The role of location and social strength for friendship prediction in location-based social networks

Jorge Valverde-Rebaza, Mathieu Roche, Pascal Poncelet, Alneu de Andrade Lopes

The role of location and social strength for friendship prediction in location-based social networks

Jorge C. Valverde-Rebaza[a,*], Mathieu Roche[c,1], Pascal Poncelet[b], Alneu de Andrade Lopes[a]

[a] *Department of Computer Science, ICMC, University of São Paulo*
C.P. 668, CEP 13560-970, São Carlos, SP, Brazil
[b] *LIRMM, University of Montpellier*
860 rue de Saint Priest F-34095 Montpellier, France
[c] *CIRAD - TETIS*
F-34398, Montpellier, France
[d] *TETIS, University of Montpellier, AgroParisTech, Cirad,*
CNRS, Irstea, Montpellier, France

Abstract

Recent advances in data mining and machine learning techniques are focused on exploiting location data. There, combined with the increased availability of location-acquisition technology, has encouraged social networking services to offer to their users different ways to share their location information. These social networks, called location-based social networks (LBSNs), have attracted millions of users and the attention of the research community. One fundamental task in the LBSN context is the friendship prediction due to its role in different applications such as recommendation systems. In the literature exists a variety of friendship prediction methods for LBSNs, but most of them give more importance to the location information of users and disregard the strength of relationships existing between these users. The contributions of this article are threefold, we: 1) carried out a comprehensive survey of methods for friendship prediction in LBSNs and proposed a taxonomy to organize the existing methods; 2) put forward a proposal of five new methods addressing gaps identified in our survey while striving to find a balance between optimizing computational resources and improving the predictive power; and 3) used a comprehensive evaluation to quantify the prediction abilities of ten current methods and our five proposals and selected the top-5 friendship prediction methods for LBSNs. We thus present a general panorama of friendship prediction task in the LBSN domain with balanced depth so as to facilitate research and real-world application design regarding this important issue.

Key words: Location-based social networks, Link prediction, Friendship recommendation, Human mobility, User behavior.

*Corresponding author
Email addresses: jorge.carlos14@gmail.com (Jorge C. Valverde-Rebaza)

1. Introduction

In the real world, many social, biological, and information systems can be naturally described as complex networks in which nodes denote entities (individuals or organizations) and links represent different interactions between these entities. A social network is a complex network in which nodes represent people or other entities in a social context, whilst links represent any type of relationship among them, like friendship, kinship, collaboration or others [1].

With the growing use of Internet and mobile devices, different web platforms such as Facebook, Twitter and Foursquare implement social network environments aimed at providing different services to facilitate the connection between individuals with similar interests and behaviors. These platforms, also called as online social networks (OSNs), have become part of the daily life of millions of people around the world who constantly maintain and create new social relationships [2, 3]. OSNs providing location-based services for users to check-in in a physical place are called location-based social networks (LBSNs) [4, 5, 6, 7].

One fundamental problem in social network analysis is link prediction, which aims to estimate the likelihood of the existence of a future or missing link between two disconnected nodes based on the observed network information [8, 9, 10, 11]. In the case of LBSNs, the link prediction problem should be dealt with by considering the different kinds of links [12, 2, 13]. Therefore, it is called *friendship prediction* when the objective is to predict social links, i.e. links connecting users, and *location prediction* when the focus is to predict user-location links, i.e. links connecting users with places [6, 14, 15].

Since location information is a natural source in LBSNs, several techniques have been proposed to deal with the location prediction problem [12, 2]. However, to the best of our knowledge no studies have analyzed the performance of friendship prediction methods in the LBSN domain.

In this paper, we review existing friendship prediction methods in the LBSN domain. Moreover, we organize the reviewed methods according to the different information sources used to make their predictions. We also analyze the different gaps between these methods and then propose five new friendship prediction methods which more efficiently explore the combination of the different identified information sources. Finally, we perform extensive experiments on well-known LBSNs and analyze the performance of all the friendship prediction methods studied not only in terms of prediction accuracy, but also regarding the quality of the correctly predicted links. Our experimental results highlight the most suitable friendship prediction methods to be used when real-world factors are considered.

