic spatio-temporal profile can be represented as follows.

A spatiotemporal (ST) user profile $P_{Rt_c}(u)$ of a user u is deduced from the set of places that u visited or annotated directly.

 ISBN: 978-1-61208-617-0

$$(P_R(u))_{t_c} = \{(r, w(u,r)_{t_c}) | (u,g,r) \in Y, \\ w(u,r)_{t_c} = |\{g_{t_c} \in G : (u,g,r) \in Y\}|$$

$w(u,r)_{t_c}$ is the number of tag assignments in the time slot t_c, where user u assigned some tag g to place r through the action of checking-in or annotation.

B. User Profile Example

Here an example is given of a sample user profile created from the experiment data set used in this work. This user checked in 600 different venues, with associated 400 venue categories.

Figure 2 shows the spatial profile for this user. The dots in the figure represent the weight assigned to place (representing the edge between the user and the place) in the profile. Weights are clustered into 4 equally spaced groups and are mapped to the three noted privacy levels. A simple function for splitting the range of levels is used in this case. However, more intelligent methods for identifying this function can be envisaged, particularly when considering the temporal dimension of the data.

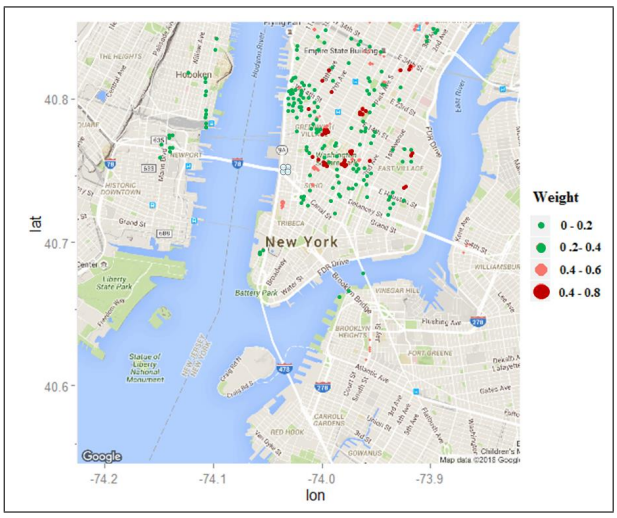

Figure 2. A sample spatial user profile and the corresponding privacy levels.

V. Conclusion

The proliferation of GeoSNs and the large-scale uptake by users suggest the urgency and importance of studying privacy implications of personal information collected by these networks. User profiling is a common method used by online applications to understand users's behaviour and preferences for the purpose of improving their quality of service. However, information implicit in location-based user profiles can reveal personal information about users that can pose real privacy risks. This paper highlights the importance of raising users' awareness of the information they share on GeoSNs. A proposal is made to extend user profiles by explicit representation of the level of risk to personal privacy associated with the information they contain. It is suggested that the level of threat is directly related to the strength of association between the data elements contained in these profiles and that a simple model reflecting this degree of association will be helpful in raising the user awareness of privacy implication of location disclosure. Future work will consider the following:

- Evaluating the proposed methods using realistic sample data sets.
- Exploring different methods of defining the thresholds for the defined levels of risk, e.g. by considering the patterns of association, in addition to the frequency.
- In-depth treatment of the temporal dimension and how to represent dynamic change of the proposed model.

References

[1] S. Mohamed and A. I. Abdelmoty, "Spatio-semantic user profiles in location-based social networks," International Journal of Data Science and Analytics, vol. 4, no. 2, 2017, p. 127142.

[2] F. Alrayes and A. Abdelmoty, "Privacy concerns in location-based social networks," in GEOProcessing 2014: The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services. IARIA, 2014, pp. 105–114.

[3] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating Twitter users," in Proceedings of the 19th ACM international conference on Information and Knowledge Management CIKM '10, 2010, pp. 759–768.

[4] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida, "We know where you live?: privacy characterization of foursquare behavior," in UbiComp '12 Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012, pp. 898–905.

[5] A. Sadilek, H. Kautz, and J. Bigham, "Finding your friends and following them to where you are," in Proceedings of the fifth ACM international conference on Web Search and Data Mining, WSDM '12, 2012, pp. 723–732.

[6] B. Carbunar, M. Rahman, N. Rishe, and J. Ballesteros, "Private location centric profiles for geosocial networks," in Proceedings of the 20th International Conference on Advances in Geographic Information Systems. ACM, 2012, pp. 458–461.

[7] J. Vosecky, D. Jiang, and W. Ng, "Limosa: A system for geographic user interest analysis in twitter," in Proceedings of the 16th International Conference on Extending Database Technology. ACM, 2013, pp. 709–712.

[8] L. Rossi and M. Musolesi, "It's the way you check-in: identifying users in location-based social networks," in Proceedings of the second edition of the ACM conference on Online social networks. ACM, 2014, pp. 215–226.

[9] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie, "You are where you go: Inferring demographic attributes from location check-ins," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015, pp. 295–304.

[10] M. J. Chorley, G. B. Colombo, S. M. Allen, and R. M. Whitaker, "Visiting patterns and personality of foursquare users," in Cloud and Green Computing (CGC), 2013 Third International Conference on. IEEE, 2013, pp. 271–276.

 ISBN: 978-1-61208-617-0