On the Use of Mobility Data for Discovery and Description of Social Ties

Mitra Baratchi   Nirvana Meratnia   Paul.J.M. Havinga

Department of Computer Science
University of Twente
Enschede, The Netherlands
{m.baratchi, n.meratnia, paul.havinga}@utwente.nl

Abstract— **Ever-increasing emergence of location-aware ubiquitous devices has facilitated collection of time-stamped mobility data. This large volume of data not only provides trajectory information but also information about social interaction between individuals. Unlike trajectory representation and discovery, discovery of social ties and interactions hidden in mobility data has not yet been fully explored. To identify such interaction, social network analysis has been recently used. However, compared with data from emails, phone calls, and messages, which are commonly used for social network analysis, mobility data convey less information about interaction between entities. Therefore, identifying the type of tie between two entities using only mobility data is a great challenge. In this paper, we propose a method for measuring the strength and type of social ties between people only based on their spatio-temporal correlations. Using mutual information metric, we propose utilization of two types of measures for identifying the purpose of being in a certain location. Our experimental results using a location-aware sensing device show that our method can identify different social ties between various entities successfully.**

Keywords—Mobility data; social ties; link description; link prediction; social networks

I. INTRODUCTION

Ever-increasing availability of location-aware technologies such as GPS enabled sensor nodes, and mobile phones has made it possible to collect large volumes of mobility data from various moving entities [1]. Recently people are sharing their mobility data through different location based services such as Foursquare [2], and My tracks [3]. Mining these mobility data provides the opportunity of extracting valuable knowledge regarding behavioral habits such as frequent paths [4]. This knowledge leads to design of various environmental or healthcare services [5]. An interesting topic in analyzing such patterns is finding the existence of social ties between entities and describing the purpose of such links. Social ties between people are formed due to different classes of relationships such as friendship, work related acquaintance, and family membership. These ties convey different information regarding the habits, interactions, and information exchange between the entities connected through them. While friends tend to show more similarity in their interests and habits, more information is exchanged between acquainted people [6]. Therefore, correctly distinguishing between different types of social ties is essential for discovering different communities of people and understanding their interaction [7]. This social tie information can also be exploited in different recommendation systems. Similar to human studies, differentiating between social ties is important in ecological research. Differentiating among social ties between animals provides insights about their evolution and gene flow, maintenance of society, analyzing epidemic patterns, transmission of information, and social learning [8].

The success and reliability of a system for analyzing social ties greatly depends on the data it uses. The types of data which are normally used for discovery of social ties such as emails, phone calls, and data from online social networks, are rich in interaction information. Discovery of social ties using only mobility data is much more challenging due to their limited interaction content. For example, working in the same building does not guarantee that two people are friends or even know each other (two people may work in two different floors of a building). In contrast, the fact that two people post on each other's Facebook wall, send an SMS/email to each other, or talk on the phone indicates existence of a direct interaction between them. Furthermore, imprecision of mobility data acquired by existing technologies, which may be in order of tens of meters, makes the discovery of social ties even more difficult.

Motivated by the fact that entities having social ties to some degree share, spatio-temporal context [9, 10], our contribution in this paper can be summarized as:

- Using mobility data for identifying social ties between people with daily behaviors of different entropies.

- Proposing two information theory-based indicators to measure the correlation between people based on their purpose of visiting different places.

- Identifying the nature of social ties between two people based on the above mentioned indicators.

- Successfully discovering social ties among a group of people carrying a custom designed GPS-enabled sensor node.

The rest of the paper is organized as follows. The problem statement and background information are described in Section II and Section III, respectively. The detailed description of our approach is provided in Section IV. Evaluation results are presented in Section V. Section VI and VII are the related works and conclusions.

II. PROBLEM DEFINITION

Let $D = \{P_1, P_2 \ldots P_N\}$ represent a set of mobility data collected from N number of people. Assume that for each person i there exists a list of time-stamped measurements denoted by $Pi = \{Ts_1, Ts_2, \ldots, Ts_m\}$ over observation duration of m time stamps where Ts_k is a two-dimensional spatial coordinate.

Having D, we are interested in inferring the type of social tie between two people denoted by i and j. We define these social ties to be *acquaintance, friendship (ordinary or buddy), cohabitance,* and *no-relationship*. Acquaintances are people who know each other due to an un-emotional reason. The social tie between colleagues is an example of this type. Friends have special emotional relationship. While ordinary friends only have emotional relationships, buddies have both un-emotional reasons and emotional ones (for instance, they work or study in the same place). By those who cohabit, we mean people who live in the same place. For example, families are a subset of this group. There are also people who do not fall under any of these categories and have no relationships with each other.