The remainder of this paper is organized as follows. Section 2 briefly describes the formal definition of an LBSN. Section 3 formally explains the link prediction problem and how it is dealt with in the LBSN domain. This section also presents a survey of different friendship prediction methods from the literature. Section 4 presents our proposals with a detailed explanation on how

they exploit different information sources to improve the friendship prediction accuracy. Section 5 shows experimental results obtained by comparing the efficiency of existing friendship prediction methods against our proposals. Finally, Section 6 closes with a summary of our main contributions and final remarks.

2. Location-Based Social Networks

A location-based social network (LBSN), also referred to as *geographic social network* or *geo-social network*, is formally defined as a specific type of social networking platform in which geographical services complement traditional social networks. This additional information enables new social dynamics, including those derived from visits of users to the same or similar locations, in addition to knowledge of common interests, activities and behaviors inferred from the set of places visited by a person and the location-tagged data generated during these visits [2, 16, 12, 6, 17].

Formally, we represent an LBSN as an undirected network $G(V, E, \mathcal{L}, \Phi)$, where V is the set of users, E is the set of edges representing social links among users, $\mathcal{L}$ is the set of different places visited by all users, and Φ is the set of check-ins representing connections between users and places. This representation reflects the presence of two types of nodes: users and locations, and two kinds of links: user-user (social links) and user-location (check-ins), which is an indicator of the heterogeneity of LBSNs [2, 18, 19]

Multiple links and self-connections are not allowed in the set E of social links. On the other hand, only self-connections are not allowed in the set Φ of check-ins. Since a user can visit the same place more than once, the presence of multiple links connecting users and places is possible if a temporal factor is considered. Therefore, a check-in is defined as a tuple $\theta = (x, t, \ell)$, where $x \in V$, t is the check-in time, and $\ell \in \mathcal{L}$. Clearly, $\theta \in \Phi$ and $|\Phi|$ defines the total number of check-ins made by all users.

3. Link Prediction

In this section, we formally describe the link prediction problem and how this mining task is addressed in the LBSN domain. Moreover, we also review a selected number of friendship prediction methods for LBSNs.

3.1. Problem Description

Link prediction is a fundamental problem in complex network analysis [1, 9], hence in social network analysis [11, 20, 21, 22]. Formally, the link prediction problem aims at predicting the existence of a future or missing link among all possible pairs of nodes that have not established any connection in the current network structure [8].

Consider as *potential link* any pair of disconnected users $x, y \in V$ such that $(x, y) \notin E$. U denotes the universal set containing all potential links between pairs of nodes in V, i.e. $|U| = \frac{|V| \times (|V|-1)}{2}$ since G is an undirected network.

Also consider a *missing link* as any potential link in the set of nonexistent links $U - E$. The fundamental link prediction task here is thus to detect the missing links in the set of nonexistent links, while scoring each link in this set. Thus, a *predicted link* is any potential link that has received a score above zero as determined by any link prediction method. The higher the score, the more likely the link will be [1, 8, 10].

From the set of all predicted links, L_p, obtained by use of a link prediction method, we assume the set of *true positives* (TP) as all correctly predicted links, and the set of *false positives* (FP) as the wrongly predicted links. Thus, $L_p = TP \cup FP$. Moreover, the set of *false negatives* (FN) is formed by all truly new links that were not predicted.

Therefore, evaluation measures as the *imbalance ratio*, defined as $IR = \frac{|L_p|}{|TP|}$, *precision*, defined as $P = \frac{|TP|}{|TP|+|FP|}$, and *recall*, defined as $R = \frac{|TP|}{|TP|+|FN|}$, can be used as well as the harmonic mean of precision and recall, the *F-measure*, defined as $F_1 = 2 \times \frac{P \times R}{P+R}$ [23, 24]. However, most of the researches in link prediction consider that these evaluation measures do not give a clear judgment of the quality of predictions. For instance, a right predicted link could not be considered as a true positive if any link prediction method gives it a low score. To avoid this fact, two standard evaluation measures are used, AUC and precisi@L [6, 9].

