Familiar Strangers detection in online social networks

Charles PEREZ, Babiga BIRREGAH, Marc LEMERCIER
ICD (Charles Delaunay Institute) - UMR CNRS STMR 6279
University of Technology of Troyes,
12 rue Marie Curie,
10 010 Troyes Cedex
{charles.perez, babiga.birregah, marc.lemercier}@utt.fr

Abstract—Online social networks and microblogging platforms have collected a huge number of users this last decade. On such platforms, traces of activities are automatically recorded and stored on remote servers. Open data deriving from these traces of interactions represent a major opportunity for social network analysis and mining. This leads to important challenges when trying to understand and analyse these large-scale networks better. Recently, many sociological concepts such as friendship, community, trust and reputation have been transposed and integrated into online social networks. The recent success of mobile social networks and the increasing number of nomadic users of online social networks can contribute to extending the scope of these concepts. In this paper, we transpose the notion of the Familiar Stranger, which is a sociological concept introduced by Stanley Milgram. We propose a framework particularly adapted to online platforms that allows this concept to be defined. Various application fields may be considered: entertainment, services, homeland security, etc. To perform the detection task, we address the concept of familiarity based on spatio-temporal and attribute similarities. The paper ends with a case study of the well-known microblogging platform Twitter.

Index Terms—Familiar Stranger, Social Network Analysis, Nomadism, Online Social Networks, Smartphones, Geo-location, Twitter

I. INTRODUCTION

On social networking sites, a user can create a virtual identity and interact online with other users. By definition, social networking sites can allow the user to: (1) construct a public or semi-public profile within the system, (2) manage a list of other users with whom they share a connection and (3) view and traverse their list of connections [1]. Although this definition only contains basic features, social networking sites have been enriched by many other services such as text, picture and video publishing or geolocation services. With the increase in the number of participants, these networks become more and more complex and can easily integrate a wide range of sociological concepts such as friendship, neighbourhood, community, prestige, etc. Figure 1 highlights some concepts that apply to both the virtual and the physical worlds. Depending on the context, each concept has relatively similar meanings.

Geosocial data represent a good example of the connections between the virtual and physical worlds [2]. Geosocial data can be defined as geolocated or geotagged data that are generated from a social platform. These data represent traces of interactions that help to reconstitute networks in both virtual and physical worlds. A message sent online by a user with a smart device (e.g. smartphone, smart tablet) represents a virtual interaction but also contains geolocation data. Geolocation can allow the detection of physical proximity between users which can then contribute to the construction of physical social networks. The bridge represented by this type of data can help to enlarge the possibilities of applications and should permit a better understanding of the relationship between users’ online and offline lives [3], [4], [5]. Those networks that combine measures of the physical world with human input are often referred to as cyber-physical social networks [6]. In this work, we exploit these cyber-physical social networks by introducing a framework that aims to detect Familiar Strangers (FS).

The concept of the Familiar Stranger was first introduced by S. Milgram in 1972 [7]. Our Familiar Stranger is a person whom we observe regularly but without direct interaction. An example of Familiar Stranger are people who take the same bus as us every day, whom we encounter repeatedly but without direct interaction (e.g. talking with). They are not friends, but they are more likely to become our friends than simple strangers. It is important to emphasise that this concept is sociological and involves several dimensions when adapted to the online sphere (behavioural, spatial and temporal, etc.). The growth of digital social networks offers a good opportunity for the investigation of the different dimensions of this phenomenon with several theoretical and applied challenges. Various application fields may be considered: entertainment,
services, homeland security, etc. [8], [9].

The remainder of this paper is organised as follows. Section II provides some definitions of the Familiar Stranger and discusses their limitations with respect to the original concept introduced by S. Milgram. Section III presents an overview of the multi-dimensional model and its usefulness to address the FS detection. After some preliminary definitions, in Section IV we introduce a new definition of FS in the context of Online Social Networks (OSN). Section V presents an algorithm for detecting FS. An application to Twitter is presented in section VI and the last section concludes this paper.

II. THE FAMILIAR STRANGER: CONCEPT AND RELATED WORKS

The Familiar Stranger concept, as described in the reference literature, has been adapted to many situations. In this section, we present the most relevant contributions for detecting the Familiar Stranger.