III. BACKGROUND

A. Social ties and stay points

A trajectory is composed of transition lines and transition endpoints. Transition endpoints are places that moving entities stay for a considerable amount of time while the transition lines are the paths which they traverse to reach one transition endpoint from another one. Most of people's social ties are formed in places where they stay rather than on paths they traverse to get to these places. Inspired by this observation, we use the mobility data in transition endpoints (stay points) for describing the type of social tie between people. Based on the theory of homophily [11], people tend to build social ties with whom they have more similarity. Therefore, correlation and similarity of people in visit to stay points can be used to describe their social tie information.

In this paper, we use the idea of an information theory-based measure called mutual information for measuring the correlation between people at stay points. In what follows, we will first give some background information on this measure.

B. Mutual information

Information theory-based measures relate the information content of events to their probability of happening. Mutual information [12] (MI) metric measures the dependence of two random variables on each other in terms of the amount of information they share. Given two random variables X and Y with marginal probability mass function $p(x)$ and $p(y)$, and the joint probability mass function of $p(x, y)$, their mutual

information is defined as the relative entropy between the joint distribution and their product distribution as stated below [12]:

$$MI(x, y) = \sum_{x \in X, y \in Y} p(x, y) log \frac{p(x,y)}{p(x).p(y)} \tag{1}$$

The unit of mutual information is bit and if two random variables are independent of each other, their mutual information will be equal to 0 bits.

An extension of mutual information is normalized mutual information [13] (NMI) which scales the above mentioned measure between 0 and 1 where $H(x)$ and $H(y)$ are the entropy of x and y, respectively. Normalized mutual information is calculated as follows:

$$NMI(x, y) = \frac{2MI(x,y)}{H(x)+H(y)} \tag{2}$$

This measure shows how predictable one random variable is from another one and its advantage is quantifying the information content of events by their probability. Thereby, an event which is less likely to happen, contains more information than the one which is more likely. This property can be exploited in distinguishing different types of social ties. A short visit of two friends should bring more information about their social tie than a frequent visit of two colleagues at work. Mutual information metric is extensively studied in different domains of science such as biology [14]. However, its potential to identify the social tie between people from mobility data has not yet been fully explored.

IV. DISCOVERY OF SOCIAL TIES

A. A naïve approach for using mutual information

In this section, we explain that how we can utilize the abovementioned theory to distinguish between people's social ties.

If we assume that mutual information can be used for measuring the similarity between two people i and j, then a naïve idea (close to [15]) will be to first compose an ordered list of time stamped stay points over a period of m timestamps denoted by $SPL = \{x_1 \ldots x_m\}$ for each person. In this list x_b is equal to identifier of stay point a, when the person is at stay point a at timestamp b.

Next, we can apply normalized mutual information on the previously defined ordered list of time stamped stay points of two people (SPL_i, SPL_j) and then take the measured value as an indicator of strength of their social tie. Although simple, the naïve approach suffers from a number of shortcomings highlighted using the following example.

Example 1: Let us consider four people, Alice, Bob, Chuck, and Linda. Alice and Bob are ordinary friends. Bob and Chuck are colleagues and work in the same building. Linda is Chuck's wife and they live together. Every 8 hours, we collect data from places that these four people visit for a period of 3 weeks. Let us consider the activity of visiting places as listed in Table 1-2. All these people go to work every weekday. One weekend Alice and Bob go to a musical and the

next weekend Chuck and Linda go to the same musical. We give an identifier to each visited place (see Table I) and represent the list of time stamped stay points in Table II. Table III shows mutual information measured between these four people on the set of visited places.

TABLE I. LIST OF PLACES

Place	Code
Alice's house	1
Alice's office	2
Bob's house	3
Bob and chuck's office	4
Musical	5
Chuck and Linda's house	6
Linda's office	7

TABLE II. ORDERED LIST OF STAY POINTS DURING A PERIOD OF 3 WEEKS COLLECTED EVERY 8 HOURS.

Person	String
Alice	1211211211211211151111211211211211211111111112112 1121121121111111
Bob	3433433433433433353333433433433433433333333334334 3343343343333333
Chuck	6466466466466466666666466466466466466456666664664 6646646646666666
Linda	6766766766766766666666766766766766766566666667667 6676676676666666

TABLE III. NORMALIZED MUTUAL INFORMATION (NMI) COMPUTED USING EQUATION (2) MEASURED OVER THE SET OF VISITED PLACES SHOWN IN TABLE II.