The *area under the receiver operating characteristic curve* (AUC) is defined as $AUC = \frac{n_1 + 0.5 \times n_2}{n}$, where from a total of n independent comparisons between pairs of positively and negatively predicted links, n_1 times the positively predicted links were given higher scores than negatively predicted links whilst n_2 times they were given equal scores. If the scores are generated from an independent and identical distribution, the AUC should be about 0.5; thus, the extent to which AUC exceeds 0.5 indicates how much better the link prediction method performs than pure chance. On the other hand, precisi@L is computed as $precisi@L = \frac{L_r}{L}$, where L_r is the number of correctly predicted links from L top-ranked predicted links.

3.2. Friendship Prediction in LBSNs

LBSNs provide services to their users to enable them to take better advantage of different resources within a specific geographical area, so the quality of such services can substantially benefit from improvements in link prediction [5, 6]. Therefore, considering the natural heterogeneity of LBSNs, the link prediction problem for this type of network must consider its two kinds of links [12, 2], i.e. friendship prediction involves predicting user-user links [25, 26, 6] whilst location prediction focuses on predict user-location links [14, 15, 27].

Friendship prediction is a traditional link prediction application, providing users with potential friends based on their relationship patterns and the social structure of the network [3]. Friendship prediction have been widely explored in LBSNs since it is possible to use traditional link prediction methods, such as common neighbors, Adamic-Adar, Jaccard, resource-allocation and preferential attachment, which are commonly applied and have been extensively studied

in traditional social networks [9, 10]. However, as location information is a natural resource in LBSNs, different authors have proposed friendship prediction methods to exploit it. Therefore, some methods use geographical distance [28], GPS and/or check-in history [29], location semantics (tags, categories, etc.) [30] and other mobility user patterns [18, 24, 31] as information sources to improve the effectiveness of friendship prediction in LBSNs.

The friendship prediction task in LBSNs is still an open issue where there are constant advances and new challenges. Furthermore, the importance of the friendship prediction task is not only due to its well known application in friendship recommendation systems, but also because it opens doors to new research and application issues, such as companion prediction [32], local expert prediction [33, 34, 35], user identification [36, 37] and others.

3.3. Friendship Prediction Methods for LBSNs

Most existing link prediction methods are based on specific measures that capture similarity or proximity between nodes. Due to theirs low computational cost and easy calculation, link prediction methods based on similarity are candidate approaches for real-world applications [9, 38, 17].

Although there is abundant literature related to friendship prediction in the LBSN context, there is a lack of well organised and clearly explained taxonomy of existing methods in the current literature. For the sake of clearly arranging these existing methods, this study proposes a taxonomy for friendship prediction methods for LBSNs based on the information sources used to perform their predictions. Figure 1 shows the proposed taxonomy.

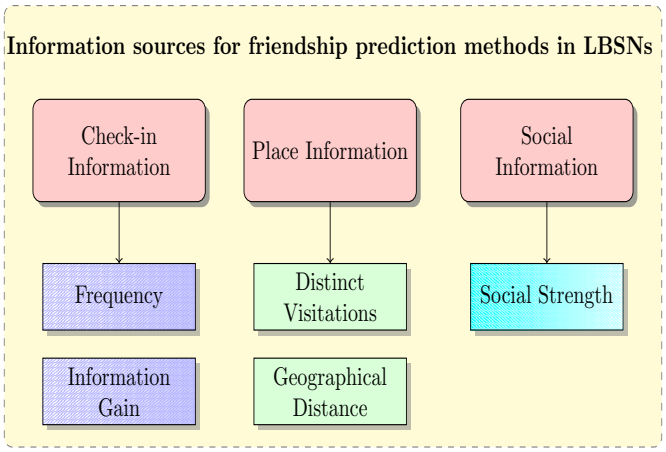


Figure 1: Information sources and the different similarity criteria used by existing methods to perform friendship prediction in LBSNs.

