A new framework to define Familiar Strangers in online social networks: spacio-temporal challenges

A proposal from the DARSI (ADSNI) project team:

Charles PEREZ, Babiga BIRREGAH, Marc LEMERCIER, Alain CORPEL, Patrick LACLEMENCE, Eric CHATELET
Charles Delaunay Institute, UMR (CNRS) 6279 STMR, Global Security Department
Centre for Research in Interdisciplinary Studies in Sustainable Development
University of Technology of Troyes, 12 Rue Marie Curie, 10010 Troyes, France. Tel: +33(0) 325717600
(Corresponding author: babiga.birregah@utt.fr)

Abstract

With the exponential growth of social networks of Internet the identity of an individual has become numerical and thus diffused in both time and space dimensions. We cannot ignore that online social network has become a new platform for people to communicate and interact with each other. In our study we propose to focus on the Familiar Stranger notion in order to adapt it to virtual communities such as online social networks and micro-blogging platforms. Familiar Strangers are critical in the understanding of our society (as they can become friend easily) and interesting for a lot of improvement of our community. In this study we propose to adapt and improve the actual formalization of Familiar Stranger applied to Social Networks by injecting both time and spatial constraint. This framework opens on some applications such as identifying the set of Familiar Strangers (FS) of a given micro-blogger in twitter.

Introduction

S. Milgram introduced in 1972 the definition of the concept of Familiar Stranger (FS) (Milgram 1977). Our Familiar Stranger is a person whose face is familiar to us but with whom we do not have direct interaction. An example of our FSs are people who take the same bus with us everyday, who we encounter repeatedly but without direct interaction (e.g. talking). Typically they are not our friends, but they are more likely to become our friends, as explained by S. Milgram. They look familiar to us and they share some common characteristics (hobbies, interests, etc.) with us. Recent studies have attempted to formalize some algorithms for Familiar Strangers detection (Agarwal, et al. 2009). The starting point of these algorithms is the building of a reference set of attributes named "Goal". In this work we present a new framework that revisits the definition of this set to better take into account spatial-temporal constraints. The first section of this paper presents the current definition of familiar stranger and discusses its limitations. Then in section 2 we propose a new definition that takes into account time and space. Finally section 3 presents the challenges of such approach.

1-Familiar Stranger definition on graphs

A social network can be modeled by a graph G(N,E,A) where N is a set of nodes (persons) linked by edges from the set E (relationships). Each node has a subset of attributes from a collection of attributes A={characteristics, hobbies, interests, ...}. Considering a graph G representing a social network the notion of Familiar stranger is defined as follows:

Definition 1[(Agarwal, et al. 2009)**]:**
The set of nodes T_u, FS of a node u respect two conditions:
- Stranger condition (H1):$\forall\, w \in T_u, edge(w,u) = 0$
- Familiar condition (H2):

$\forall\, w \in T_u, A_w \cap \gamma \neq \emptyset \; and \; \gamma \cap A_u = \gamma,$
Where:
$edge(w,u)$ *is a Boolean function revealing the existence of a link between nodes w and u*
A_u *is the set of attributes of a node u*
γ *is the Goal for the detection subset of attributes.*

Depending on the purpose, the detection of Familiar Stranger can be time consuming, especially if one wants to detect all the familiar strangers of any node without any limitation. A first approach can be to fix a Goal that is a subset of attributes (Academic, Arts, Business, News, Political, Geographical location, etc.) and to look for nodes that share the same attributes (H2) but that are not directly connected (H1). Each approach will therefore focalize on the way to search these individuals in the graph with some extensions of the concept. For more details about current techniques of detection and applications of this concept, the reader can refer to (Paulos et Goodman 2004, Perez, et al. 2010). Although the goal (as defined above) can contain attributes in relation with geographical location and activities over time, it does not take into account time and space constrains such as the geographical notions of neighborhood, proximity, distance, and their consistency over time.

2- A New approach to compute FS detection

Since no conditions imply directly a geographical distance between a person and his FS, we propose to improve the definition 1 with the geographical criteria that could approach the sociological definition of S.Milgram. (Liben-Nowell, et al. 2005) analyzes the relation between friendship and location of individuals. His study was about the LiveJournal social network. He concludes that one of the best approaches to link the geographical location with friendship in the network has to take into account a Rank Based parameter. This parameter is computed as the number of persons connected to the network that appear (based on geographical position) between two persons *u* and *v*.
Defining the rank as:

$$rank(u,v) = \{w \in N |\, d(u,w) \; < d(u,v)\,\},$$

where $d(u,v)$ is the Euclidian distance between u and v. The probability for u to interact with v on the online social network is the inverse of the rank:

$$Pr[(u,v) \in E] = \frac{1}{|rank(u,v)|}$$

This probability can be directly related to the location criterion of S.Milgram that is "an individual who is recognized from regular geographical position". We can consider that two persons are more likely to be familiar if they have a high probability to interact. This justifies the fact that two familiar strangers have a high probability to interact and that for example they can share the same way of moving from home to work. If the

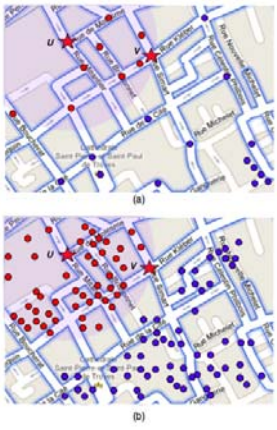

Figure 1. Snapshots of the sub-community of a given network

probability $Pr[(u,v) \in E]$ is high, $|rank(u,v)|$ is low and thus there are few people who at the time of observation network, are located (geographically) between us. Therefore U is more likely to notice the presence (or recognized from regular activities) of V (figure 1). We propose to add a second "familiar condition" (H2.2) to the Familiar Stranger formulation in order to better fit with the original sociological concept.