NMI	Alice	Bob	Chuck	Linda
Alice		1	0.87	0.87
Bob			0.87	0.87
Chuck				1
Linda				

From this simple example, we can conclude that although the normalized mutual information can say how predictable behavior of a person is from another one's, it does not well indicate how people are socially connected. Firstly, All these people have relatively high normalized mutual information with each other while they have different social ties. Furthermore, there is no distinction between the two pairs 'Alice-Bob' and 'Chuck-Linda', while the first pair are ordinary friends who only visited each other once and the second pair live together. Furthermore, Alice has not visited Chuck and Linda in any place but her normalized mutual information with them is as high as the normalized mutual information between Bob and Chuck, who work together.

A disadvantage of this measure is that it does not consider the fact that social tie between people is (mainly) formed due to their co-existence in the same place. The fact that people follow similar daily patterns in distinctive places cause their normalized mutual information to be high. Perhaps this is one of the reasons why normalized mutual information metric has not yet been fully explored in describing the social ties. Another drawback of the naïve approach is that it computes the normalized mutual information using the entire set of visited places without considering the purpose of these visits.

To this end, it is not logical to use predictability of people over a long period of time (their entire life span) to measure their social ties.

B. A heuristic based approach

Having the disadvantages of the naïve approach highlighted, in what follows we present our proposed heuristic approach, which exploits the advantage offered by normalized mutual information measure and at the same time deals with the two above mentioned drawbacks.

Fig. 1 abstractly shows the process of extracting the type of social tie between two people by this approach. The heart of this approach is proposing two indicators which show the interest in common places and interest in person.
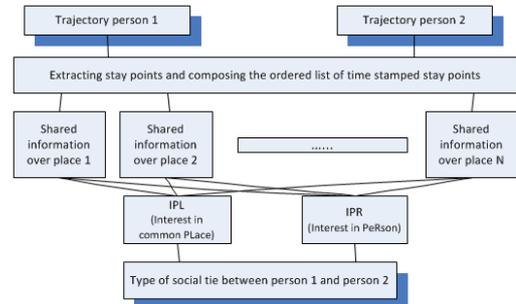
Fig. 1. The process of extracting the type of social ties with the heuristic approach.

Before we define these indicators, we explain two key observations related to the social behavior of people which helped us in defining this approach:

Observation 1: The social tie between people may cause correlation in their visit only to some places and not all the places that they visit (by correlation, we mean simultaneous absence and presence at the place). For example, ordinary friends, may have correlation in visit to places such as cafes, and restaurants (not at work) while people who work in the same place only have correlation in visiting their working place and not in visiting other places. When the visit of two people to their work place is correlated such that they are present at work on the same days, are absent on the same days, and work late on the same days, this is an indication that they may be socially related (e.g. they work on the same project). The fact that these people visit different places when they are absent is not important in deriving any conclusion about their social tie. Therefore, it is better to define the mutual information of people for each single place separately and to ignore the information content of correlation between two people on the entire set of stay points.

Observation 2: People's intention of visiting a place is related to the social tie they have with other people who visit that place. In friendship social ties, being with a friend is one of the primary reasons of visiting a place (two people usually go to a café to be with each other i.e. there is a friendship before going to the café). These correlated visits to different places are normally of low frequency, and short duration. Acquaintances, however, come to know other people as a

consequence of their intention of being in a special place and not because they intend to be with those people (people do not start working to be with their colleagues i.e. there is no acquaintance before starting to work). These correlated visits to one place (for example to a work place) normally happens with high frequency. Therefore, in order to be able to distinguish between different classes of social ties from a set of places, a solution can be using measures that makes a distinction between these two types of interests by considering the frequency of correlated visits (interest in a place (frequently visited places) versus interest in a person (infrequently visited places)).

Based on these two observations, we propose to compute shared information content of two people on each stay point separately and then use the results in computing two indicators which show the interest in person or interest in place. Each indicator will accentuate the correlation of two people in visit to a specific type of stay point. One indicator will emphasize on correlation of two people on the frequently visited stay points and the other one will emphasize on correlation of two people on infrequently visited ones. In other words, the first indicator is an interest indicator for places(*IPL*), implying that places are the focus of the visit, while the second one (*IPR*) is the interest indicator for a person, implying that the person is the reason behind the visit. Using different combination of these two indicators we can discover people's social tie. We continue this section by providing a number of definitions which will later help us in defining these two indicators.