In his original experiment, S. Milgram proposes a simple way to highlight the existence of Familiar Strangers in a real-life social network [10]. He proposes to his students of the University of New York that they go to a train station at a particular time in the morning and take pictures of people waiting there. A week later, he asks his students to show their photographs to the people in the picture and ask them who they recognise and with whom they ever interact. This experiment shows that most people are able to clearly identify many individuals with whom they never interact but whose faces are familiar. These individuals are neither friends nor strangers, but Familiar Strangers.

A first approach to automatically detecting such individuals was proposed by [11]. This approach revisits the S. Milgram experiment with the use of Bluetooth devices called Jabberwockies. These devices are worn by individuals or placed in static locations such as bus stops or in train stations. They allow the detection of Familiar Strangers based on both the neighbourhood of an individual (within 20 meters) and proximity to a static set of chosen locations. These locations are chosen based on the places where Familiar Strangers are more likely to meet (e.g. bus stop, train station).

A drawback of this experiment is the need to place specific devices on both individuals and locations that are observed. This implies that the experiment is performed on a specific set of individuals and in a predefined spatio-temporal context.

[12] have presented a mobility model that takes into account the duration and frequency of contacts between people to compute a familiarity metric. This work states that Familiar Strangers are all pairs of individuals that meet regularly but do not spend time with each other. The proposed framework is based on real human mobility datasets and thus fully takes into account the spatio-temporal aspects. However, this approach reduces the problem of FS detection purely to spatio-temporal considerations.

A social network approach based on social identity has also been proposed in order to formalise the Familiar Stranger concept [13]. This approach models a social network as a graph \( G(N, E, A) \), where \( N \) is a set of nodes (individuals) linked by a set \( E \) of connections (relationships). Each node possesses a subset of attributes \( A_u \) from a collection of attributes \( A \). The approach is based on the analysis of these attributes by taking into account proximity as a key factor. These attributes can be generated from the content of social identity and interactions such as phone conversations, sent mails, etc. In this context, the notion of Familiar Stranger is defined based on two requirements. The first requirement (stranger) aims to eliminate all connections of the targeted individual with his set of Familiar Strangers. The second requirement (familiar) ensures that a familiar node possesses a set of attributes that are required and contained in a goal. This goal depends on the individual for whom we are looking for Familiar Strangers. Depending on the purpose, the work of [13] can be time-consuming, especially if the aim is to detect all of the Familiar Strangers of any node without limitations.

Although this approach remains focused on attributes that may contain geographical locations and activities over time, it does not take into account time and space constraints such as the geographical notions of neighbourhood, proximity, distance, and their consistency over time.

The different approaches are classified in Table I based on five important parameters of detection. The social network used for the detection (SN: physical or digital). The spatio-temporal (ST) aspect that is or not considered in the detection. The attribute parameter (Att) that distinguishes the works that may contain geographical locations and activities over time. Finally, we distinguish the detection approaches that rely on the use of a particular device (Dev) from those that do not.

Table II indicates, for each contribution, the meaning given to the Familiar and Stranger aspects of the Familiar Stranger. It is clear that many distinctions exist between the interpretation of the concept depending on the context of the research.

With the recent success of nomadism, geolocation based services and social networks, our study aims to build a detection algorithm that only depends on data generated by users through mobile devices (smartphones, smart tablets). This approach is based on a specific combination of technological devices (smartphones) and usage practice (users that enable geolocation of their statuses).
The above mentioned works regarding Familiar Stranger detection take into account the spatio-temporal or attribute similarity but do not combine these factors in the detection process. They also require software or particular hardware to perform this detection. The approach proposed in this paper takes into account spatio-temporal parameters but also the social proximity (i.e. similarity induced by node attributes) of individuals. To the best of our knowledge, this approach is the first attempt to detect the Familiar Stranger based only on the data generated by mobile social network users.