Friendship prediction methods for LBSNs use three information sources to compute the similarity between a pair of users: check-in, place, and social information. In turn, each information source has specific similarity criteria. There-

fore, *methods based on check-in information* explore the frequency of visits at specific places and information gain. *Methods based on place information* commonly explore the number of user visits, regardless of frequency, to distinct places as well as the geographical distance between places. Finally, *methods based on social information* explore the social strength among users visiting the same places.

Here, we will give a systematic explanation of popular methods for friendship prediction in LBSNs belonging to each one of the proposed categories.

3.3.1. Methods based on Check-in Information

User mobility behaviors can be analyzed when the time and geographical information about the location visited are record at check-ins. The number of check-ins may be an indicator of users' preference for visiting a specific type of places and therefore, the key to establishing new friendships. Two of the most common similarity criteria used by methods based on check-in information are the check-in frequency and information gain.

Methods based on check-in frequency consider that the more check-ins at same places have made two users the more likely they will establish a friendship relationship. Some representative methods based on check-in frequency are the collocation, distinct collocation, Adamic-Adar of places, preferential attachment of check-ins, among others [39, 18, 40]. Bellow, we present the definition of two well-known friendship methods for LBSNs based on check-in frequency.

Collocation (Co). This is one of the most popular methods based on the check-in frequency. The collocation method, also referred to as the *number of collocations* or *common check-in count*, expresses the number of times that users x and y visited some location at the same period of time. Thus, for a pair of disconnected users x and y, and considering a temporal threshold of time, $\tau \in \mathbb{R}$, the Co method is defined as:

$$s^{Co}_{x,y,\tau} = |\Phi_{Co}(x, y, \tau)|, \tag{1}$$

where, $\Phi_{Co}(x, y, \tau) = \{(x, y, t_x, t_y, \ell) \mid (x, t_x, \ell) \in \Phi(x) \ \wedge \ (y, t_y, \ell) \in \Phi(y) \ \wedge \ |t_x - t_y| \leq \tau\}$, is the set of check-ins made by both users x and y at the same place and over the same period of time, and $\Phi(x) = \{(x, t, \ell) \mid x \in V : \ (x, t, \ell) \in \Phi\}$ is the set of check-ins made by the user x at different places.

Adamic-Adar of Places (AAP). This is based on the traditional Adamic-Adar method but considering the number of check-ins of common visited places of users x and y. Thus, for a pair of users x and y, AAP is computed as:

$$s^{AAP}_{x,y} = \sum_{\ell \in \Phi_{\mathcal{L}}(x,y)} \frac{1}{\log |\Phi(\ell)|}, \tag{2}$$

where $\Phi_{\mathcal{L}}(x, y) = \Phi_{\mathcal{L}}(x) \cap \Phi_{\mathcal{L}}(y)$ is the set of places commonly visited by users x and y, $\Phi_{\mathcal{L}}(x) = \{\ell \mid \forall \ell \in \mathcal{L} : (x, t, \ell) \in \Phi(x)\}$ is the set of distinct

places visited by user x, and $\Phi(\ell) = \{(x,t,\ell) \mid \ell \in \mathcal{L} : (x,t,\ell) \in \Phi\}$, is the set of check-ins made by different users at location ℓ.

Although the number of check-ins may be a good indicator for the establishment of friendship between users, the fact that they have many check-ins at visited places may, on the contrary, reduce their chances of getting to know each other. To avoid this situation, some researchers have used the information gain of places as a resource to better discriminate whether a certain place is relevant to the formation of social ties between its visitors [39, 23, 25, 40]. Some methods based on information gain of places are min entropy, Adamic-Adar of entropy, location category and others. Bellow, we present two well-known friendship methods for LBSNs based on information gain.