Definition 1:

Shared information content (SI) *of two people i and j for the time they spent at stay point a is defined as:*

$$SI_a(i,j) = log \frac{p_a(i,j)}{p_a(i)p_a(j)} \qquad (3)$$

Where $p_a(i)$ is the marginal probability mass function of person i for being at stay point a while $p_a(i,j)$ is the joint probability mass function of two people i and j for being at stay point a. Please note that this measure is different from the original mutual information. As opposed to mutual information, we only measure the information of simultaneous visits of two people to the same stay point and not the combinations where one or both of the people are absent in the stay point. Therefore, we named this measure as shared information content.

Definition 2:

Normalized shared information (NSI) *content of two people i and j for the time they spent at stay point a is defined as follows:*

$$NSI_a(i,j) = \frac{2*p_a(i,j)*SI_a(i,j)}{H_a(i,j)} \qquad (4)$$

Where $H_a(i,j)$ is computed as follows:

$$H_a(i,j) = log(p_a(i)) + log(p_a(j)) \qquad (5)$$

We use $SI_a(i,j)$ and $NSI_a(i,j)$ to define two indicators for shared information due to the interest in (i) common places and (ii) person. Considering that different set of stay points provide different information about social ties, each of these indicators accentuate on the value of shared information content from the relative important group of stay points (frequently visited stay points vs. infrequently visited stay points). The maximum value of each of these indicators will be 1.

Definition 3:

*The indicator of shared information due to **Interest in common Places (IPL)** for two people i and j over a set of stay points A={$a_1, ..., a_N$ } with ($NSI_a(i,j) > Th$), where Th is a predefined threshold, is defined as:*

$$IPL = \sum_{a \in A} NSI_a(i,j) \qquad (6)$$

As seen in definition 3, for computing $NSI_a(i,j)$, the fraction of time that people spend together at each stay point denoted by $p_a(i,j)$ is scaled by the information they share at that stay point $SI_a(i,j)$. In this way we put more focus on the shared information content over stay points which are visited regularly and have higher $p_a(i,j)$. As mentioned before, regular visit is an indication of interest in a place. The value of *IPL* indicator should be high for people who work, study, or live together. This indicator represent the information that two people share over the whole observation time. The longer the amount of time that two people spend together, the higher the effect of their shared information content on *IPL*. The condition ($NSI_a(i,j) > Th$) is added to prevent highly ranking the low shared information content of two people over a stay point because of the low probability of happening.

Definition 4:

*The Indicator of shared information due to **Interest in Person (IPR)** between two people i and j over a set of stay points A={$a_1, ..., a_N$ } with ($NSI_a(i,j) > Th$) is defined as:*

$$IPR = \frac{p_{min}}{N*SI_{max}} \sum_{a \in A} \frac{SI_a(i,j)}{P_a(i,j)} \qquad (7)$$

In this equation, p_{min} is the lowest probability possible for the visit of a person over a stay point ($1/t_{stop}$) where t_{stop} is the minimum stay time used to extract the stay points and SI_{max} is $log(1/p_{min})$.

IPL indicator scales the shared information content of two people by the fraction of time they spend in that stay point to focus on information shared over frequently visited stay points. In contrast, *IPR* indicator, divides shared information content of two people $SI_a(i,j)$ over the fraction of time they spend together $p_a(i,j)$, to represent the information two people share relative to the time they spend together. *IPR* accentuates the shared information content of two people in stay points which are visited less frequently (with lower $p_a(i,j)$). More specifically, this measure shows that each bit of shared information is extracted from what amount of time that two people spent together. The shorter the amount of time

that two people spend together, the higher the effect of their shared information content on *IPR* will be.

C. Identifying ties based on IPL and IPR indicators

Different combinations of two above indicators can show various types of social ties between two people. The correlation of people in regular visits to their working place will be shown in their high *IPL*. These people may visit some random places together as well. For example, imagine a group of people who work in the same building. These people all have relatively high *IPL* with each other. Among these people, those who work in the same group may spend sometimes out for a social activity. This will cause their *IPR* to increase slightly. This small amount of *IPR* which is measured due to some occasions will help in distinguishing the members of this group from those who work in the same building and have lower probability of acquaintance. Cohabitees or buddies (those who work or study together as well as performing non-frequent activities) might have both high *IPR* and *IPL*. The difference between these two groups is of course distinguishable if the time of day, when activities due to interest in place are performed, is also taken into account. For example, cohabitees will have high *IPL* during night-time and day-time, buddies will have high *IPL* only in day-time. Ordinary friends who do not work or study together may only visit each other once in a while and in some random places. Their correlation in such stay points will cause their *IPR* measure to increase considerably. We summarize combinations of these two indicators with respect to the type of social tie they represent in Table IV.