Definition 2:
The set of Familiar Strangers F_u of a given node u respect three conditions:

- Stranger condition (H1): $\forall w \in F_u, edge(w,u) = 0$
- Familiar conditions:
  (H2.1) $\forall w \in F_u, A_w \cap \gamma \neq \emptyset$ and $\gamma \cap A_u = \gamma$
  (H2.2) $\forall w \in F_u, Pr[(u,w) \in E] \geq K$

Where:
K is the ***familiarity threshold****.*

As illustration, we provide in figure 1, two snapshots of the sub-community of a given network that have been captured at two different periods on the same geographical area. For convenience we have not displayed the links between the individuals. At the opposite of case (b), V shares his regular activities with less people in case (a) and therefore he is more likely to be "identified" by U.

3- Application and Challenges of this approach

We end this study by discussion some key issues emerging in the application of such framework.

Data collection

In order to compute the detection of Familiar Stranger, one needs to access a huge amount of data on the network (Relations,Messages,Location, etc.). The extraction of these data can be done through the APIs (Application Programming Interfaces) provided by some online social networks. These APIs provide a set of methods and documentation in order to extract, store, update (over the time) and analyze data in the limit of privacy level granted by users. The Graph G is built by extracting topology (friends, followers) and attributes are generated from messages, profiles with statistical text-mining methods like tf-idf (Tan 1999, D\"{o}rre, Gerstl et Seiffert 1999). We propose to apply our Familiar Stranger Detection Algorithm on the Graph of the twitter micro-blogging platform. For more information on the twitter API the reader can refer to www.apiwiki.twitter.com. One can notice that some extractions and network analysis have been previously done on twitter and other online social networks (Java, et al. 2007, Mislove, et al. 2007). We choose Twitter because it is used worldwide and it proposes geolocation services. For example when a user updates a new status the status can be geotagged with longitude and latitude of the user location. Figure 2 shows a status extracted in xml format with the

```
1  <user>
2    <id>Permanent unique id referencing a user</id>
3    <name>User specified name for a saved search</name>
4    <...
5    ....
6    ....>
7    <status>
8      <created_at>UTC timestamp of status creation</created_at>
9      <id>Permanent unique id referencing a status</id>
10     <text>Status body</text>
11     <source>Application that sent a status web/iphone/android</source>
12     <geo>Object that may contain GeoRSS or GeoJSON data for a point</geo>
13   </status>
14 </user>
15
```

Figure 2. Xml sample file of a user and status data associated timestamp and geotag on line 8 and 12.

Fast algorithm to compute the familiar stranger

The proposed framework is quite simple, since one can follow the public timeline of twitter through the API (i.e receive new statuses of public users), we propose to apply a FS detection every time that the targeted user (Node u) publishes a new geotagged tweet. Then we store in a list (F_u) the nodes that respect the 3 conditions of definition 2. However one of the obvious difficulties that may appear is the online extraction and time data. Due to combinatorial explosion issues one needs high speed processors and a high storage capacity to perform such an algorithm.

Conclusion

We have proposed a framework to improve the adaptation of the concept of Familiar Stranger to social network of Internet by adding Spatio-Temporal constraints. This proposal, specifically adapted for online social networks, attempts to better fit the sociological requirements as postulated by S. Milgram in his first studies. The resulting detection algorithm can be applied to any social networks or microblogging platforms that support geolocation APIs.

References

Agarwal, Nitin, Huan Liu, Sudheendra Murthy, Arunabha Sen, et Xufei Wang. «A social identity approach to identify familiar strangers in a social network.» 2009.

D\"{o}rre, Jochen, Peter Gerstl, et Roland Seiffert. «Text mining: finding nuggets in mountains of textual data.» ACM, 1999. 398-401.

Java, Akshay, Xiaodan Song, Tim Finin, et Belle Tseng. «Why we twitter: understanding microblogging usage and communities.» ACM, 2007. 56-65.

Liben-Nowell, David, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, et Andrew Tomkins. «Geographic routing in social networks.» *Proceedings of the National Academy of Sciences of the United States of America* 102 (2005): 11623-11628.

Milgram, Stanley. «The individual in a social world.» Chap. The Familiar Stranger: An Aspect of Urban Anonymity, 51-53. Addison-Wesley, 1977.

Mislove, Alan, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, et Bobby Bhattacharjee. «Measurement and analysis of online social networks.» ACM, 2007. 29-42.

Paulos, Eric, et Elizabeth Goodman. «The familiar stranger: anxiety, comfort, and play in public places.» ACM, 2004. 223-230.

Perez, Charles, Babiga Birregah, Patrick Laclémence, et Eric Chatelêt. «A combined approach for suspicious networks detection in graphs.» 2010.

Tan, Ah-hwee. «Text Mining: The state of the art and the challenges.» 1999. 65-70.