### III. The Multi-Dimensional Model

The FS definition applied by [13] to social networks is mainly based on the conditions of stranger and familiarity. We propose to keep these conditions and modify them to perform FS detection. The notion of familiarity as defined by Stanley Milgram is multi dimensional: societal, behavioural, spatial, temporal, etc. In this context, the main dimensions to retain in the definition of Familiar Stranger can be found in the two following statements:

**S1.** *Our FS do not have direct interaction with us*
**S2.** *FS are people who seem familiar*

S1 requires two familiar people not to have direct interaction; this means that they should not be friends. S2 requires that they frequent the same neighbourhood regularly and share some common characteristics.

Some researches such as [14] and [15] have proposed a multi-dimensional framework for friend recommendation systems. The main dimensions of these approaches are presented in Figure 2 and the dimensions underlying Familiar Stranger behaviour can be represented and analysed using this model.

The model takes into account the three dimensions that are involved in mobile social networks and that can contribute to Familiar Stranger detection online. Layer I represents the spatio-temporal patterns (i.e. set of positions over time $T = t_0, t_1, t_2$) that are generated by users who have enabled the mobile geolocation service proposed on most of the platforms. The second layer represents the online social graph that reveals the connections between profiles on a given platform (e.g. friends). The third layer represents the data that are generated by users and more specifically the connections among content that can be extracted from these data. In the case of Twitter users, the connections between individuals can be deduced based on the fact that they use the same hashtags (#) or reference the same profiles (@) in their messages.

The concept of Familiar Stranger as addressed in this work hinges on the concepts of strangers (as opposed to friends) contained in layer II. The concept of familiarity is associated with the similarity of attributes in layer III but also on the spatio-temporal similarity of layer I.

### IV. Familiar Stranger Definition in Online Social Networks

In this section, we present and define a set of concepts that are required to identify a FS. First, we introduce the concept of friends and strangers; second, we consider attribute similarity and third, we address the spatio-temporal dimension.

#### A. Friends and strangers in online social networks

The notion of friendship, based on the representation of social ties, exists in both virtual and physical worlds. In this work, we only consider online friendship based on the existence of virtual connections. However, many works have highlighted the correlations between the online and offline social network of a user [16], [17], [14], [18].

Since most of the online social platforms require the creation of a link between two people before they can interact, we propose the identification of strangers based on the existence or not of an edge between them. We can identify two different types of platform: the one that permits the creation of directed edges (e.g. Twitter, LiveJournal) and the one that does not (e.g. Facebook, LinkedIn). In most cases the first category does not need mutual agreement for creating edges, while the second requires the consent of both nodes involved. We introduce some preliminary definitions below.

---

### Table II

**Comparison of Familiar Stranger concept adaptations**

<table>
<thead>
<tr>
<th></th>
<th>Familiar</th>
<th>Stranger</th>
</tr>
</thead>
<tbody>
<tr>
<td>[10]</td>
<td>Observed repeatedly</td>
<td>No direct interactions</td>
</tr>
<tr>
<td>[11]</td>
<td>We repeatedly observe</td>
<td>Do not directly interact with</td>
</tr>
<tr>
<td>[12]</td>
<td>High number of contacts and low contacts duration</td>
<td>Not directly connected</td>
</tr>
<tr>
<td>[13]</td>
<td>Exhibit similarity</td>
<td>Not directly connected</td>
</tr>
</tbody>
</table>

---

**Fig. 2. Multi-layer model**
Definition 1: Friends in OSN
Two nodes \((u, v) \subset N^2\) are friends if and only if:
\[
\{(u, v), (v, u)\} \subset E^2
\]
Where \((u, v)\) is an arc from node \(u\) to node \(v\), \(E\) is the set of edges and \(N\) the set of Nodes.

On Twitter, if a node denoted \(u\) follows a node \(v\) and \(v\) follows \(u\) then \(u\) and \(v\) are considered mutual friends. The friendship link such as on Facebook is considered as an arc from \(u\) to \(v\) (\(u\) invites \(v\) to be friends) and an arc from \(v\) to \(u\) (\(u\) accepts the request) or vice versa. Although some studies consider that two nodes linked by a unilateral arc are friends (e.g. [19]) this is not the case in this work. Basically, we will consider that being strangers is the contrary of being friends.