Adamic Adar of Entropy (AAE). This also applies the traditional Adamic-Adar method while considering the place entropy for common locations of a pair of users x and y. Therefore, the AAE method is defined as:

$$s_{x,y}^{AAE} = \sum_{\ell \in \Phi_{\mathcal{L}}(x,y)} = \frac{1}{\log \mathcal{E}(\ell)}, \tag{3}$$

where $\mathcal{E}(\ell) = -\sum_{x \in \Phi_V(\ell)} q_{x,\ell} \log(q_{x,\ell})$ is the place entropy of location ℓ, $q_{x,\ell} = \frac{|\Phi(x,\ell)|}{|\Phi(\ell)|}$ is the relevance of check-ins of a user, $\Phi(x,\ell) = \{(x, t, \ell) \mid (x,t,\ell) \in \Phi(x) \ \wedge \ \ell \in \Phi_{\mathcal{L}}(x)\}$ is the set of check-ins of a user x at location ℓ, and $\Phi_V(\ell) = \{x \mid (x,t,\ell) \in \Phi(x) \ \wedge \ \ell \in \Phi_{\mathcal{L}}(x)\}$ is the set of visitors of location ℓ.

Location Category (LC). This calculates the total sum of the ratio of the number of check-ins of all locations visited by users x and y to the number of check-ins of users x and y at these locations while disregarding those with a high place entropy. Therefore, considering an entropy threshold $\tau_{\mathcal{E}} \in \mathbb{R}$, the LC method is defined as:

$$s_{x,y}^{LC} = \sum_{\ell \in \Phi_{\mathcal{L}}(x) \ \wedge \ \mathcal{E}(\ell) < \tau_{\mathcal{E}}} \ \sum_{\ell' \in \Phi_{\mathcal{L}}(y) \ \wedge \ \mathcal{E}(\ell') < \tau_{\mathcal{E}}} \frac{|\Phi(\ell)| + |\Phi(\ell')|}{|\Phi(x,\ell)| + |\Phi(y,\ell')|}. \tag{4}$$

3.3.2. Methods based on Place Information

Friendship prediction methods based on place information consider that locations are the main elements on which different similarity criteria can be used. Two of the most common similarity criteria used by methods based on place information are the number of distinct visitations and geographical distance.

Methods based on distinct visitations consider specific relations among the different visited places by a pair of user as the key to compute the likelihood of a future friendship between them. Some representative methods based on distinct visitations at specific places are the common location, jaccard of places, location observation, preferential attachment of places, among others [39, 41, 42, 23]. Below, we present two of the most representative friendship prediction methods for LBSNs based on distinct visitations.

Common Location (CL). This is inspired by the traditional common neighbor method and constitute the simplest and most popular method based on distinct visitations at places to determine the homophily among pairs of users. Common location method, also known as *common places* or *distinct common locations*, expresses the number of common locations visited by users x and y. Thus, CL is defined as:

$$s_{x,y}^{CL} = |\Phi_{\mathcal{L}}(x,y)|, \tag{5}$$

where, $\Phi_{\mathcal{L}}(x,y) = \Phi_{\mathcal{L}}(x) \cap \Phi_{\mathcal{L}}(y)$ is the previously defined set of common visited places of a pair of users x and y.

Jaccard of Places (JacP). This is inspired by the traditional Jaccard method. Jaccard of places method is defined as the fraction of the number of common locations and the number of locations visited by both users x and y. Therefore, JacP is computed as:

$$s_{x,y}^{JacP} = \frac{|\Phi_{\mathcal{L}}(x,y)|}{|\Phi_{\mathcal{L}}(x) \cup \Phi_{\mathcal{L}}(y)|}. \tag{6}$$

On the other hand, since different studies have shown the importance of geographical or geospatial distance in the establishment of social ties, many authors have proposed to exploit this fact to improve friendship prediction. Some of the most representative methods based on geographical distance are the min distance, geodist, weighted geodist, Hausdorff distance and adjusted Hausdorff distance [23, 40, 28, 43]. Below, we discuss two representative friendship prediction methods for LBSNs based on geographical distance.

GeoDist (GeoD). This method is the most common of those based on geographical distance. Consider as the "home location" of user x, ℓ_x^h, relative to the most checked-in place. Therefore, GeoD computes the geographical distance between the home locations of users x and y. Thus, GeoD is calculated as:

$$s_{x,y}^{GeoD} = dist(\ell_x^h, \ell_y^h), \tag{7}$$

where $dist(\ell, \ell')$ is simply the well-known *Haversine formula* to calculate the great-circle distance between two points ℓ and ℓ' over the Earth's surface [44]. It is important to note that for this case, two users are more likely to establish a friendship if they have a low GeoD value.