TABLE IV.   LINK TYPES BASED ON *IPL* AND *IPR* INDICATORS

Link type	IPL	IPR
Acquaintances (with high probability)	High	Low
Acquaintances (with low probability)	High	Zero, Extremely low
Cohabitees	High (Night time)	High
Friends (buddies)	High (Day time)	High
Friends (ordinary)	Low	High
No relation	Zero, High	Zero-low

The pseudocode of our approach to discover the social ties is presented below:

```
Algorithm 1: LinkDescription
INPUT: A data set of trajectories from people D = {P₁, P₂ ... Pₙ}
OUTPUT: A set of link types for each pair L = {L₁,₂ ... L_{N-1,N}}
ALGORITHM:
1: For each (Pᵢ ∈ D) do
2:    Extract the stay points and form the ordered list of stay points SPL
3: For each (Pᵢ, Pⱼ ∈ D) do
4:    For each a ∈ SPL do
5:        measure SIₐ(i, j)
6:    Measure IPL and IPR
7:    Set L_{i,j} based on IPL and IPR
```

Considering *Example* 1, after measuring *IPL* and *IPR* indicators we will have the values presented in Table V and VI. An important point that should be considered in interpreting these results is considering the role of time. As the observation time increases, some of the above values change. As seen, in this example the *IPL* indicator measure of Linda and Chuck is higher than that of Alice and Bob and that of Bob and Chuck. By extending the time of observation this indicator will decrease rapidly for Alice and Bob while it stays the same for Chuck and Linda. The high value of *IPL* for Linda and Chuck is due to their high shared information content for the time they spend at home, while Alice and Bob have different working and living habits. Bob and Chuck have 0.23 *IPL* correlation which is also a good indicator of their correlation at work. This value will stay the same by increasing the observation duration. The indicator of interest in person for two couples (Alice-Bob and Chuck-Linda) is 37 times more than that of Bob and Chuck. If we extend the time of observation and Bob and Chuck keep working with each other without visiting any random places together while they visit random places with their partners, then even this small interest between them will disappear. While by increasing the time of observation, the amount of correlation between the partners will stay the same.

TABLE V.   *IPL* INDICATOR FOR EXAMPLE 1.

IPL	Alice	Bob	Chuck	Linda
Alice		0.01	0	0
Bob			0.23	0
Chuck				0.76
Linda				

TABLE VI.   *IPR* INDICATOR FOR EXAMPLE 1.

IPR	Alice	Bob	Chuck	Linda
Alice		0.14	0	0
Bob			0.003	0
Chuck				0.14
Linda				

Using the combination of these two indicators based on Table IV, we can classify the social tie between these four people. The results are presented in Table VII.

TABLE VII.   SOCIAL TIES OF EXAMPLE 1 BASED ON *IPL* AND *IPR* INDICATORS.

Link	Alice	Bob	Chuck	Linda
Alice		Friend (ordinary)	No-relation	No-relation
Bob			Acquainted (Low probability)	No-relation
Chuck				Cohabitee
Linda				

V.   EVALUATION

A. Dataset

As we did not find any previous work with the purpose of identification of different classes of social ties between people, we do not compare this method with any other. For evaluation of our approach, we used a custom designed GPS node (shown in Fig. 2) which was carried around by a number of colleagues of our research group for a period of 21 days. The study group was composed of two couples (#1&#2 and #4&#5) and three other colleagues, all of whom work in the same building. #1 and #2 mostly visit different places together. They only have

very little difference in working hours because one of them works later most of the time. Couple #4 and #5 have very similar activities at work but normally one of them does some activities such as shopping alone. This couple have visited several random places together. #6 works one day less than the other 5 candidates and lives in another city. The two couples once visited #6 at his home. Candidate #5 is a visiting researcher who does not have any special social tie with the other five people. He has only been at the same stay point with #1 and #2 accidently once in a super market.

The device was set to take a measurement every minute but the data retrieved had a great deal of missing measurements. We extracted the time-stamped GPS measurements in form of latitude and longitude coordinates and used interpolation to replace missing measurements. Considerable amount of noise existed in data as well, especially when the nodes were used inside a building, near the window. We tried to remove such noises by considering a speed threshold. Next, we used the method proposed in [16] to extract the stay points. Each stay point is a group of spatial locations with the maximum radius of 100 meters where people had stayed more than 1 hour. We later merged the stay points which were also closer than 100 meters. Due to high density of places, sometimes one stay point does not necessarily show one specific attraction but a group of them (a shopping center rather than a shop). We extracted 23 places as stay points, 11 of which were at least visited by 2 people.