Definition 2: Strangers in OSN
Two nodes \((u, v) \subset N^2\) are strangers if and only if:
\[
\{(u, v), (v, u)\} \not\subset E
\]
The definition of strangers on platforms with undirected links is straightforward since two unconnected nodes will be considered strangers. On directed platforms such as Twitter two nodes are strangers if they are not connected, or if only one arc exists between them (\(u\) follows \(v\) or \(v\) follows \(u\) but not mutually).

B. Content-based similarity
Many techniques for attributes generation and similarity measures can be found in the literature [20], [21], [22], [23], [24]. The similarity of interests is computed as a content-based attribute similarity between two individuals. Since we require no additional features in our detection approach, this indicator is necessarily based on the information that can be publicly retrieved online. Proportional frequencies is one of the most common and generic way to represent such patterns [25]. To obtain such a representation for a social networking site user, it is necessary to define a set of possible attributes whose values are discrete and finite (e.g. \(A=\{\text{sport, science, literature}\}\)). These attributes can be chosen depending on the expected outcomes or can be adapted to the information retrieved on the platform. Each node \(u \in N\) is represented by a set \(A_u\) of \(n\) attributes whose values belong to \(A\) and that may include duplicates when occurrences are multiple (e.g. \(A_u=\{\text{sport, sport, science}\}\)). This set of elements is basically built up from the occurrences of terms observed in a subset of collected public messages but many alternatives exist [22], [20], [21]. \(A_u\) allows us to build a histogram where each element of \(A\) is associated with its number of occurrences (e.g. \(H_{\text{sport}}=2, H_{\text{science}}=1, H_{\text{literature}}=0\) ). Dividing each of these occurrences by the number \(n\) of attributes permits us to create proportional frequencies. In this work, we only consider information extracted from public messages (i.e. activity traces), since all other types of information (e.g. self descriptions) can be incomplete, false or absent from the platform. We also assume that significant information is more likely to be contained in traces of activities (i.e. chats, talks) than in static content that is often obsolete. Many different similarity measures can be used to compare a couple of users \((u, v)\) from their proportional frequencies \((P, Q)\). As stated in definition 3, we propose to evaluate the interest similarity based on the Jaccard’s coefficient.

Definition 3: Interest similarity
The interest similarity between two nodes \((u, v) \subset N^2\) is defined as:
\[
S_I(u, v) = \frac{\sum_{i=1}^{d} P_iQ_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_iQ_i}
\]
Where \((P, Q)\) are the proportional frequencies of nodes \((u, v)\) and \(d\) is the size of the set of possible attributes.

In Section V we present the integration of this indicator into the Familiar Stranger algorithm.

C. Spatio-temporal metric
With the success of online platforms, users’ spatio-temporal footprints are increasing and become accessible for analysis. Since the beginning of online social networks, the temporal aspect has been naturally identified by a timestamp. This timestamp is determined by the computer or the device’s internal clock. The spatial aspect started to emerge in recent years with the use of geolocation by GPS and Wi-Fi and has been integrated into online social networks [26].

A message can automatically be associated with a time and a location. This geolocation usually requires the use of a smartphone or a smart tablet that offers GPS, and the agreement of the user. When agreement is given, the longitude and latitude of the user are automatically sent within the metadata of his or her messages. This feature is now available on Twitter and Facebook and is the central feature of Mobile Social Softwares (a.k.a. MoSoSo [27]) such as Foursquare. The precision of this geolocation is between 50 and 300 feet.

Many approaches can be performed to deduce a spatio-temporal relation between social network actors. In the literature, many proposed analysis of spatio-temporal relations between actors reposes on a similarity of their spatio-temporal patterns [9], [6], [8]. The top scores of similarities are identified as the best candidates for establishing the relation. The calculation usually refers to a similarity score that is computed between two patterns. Then, a similarity graph is computed where nodes are patterns and weighted links are scores between each pair of users. The final step is to apply an algorithm for estimating the relative importance of the feature in the network [28].

In this work, we investigate a specific heuristic approach identified by Stanley Milgram as individuals who seem familiar. The described assumption requires two individuals to meet each other regularly. This relation is clearly related to the Meeting heuristic as presented in [8] and [29] but with an additional regularity constraint.