Adjusted Hausdorff Distance (AHD). This method is based on the classic Hausdorff distance but applying an adjustment to improve the friendship prediction accuracy. The AHD method is thus defined as:

$$s_{x,y}^{AHD} = \max\{\sup_{\ell \in \Phi_{\mathcal{L}}(x)} \inf_{\ell' \in \Phi_{\mathcal{L}}(y)} dist_{adj}(\ell, \ell'), \sup_{\ell' \in \Phi_{\mathcal{L}}(y)} \inf_{\ell \in \Phi_{\mathcal{L}}(x)} dist_{adj}(\ell, \ell')\}, \tag{8}$$

where $dist_{adj}(\ell, \ell') = dist(\ell, \ell') \times \max(diversity(\ell), diversity(\ell'))$ is the adjusted geographical distance between two locations ℓ and ℓ', $diversity(\ell)$ $= \exp(\mathcal{E}(\ell))$ is the location diversity used to represent a location's popularity, and sup and inf represent the *supremum* (least upper bound) and *infimum* (greatest lower bound), respectively, from the set of visited places of a user x. Also similar to GeoD method, two users will be more likely to establish a relationship if they have a low AHD value.

3.3.3. Methods based on Social Information

Despite the fact that most of previously described methods capture different social behavior patterns based on the visited places of users, they do not directly use the social strength of ties between visitors of places [6].

In the last years, some methods have been proposed to compute the friendship probability between a pair of users based on the places visited by their common friends. Some methods based on social strength are common neighbors within and outside of common places, common neighbors of places, common neighbors with total and partial overlapping of places and total common friend common check-ins [6, 40]. Below, we describe two representative friendship prediction methods for LBSNs based on social strength.

Common Neighbors of Places (CNP). This indicates that a pair of users x and y are more likely have a future friendship if they have many common friends visiting the same places also visited by at least x or y. Thus, the CNP method is defined as:

$$s_{x,y}^{CNP} = |\Lambda_{x,y}^{\mathcal{L}}|, \tag{9}$$

where $\Lambda_{x,y}^{\mathcal{L}} = \{z \in \Lambda_{x,y} \mid \Phi_{\mathcal{L}}(x) \cap \Phi_{\mathcal{L}}(z) \neq \varnothing \ \vee \ \Phi_{\mathcal{L}}(y) \cap \Phi_{\mathcal{L}}(z) \neq \varnothing\}$ is the set of common neighbors of places of users x and y, and $\Lambda_{x,y} = \{z \in V \mid (x,z) \in E \ \wedge \ (y,z) \in E\}$ is the traditional set of common neighbors of pair of users x and y.

Common Neighbors with Total and Partial Overlapping of Places (TPOP). This considers that a pair of users x and y could develop a friendship if they have more common friends visiting places also visited by both users than common friends who visited places also visited by only one of them. Therefore, the TPOP method is defined as:

$$s_{x,y}^{TPOP} = \frac{|\Lambda_{x,y}^{TOP}|}{|\Lambda_{x,y}^{POP}|}, \tag{10}$$

where, $\Lambda_{x,y}^{TOP} = \{z \in \Lambda_{x,y}^{\mathcal{L}} \mid \Phi_{\mathcal{L}}(x) \cap \Phi_{\mathcal{L}}(z) \neq \varnothing \wedge \Phi_{\mathcal{L}}(y) \cap \Phi_{\mathcal{L}}(z) \neq \varnothing\}$ is the set of common neighbors with total overlapping of places, and $\Lambda_{x,y}^{POP} = \Lambda_{x,y}^{\mathcal{L}} - \Lambda_{x,y}^{TOP}$ is the set of common neighbors with partial overlapping of places.