Fig. 3 compares the 5 candidates in terms of the time they have stayed in 11 stay points. We did not represent the visit to the other 12 stay points as they do not convey information about the social ties (being visited by only one person). Stay points 1, 5 and 8 are the houses of 2 couples and candidate #6 respectively. Stay point 2 is the place the candidates work, stay point 4 is an area in the city center with shopping centers, stay point 10 is a gym and the rest of stay points are random places where at least 2 people had stayed in.

As seen in Fig. 3, the distinction between low and high frequency visited stay points is evident. The only stay point that all the candidates have visited is their working place (stay point 2) and their visit to this stay point has been relatively high but less than their houses.

Fig. 2. The custom designed GPS data logger.

B. Experimental results

After measuring the shared information content of each pair of people based on definition 1 over all stay points, we used the results to measure two indicators of *IPR* and *IPL* for

each pair of candidates. Fig. 4 and Fig. 5 compare the *IPR* and *IPL* measurement results. Since the results of these indicators are symmetric, we have only showed the values over the diagonal line in Fig. 4 and Fig. 5. Furthermore, considering the fact that use of these indicators is meaningless for comparing one person to himself, we also omitted the values on the diagonal line for better visibility.
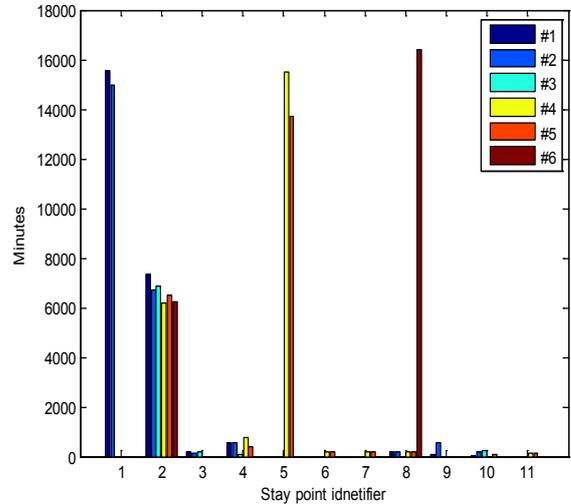

Fig. 3. The amount of time spent in stay point by different candidates.

As seen in Fig. 4, all these candidates have an *IPL* value more than 0.2 bits. This is due to the considerable amount of time they spend in stay point 2. Imagine that we represent the information content of the visit of one person to different places over 21 days in 1 bit. The *IPL* value of 0.2 bits between two people means that they share 0.2 bit of this information together. Therefore, considering that this 0.2 bits represents the information over the whole 21 days this value is relatively high. This measure is higher for the pair (#1&#2) and pair (#4&#5) due to high amount of time they spend together living in the same place (stay point 1 and 5, respectively). The *IPR* indicator shows the information that two people share relative to the time they spend together. In this case the effect of information that people share in frequently visited stay points such as stay point 1,2 and 5 will be degraded. Looking at *IPR* indicator measure in Fig. 5, we see that the level of *IPR* is higher between two couples. This high *IPR* value is due to shared information content in the low probability visited stay points such as stay points 3, 4 and 6-11. The couple (#4&#5) have visited more random stay points than the first couple and naturally their *IPR* is higher. The indicator of interest between #2 and #5 is also high due to their correlated random visit to a gym and their visit to the house of #6.

An interesting point is that, the time that two couples spent with #6 at his house (stay point 8) has only brought information on their social tie with each other while as seen in Fig. 5, the *IPR* value of #6 with #1, #2, #4, #5 is still low. Looking at Fig. 3, it is seen that, the probability of #6 being in his house (stay point 8) is considerably high so this is not a good indicator of his interest in seeing the two couples. This

seems logical as it was also possible that two couples were in somebody else's house who lived in the same building where #6 lives. However, if #6 had spent more time with any of these 4 candidates in a random stay point, then the shared information was more helpful in identifying their social tie.
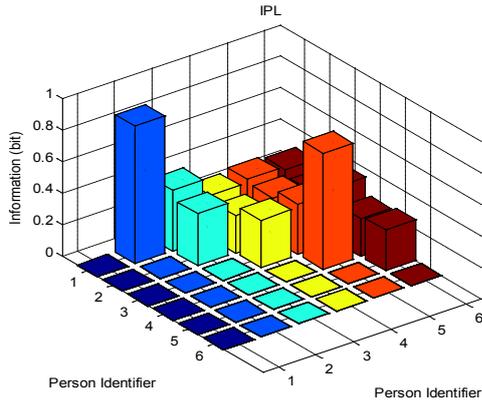