We propose the definition of a context-matching function between two persons. This function evaluates when two
people are in the same spatio-temporal frame during a particular time of the experiment \( t \in [0, T] \). The geographical neighbourhood is defined by a radius \( R \) of the circle centred on one of the two individuals analysed.

**Definition 4: Geographical neighbourhood**

The basic geographical neighbourhood of a node \( u \in N \) is defined as follows:

\[
\forall t \in [0, T], \, \, \text{Geo}^+_{t}(u) = \{ v \in N \mid \min_{t+t+\delta t} d(u,v) \leq R \}
\]

Where \( d(u,v) \) is the geographical distance \( R \), \( \delta t \) the spatio-temporal constraints.

Given \( (u,v) \in N^2 \), we define the following Boolean function that identifies when two nodes meet each other.

\[
\text{Geo}^+_{t}(u,v) = \begin{cases} 
1 & \text{if Geo}^+_{t}(u) \\
0 & \text{otherwise}
\end{cases}
\]

**Definition 5: Time meeting list**

We define the spatio-temporal list of Time meeting \( (LT) \) between two nodes \( (u,v) \in N^2 \) as:

\[
\forall t \in [0, T], \, \, \text{LT}(u,v) = \{ t \in [0, T] \mid \text{Geo}^+_{t}(u,v) = 1 \}
\]

From the meeting list, we can compute the average frequency of the meetings between the two individuals and this can be used to reveal a similarity score. However, the requirements expressed by Milgram reveal the importance of the regularity of the meetings. Typically, meeting many times in the same day is not significant if no meetings are recorded after this day. This case can illustrate two individuals who may share the same entertainment over a short period of time and with strong activity online. In this regard, meeting regularly over a long period of time is more significant in our experiment. This case can be illustrated by two individuals waiting at the same bus stop every day. Thus the consistency of the relationship over time is a critical factor. The frequency of meetings is an indicator that can be used to identify whether a relationship is significant or not but it does not necessary reveal its regularity. A high frequency can hide a very high quantity of meetings in a very small time frame and no meetings in any other time frames. On the contrary, a low frequency can hide a regularity of meetings if they are scattered over a larger time frame. We propose the definition of the observed periods as follows:

**Definition 6: Observed periods**

We denote \( LT_i(u,v) \) the \( i^{th} \) element (i.e. meeting) of the set \( LT(u,v) \) and we define the \( i^{th} \) observed periods \( P_i \) between two meetings as follow:

\[
\forall i \in [0, M] \quad P_i(u,v) = |LT_{i+1}(u,v) - LT_i(u,v)|
\]

Where \( M \) is the quantity of meetings between \( u \) and \( v \) during the experiment.

Figure 3 illustrates the spatio-temporal list of meeting times \( (LT_i) \) and the periods between these meetings \( (P_i) \).

We propose the definition of a reference value that represents the period between meetings that is ideal to establish that two people are regularly meeting. We denote this value \( P_{\text{ideal}} \). A reasonable reference value for Familiar Stranger detection could stand between one day and one week, depending on the situation. We then propose a bias indicator that only measures periods of time that exceed the \( P_{\text{ideal}} \). The measure of compliance of the observed meetings with the expected value is detailed below.

**Definition 7: Compliance with ideal**

\[
\forall (u,v) \in N^2, \, \, C(u,v) = \frac{1}{T} \sum_{P_i(u,v) > P_{\text{ideal}}} P_i(u,v) - P_{\text{ideal}}
\]

Meeting so infrequently that the average time between meetings exceeds the ideal value significantly affects the assumption of familiarity.

Finally, we define the spatio-temporal similarity between two people as:

**Definition 8: Spatio-temporal similarity**

\[
\forall (u,v) \in N^2, \, \, S_{ST}(u,v) = 1 - C(u,v)
\]

The spatio-temporal similarity between two users will be equal to one if they meet frequently enough that the time between meeting is under the specified ideal \( (P_{\text{ideal}}) \). The spatio-temporal similarity will be null if no meeting is recorded during the time of the experiment.