4. Proposals

We analyzed the reviewed link prediction methods and observed that some of them use more than one information source to improve their prediction accuracy. For example, AAP is naturally a method based on check-in frequency but it also use distinct visitations at specific places as additional information source. Other example is AHD, which is naturally a method based on geographical distance but it also use check-in frequency and information gain as additional information sources. Table 1 provides an overview of different information sources used by each friendship prediction method described in Section 3.3.

Table 1: Summary of the friendship prediction methods for LBSNs, from the literature and our proposals, as well as the information sources used to make their predictions.

Method	**Check-in Information**		**Place Information**		**Social Information**
	Frequency	**Information Gain**	**Distinct Visitations**	**Geographical Distance**	
Co	✓				
AAP	✓		✓		
AAE		✓	✓		
LC	✓	✓	✓		
CL			✓		
JacP			✓		
GeoD	✓			✓	
AHD	✓	✓		✓	
CNP			✓		✓
TPOP			✓		✓
ChO	✓		✓		
ChA	✓		✓		
FAW	✓		✓		✓
CNNP			✓	✓	✓
NDA	✓	✓	✓	✓	✓

From Table 1 we found that some information sources were not combined, for instance, social strength is only combined with distinct visitations at specific places. Assuming that combination of some information sources could improve the friendship prediction accuracy, we propose five new methods referred to as *check-in observation* (ChO), *check-in allocation* (ChA), *friendship allocation within common places* (FAW), *common neighbors of nearby places* (CNNP) and *nearby distance allocation* (NDA). They are shown in bold in Table 1 and are described as follows:

Check-in Observation (ChO). This is based on both the distinct visitations at specific places and check-in frequency to perform predictions. We define ChO method as the ratio of the sum of the number of check-ins of users x and y at common visited places to the total sum of the number of check-ins

at all locations visited by these users. Thus, ChO is computed as:

$$s_{x,y}^{ChO} = \frac{\sum_{\ell \in \Phi_{\mathcal{L}}(x,y)} |\Phi(x,\ell)| + |\Phi(y,\ell)|}{\sum_{\ell' \in \Phi_{\mathcal{L}}(x)} |\Phi(x,\ell')| \ + \ \sum_{\ell'' \in \Phi_{\mathcal{L}}(y)} |\Phi(y,\ell'')|}. \tag{11}$$

Check-in Allocation (ChA). This is based on the traditional resource allocation method, ChA refines the popularity of all common visited places of users x and y through the count of total check-ins of each of such places. Therefore, ChA is defined as:

$$s_{x,y}^{ChA} = \sum_{\ell \in \Phi_{\mathcal{L}}(x,y)} \frac{1}{|\Phi(\ell)|}. \tag{12}$$

ChA heavily punishes high numbers of check-ins at popular places (e.g. public venues) by not applying a logarithmic function on the size of sets of all check-ins at these places. Similar to ChO, the ChA method is also based on both the distinct visitations at specific places and check-in frequency to work.

Friendship Allocation Within Common Places (FAW). This is also inspired by the traditional resource allocation method. Let the set of common neighbors within common visited places be $\Lambda_{x,y}^{WCP} = \{z \in \Lambda_{x,y} \mid \Phi_{\mathcal{L}}(x,y) \cap \Phi_{\mathcal{L}}(z) \neq \varnothing\}$, the FAW method refines the number of check-ins made by all common friends within common visited places of users x and y. Therefore, the FAW is defined as:

$$s_{x,y}^{FAW} = \sum_{z \in \Lambda_{x,y}^{WCP}} \frac{1}{|\Phi(z)|}. \tag{13}$$

Despite the use of check-in frequency and distinct visitations at places by FAW, we consider that this method is mainly based on social strength, due to the fact that this criterion is the filter used to perform predictions.

Common Neighbors of Nearby Places (CNNP). This counts the number of common friends of users x and y whose geographical distance between their home locations and the home location of at least one, x or y, lies within a given radio. Therefore, given a distance threshold τ_d, CNNP is computed as:

$$s_{x,y}^{CNNP} = |\{z \mid \forall z \in \Lambda_{x,y} \wedge (dist(\ell_x^h, \ell_z^h) \leq \tau_d \vee dist(\ell_y^h, \ell_z^h) \leq \tau_d)\}|. \tag{14}$$

CNNP uses full place information as well as social information to make predictions, however we consider that it is a method based on social strength due to the fact that this criterion is fundamental for CNNP to work.