Fig. 4 . Information shared due to Interest in common places (*IPL*).
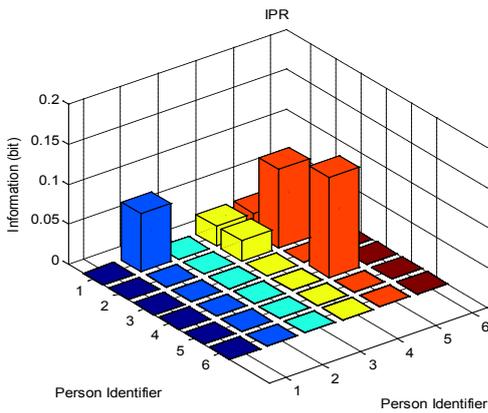
Fig. 5. Information shared due to Interest in person (*IPR*).

Another interesting point is that although #3 has spent some time in stay points 3, 4, 10 which other candidates have also visited, the value of second indicator does not show distinguishable *IPR* between this candidate and the others. The reason is that, the visit of this candidate has either happened in other times or the result of their normalized shared information was under a threshold which could be considered as an accidental co-occurrence.

The results shown in Fig. 4 and Fig. 5 show that, we can clearly make a distinction between the colleagues who only work at the same stay point with those who have performed social activity outside work.

VI. RELATED WORK

Most of the researches performed previously in analyzing and predicting social ties define binary associations (existence of a social tie versus absence of a social tie) between social entities [17-20]. These researches do not address the type of social tie between these entities. There are a number of previous works with the focus on link description based on the

data from online social networking websites [21-23], and heterogonous networks [24]. The description and prediction of social ties in these works are normally based on a number of links formed previously by the user input. In contrast, no prior information on links is available when mobility data is used. Furthermore, these works benefit from the amount of different types of "interaction" content available for each individual (number of photos tagged, number of wall posts, etc.).

More recently, identifying social ties using mobility data has been proposed. In the research presented in [25] existence of social tie between two people is inferred from the semantic similarity of their trajectories without interpreting the type of social tie. In [26], authors have used communication and mobility data from mobile phone records for finding friendships. They have used four factors, (i) campus/off campus, (ii) daytime/nighttime, (iii) weekend proximity, and (iv) phone communication for measuring the social ties. This approach is however, specific to social ties in one affiliation and does not work for people with different spatial domains. Furthermore, not all people have the same working habits that are dependent on day of the week. A number of collocation metrics are introduced in [27] to be used along with mobile phone data to measure the strength of social ties between people. These collocation metrics are based on the probability of two people being in the same place. Using mobile phone data can bring additional interaction content to the analysis process as used in [26, 27].

In this work, we consider extracting social information by just using mobility data. The major difference between our work and previous researches which have considered different social ties [21-24] is the way we describe the links. These existing works relate the strength of the tie to the strength of friendship with strong ties showing strong friendships and weak ties showing acquaintance. We however, make clear distinction between different classes of social ties, namely friends, acquaintances, and cohabitees by analyzing two indicators. Another major difference is that, all existing solutions focus on the value of joint probability of two people visiting places for measuring the strength of their social tie. As stated before, the joint probability of two people in visiting one place might be higher in acquaintances who work together than in friends who have different working and living habits. Therefore, this measure is not a good indicator of social tie nature. We consider use of the mutual information content of people over places with both high and low frequency of visit. To the best of our knowledge, there is only one previous research [15] which has considered using the mutual information content in measuring the social ties between people. In the work presented in [15], mutual information is used to measure the social tie strength in bi-partite networks. There are however, two major differences between our work and this work. Firstly this method does not make distinction between the type of tie while we proposed the *IPL* and *IPR* indicators to describe different classes of ties between people. The second difference is using location data. Authors of [15] consider measuring the social tie between people from a non-

location dataset of people who participated in selected one-time events. As shown in the naïve approach, such metric is not applicable in inferring social tie information of people form their location data.

VII. Conclusion

In this paper we proposed a method for differentiating between different types of social ties between people using mobility data. We defined shared information content metric based on mutual information to extract the information content which shows correlation between two people at a certain location. Next, we used the information content from each place in computing two indicators for each pair of people which showed the information content due to their interest in common places and their interest in each other. We further used these indicators to identify the type of social ties between two people.