**D. An improved definition of Familiar Stranger**

We propose the linear combination of the two similarities defined in the previous subsections. We define the familiarity as a linear weighted sum of Interest Similarity \( (S_I) \) and Spatio-Temporal Similarity \( (S_{ST}) \):

**Definition 9: Familiarity**

\[
\forall (u,v) \in N^2, \, \, F(u,v) = \alpha S_{ST}(u,v) + \beta S_I(u,v)
\]

with \( \alpha + \beta = 1 \)

The weights affecting \( \alpha \) and \( \beta \) depend on the situation analysed and on the expected results. We propose setting \( \alpha = \beta = 0.5 \) in order to correspond well with S. Milgram’s sociological conception of familiarity. However, it can be noted that setting \( \alpha = 0 \) reduces familiarity to interest similarity, and thus to a problem with no spatio-temporal considerations. This approach is then related to [13]. Setting
\( \beta = 0 \) reduces the problem to spatio-temporal considerations and such approaches do not need more than data generated by sensors. The interests of users are not taken into account and the approach is closer to [11] and [12].

We would lastly propose a new definition of Familiar Stranger based on constructed familiarity.

**Definition 10: Familiar Stranger in OSN**

The set of FS of a node \( u \) should respect two conditions:

- **Stranger condition:**
  \( \forall v \in FS_u, u \text{ and } v \text{ are strangers based on def. 2} \)

- **Familiar condition:**
  \( \forall v \in FS_u, F(u, v) = \alpha S_{ST}(u, v) + \beta S_I(u, v) \geq K \)
  where \( K \) is a familiarity threshold

V. Familiar Stranger detection

In this section we propose an algorithm to detect Familiar Strangers of a given individual. It is important to note that geographical constraints permit a significant reduction in the complexity of the problem. The nature of the model means that any individual who does not appear in the neighbourhood of the specified person during the time of the experiment is not analysed by our algorithm. For this reason, there is no need to crawl a full online social network to detect the Familiar Strangers of an individual. However, if a node meets this constraint at least once it will be investigated by our algorithm. The algorithm basically requires us to locate the target, track his or her movements, and analyse his or her interests and those of his or her neighbours. The accuracy of the detection will then mainly depend on the duration of the experiment and on the quality of the data and parameters of the experiment.

The inputs of the algorithm (figure 4) are: the target user \( u \), the coefficients \( \alpha, \beta \) corresponding to the spatio-temporal and interests similarities and the spatio-temporal constraints \( K \) and \( \delta t \). The output of the algorithm is a vector containing the list of top Familiar Strangers candidates for the specified target user.

During the first steps of the algorithm, the Familiar Stranger vector is initialised and attributes are generated for the target node (steps 1-3). These attributes, as described above, can be generated by different processes but on the basis of publicly available data. The algorithm enters a loop that corresponds to the full time span of the experiment. At each specified time, we collect the position of the target user and store the individuals that appear in the same spatio-temporal frame (steps 5-6). All neighbours are potential candidates to be FS and are added to the list of recorded users. When the experiment ends, we calculate the familiarity score, and finally validate the stranger condition (steps 9-12). FS candidates are ranked in the list \( FS_u \) that is returned by the algorithm.

VI. Familiar Stranger detection on Twitter

A. Selection of the platform and candidates

We identify three main requirements in order to be able to perform the FS algorithm on an online social networking platform: (1) user data must be publicly available, (2) a geolocation service should be integrated and (3) target users and candidates should be active on the platform.

Concerning the first requirement, we can only analyse platforms that provide a significant amount of public data. The second requirement is to gain access to spatio-temporal data and this is now possible with online social networks such as Twitter and Facebook and using Mobile Social Software such as Foursquare. The last requirement is mandatory to enlarge the scope and the interest of the experiment regarding the usual FS detection methods.

The analysis and comparison of these three conditions on the main platforms has led us to choose the Twitter microblogging platform to perform our algorithm. Twitter hosts about 500 million accounts which generate more than a million geolocated tweets daily and a large part of those can be collected in real time through the official Twitter streaming API.