The proposed indicators are useful in identifying the existence of a tie between two social entities as well as description of the type of tie using mobility data. By relating these two indicators with the social ties between other moving entities (e.g. animals), these metrics can also help in identifying different types of ties between those entities. Furthermore, these indicators can be used as an additional tool in improving the performance of online location-based social networks.

In primary experiments by examining a small dataset with known social ties we showed that it is possible to use this type of data for the purpose of social tie identification. However, we expect that in a larger dataset with more complicated mobility patterns and sparser data, unexpected errors happen. Therefore, our future work entails validating these results with a larger scale dataset.

REFERENCES

[1] M. Baratchi, N. Meratnia, P. Havinga, A. Skidmore, and B. Toxopeus, "Sensing Solutions for Collecting Spatio-Temporal Data for Wildlife Monitoring Applications: A Review," Sensors, vol. 13, pp. 6054-6088, 2013.

[2] Foursquare. Available: https://foursquare.com/, Accessed on 23 May 2013.

[3] My Tracks. Available: http://www.google.com/mobile/mytracks/, Accessed on 23 May 2013.

[4] M. Baratchi, N. Meratnia, and P. J. M. Havinga, "Finding frequently visited paths: dealing with the uncertainty of spatio-temporal mobility data," in IEEE ISSNIP, Melbourne, Australia, 2013.

[5] S. Aflaki, N. Meratnia, M. Baratchi, and P. J. M. Havinga, "Evaluation of Incentives for Body Area Network-based HealthCare Systems," presented at the IEEE ISSNIP, Melbourne, Australia, 2013.

[6] M. Granovetter, "The Impact of Social Structure on Economic Outcomes," Journal of Economic Perspective, vol. 19, p. 17, 2005.

[7] D. Cai, Z. Shao, X. He, X. Yan, and J. Han, "Community Mining from Multi-relational Networks," in Knowledge Discovery in Databases: PKDD 2005. vol. 3721, A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 445-452.

[8] M. K. Marsh, S. R. McLeod, M. R. Hutchings, and P. C. L. White, "Use of proximity loggers and network analysis to quantify social interactions in free-ranging wild rabbit populations," Wildlife Research, vol. 38, pp. 1-12, 2011.

[9] D. Mok, B. Wellman, and J. Carrasco, "Does Distance Matter in the Age of the Internet?," Urban Studies, vol. 47, pp. 2747-2783, November 1, 2010 2010.

[10] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA, 2011.

[11] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annual Review of Sociology, vol. 27, pp. 415-444, 2001.

[12] T. M. Cover and A. J. Thomas, Elements of Information Theory. New York: Wiley, 1991.

[13] T. O. Kvalseth, "Entropy and correlation: Some comments," IEEE Transactions on Systems, Man and Cybernetics, vol. 17, pp. 217-219, 1987.

[14] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, "Species independence of mutual information in coding and noncoding DNA," Physical Review E, vol. 61, pp. 5624-5629, 2000.

[15] W. Dong, "Mutual information: inferring tie strength and proximity in bipartite social network data with non-metric associations," Master of science, University of Illinois at Urbana-Champaign, 2011.

[16] A. T. Palma and E. al., "A clustering-based approach for discovering interesting places in trajectories," presented at the In Proc. 2008 ACM symposium on Applied computing, Fortaleza, Ceara, Brazil, 2008.

[17] H. Kashima and N. Abe, "A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction," in Data Mining, 2006. ICDM '06. Sixth International Conference on, 2006, pp. 340-349.

[18] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," presented at the Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA, 2003.

[19] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA, 2011.

[20] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media," ACM Trans. Web, vol. 6, pp. 1-33, 2012.

[21] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 2009.

[22] J. Zhuang, T. Mei, S. C. H. Hoi, X.-S. Hua, and S. Li, "Modeling social strength in social media community via kernel-based learning," presented at the Proceedings of the 19th ACM international conference on Multimedia, Scottsdale, Arizona, USA, 2011.

[23] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," presented at the Proceedings of the fourth ACM international conference on Web search and data mining, Hong Kong, China, 2011.

[24] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogenous networks," presented at the Proceedings of the fifth ACM international conference on Web search and data mining, Seattle, Washington, USA, 2012.

[25] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring social ties between users with human location history," Journal of Ambient Intelligence and Humanized Computing, pp. 1-17, 2012/12/01 2012.

[26] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," Proceedings of the National Academy of Sciences, vol. 106, pp. 15274-15278, September 8, 2009 2009.

[27] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA, 2011.