We have performed a preselection of profiles who meet requirements (1) and (3). For this purpose, we opened a stream in a specific zone and collected anyone who sent a tweet in this area during a given period of time. For each profile, we collected the last two hundred messages and calculated the frequency of activity, the ratio of geolocated tweets and the number of distinct locations associated with the user’s tweets. The individuals who met given thresholds (i.e. who are nomad users) were selected for the experiment. The threshold set filtered individuals who sent up to ten messages per day and of
whose activity, at least 75% was geolocated. These 200 tweets were necessarily associated with at least 50 distinct positions.

We performed the experiment in the San Francisco Bay area from November 2011 to April 2012. During this time period, a number of fifty thousand users have generated a number of geolocated messages equal to one million. The geographical footprints generated by the sample of these users are represented in Figure 5.

B. Measuring familiarity

In order to perform the interest similarity calculation, we collected the tweets of users and extracted the entities with the help of regular expressions and a term dictionary. We then built the proportional frequencies on the basis of the top measured entities of the sample and applied a similarity coefficient between pairs of users.

Twitter users cannot be located in one place continuously and the only information accessible on their positions is discrete geolocated tweets. To overcome this problem we set a time delay parameter ($\delta t$) that makes each position available during a specific period of time. Combined with the defined radius ($R$) of spatial proximity, this allows us to define with more or less flexibility the spatio-temporal constraints of encounters.

We have generated the spatio-temporal encounter graph for the set of selected individuals for distinct spatio-temporal parameters. On such a graph a link between two nodes means that they met at least once during the time span of the experiment. Figure 6 represents the core component of the spatio-temporal encounter graph for distinct spatio-temporal constraints. This representation gives an idea of the impact of the choice of constraints that can be used for the computation. According to the previous results, broadening the constraints leads to an increase in connections and thus to an increase in candidates based on spatio-temporal similarity between individuals.

The final step is to compute the Familiar Stranger detection. In this work, we have set equal weights for the spatio-temporal and interest similarity indices (i.e. $\alpha = \beta$). This allows better compliance with the Milgram requirements for FS.

The final result in the dataset is presented in the familiarity matrix of Figure 7. On such a matrix, each line and column corresponds to a unique analysed individual. A black pixel represents perfect familiarity while a white pixel represents completely non-familiar users. We can see that the diagonal is black, which shows that the familiarity between an individual and him or herself is always maximal. The matrix is symmetrical because the familiarity between a user $u$ and a user $v$ is equal to the familiarity between $v$ and $u$. We can see that most of the pixels are light, which means that not many users are familiar to each other. The most familiar people linked to an individual correspond to the darkest pixels encountered on the line or on the column of the individual concerned.

C. Familiar Strangers

The top Familiar Stranger candidates are deduced from this figure as the most familiar people that comply with the stranger assumption. The results confirm that the Familiar Stranger, even in the context of a single city, is not a commonly observed
phenomenon. We were able to extract serious candidates with strong similarities for a significant part of the individuals. The selection of the familiarity threshold $K$ remains important for the final selection of candidates. In our work, the $K$ parameter is set up to select the top 10% of familiar people that comply with the stranger assumption. It is important to note that this parameter may depend on each individual’s behaviour, since two different people may have a different number of Familiar Strangers. To identify a good threshold value for $K$ we could ask users to participate in regard to their Familiar Stranger. This could permit the calculation of false positives and false negatives ratios and the adaptation of the threshold with regard to the results.

VII. CONCLUSION

This proposal, specifically adapted for online social networks, attempts to better adapt the FS sociological requirements as postulated by S. Milgram in his first studies. This framework contains spatio-temporal, content-based and online social graph analysis to take into account the multidimensional aspect of the concept. Such a framework has been particularly designed to be applicable to online social networks using geolocation services and an application to Twitter has been proposed. Although the quantity and accuracy of geolocation data is still not sufficient to ensure the exhaustive nature of the results, the growth of mobile social networking applications and the success of smartphones should permit this problem to be resolved in the near future. The approach proposed in this work concerns various application fields such as entertainment, services or homeland security.

VIII. ACKNOWLEDGMENT

This work is part of the CyNIC (Cybercrime, Nomadism and IntelligenCe) CPER project supported by the Champagne-Ardenne region and European Regional Development Fund (ERDF).

REFERENCES