Nearby Distance Allocation (NDA). This refines all the minimum adjusted distances calculated between the home locations of users x and y, and

the respective home locations of all of their common neighbors of places. Therefore, NDA is defined as:

$$s_{x,y}^{NDA} = \sum_{z \in \Lambda_{x,y}^{\mathcal{L}}} \frac{1}{\min\{dist_{adj}(\ell_x^h, \ell_z^h), dist_{adj}(\ell_y^h, \ell_z^h)\}}. \quad (15)$$

NDA is the only method that uses full check-in, place and social information. However, as previously applied for the other proposals, since NDA uses social strength as the main criterion, we consider it to be a method based on social information.

5. Performance Evaluation

In this section, we present an experimental evaluation carried out for all link prediction methods previously studied. This section includes an analysis of three real-world LBSN datasets with which the experiments were performed as well as a deep analysis of the predictive capabilities of each evaluated method.

5.1. Dataset Description

The datasets used in our experiments are real-world LBSNs in which users made check-ins to report visits to specific physical locations. In this section, we describe their main properties and ways to construct the training and test datasets.

5.1.1. Dataset Selection

The datasets used for our experiments had to meet certain requirements: i) they had to represent social and location data, i.e. data defining existing connections between users as well as the check-ins of all of them at all of their visited locations, and ii) those connections and/or check-ins had to be time stamped. Based on these two criteria, we selected three datasets collected from real-world LBSNs, which are commonly used by the scientific community for mining tasks in the LBSN domain.

Brightkite. This was once a location-based social networking service provider where users shared their locations by checking-in. The Brightkite service was shut down in 2012, but the dataset was collected over the April 2008 to October 2010 period [4]. This publicly available dataset[1] consists of 58228 users, 214078 relations, 4491144 check-ins and 772788 places.

Gowalla. This is also another location-based social networking service that ceased operation in 2012. The dataset was collected over the February 2009 to October 2010 period [4] and also is publicly available[2]. This dataset contains 196591 users, 950327 relations, 6442892 check-ins and 1280969 different places.

[1] `http://snap.stanford.edu/data/loc-brightkite.html`
[2] `http://snap.stanford.edu/data/loc-gowalla.html`

Foursquare. Foursquare is one of the most popular online LBSN. Currently, this service report more than 50 million users, 12 billion check-ins and 105 million places in January 2018[3]. The dataset used for us experiments was collected over January 2011 to December 2011 period [45]. This publicly available dataset[4] contains 11326 users, 23582 relations, 2029555 check-ins and 172044 different places.

The various properties of these datasets were calculated an the values depicted in Table 2. This table is divided into two parts, the first shows topological properties [1] whilst the second shows location properties [2, 6]. Therefore, considering the first part of Table 2 we observe that the analyzed networks have a small *average degree*, $\langle k \rangle$, which suggests that the users of these networks had between 4 and 10 friends in average. This implies that the *average clustering coefficient*, C, of networks is also low. However, the low *degree heterogeneity*, $H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$, of Brightkite and Foursquare indicate that their users are less different from each other than the users of Gowalla. Also, the *assortativity coefficient* r, which measures the preference of users to attach to others, shows that only Brightkite is assortative, which is why it has a positive value, indicating the presence of few relationships among users with a similar degree. On the other hand, Gowalla and Foursquare are disassortative, since their assortativity coefficients are negative, indicating the presence of a considerable number of relationships among users with a different degree.

Table 2: The main properties of the experimental LBSNs.

	Brightkite	**Gowalla**	**Foursquare**
$\|V\|$	58228	196591	11326
$\|E\|$	214078	950327	23582
$\langle k \rangle$	7.35	9.66	4.16
C	0.17	0.24	0.06
H	8.66	31.71	7.66
r	0.01	-0.03	-0.07
$\|\Phi\|$	4491144	6442892	202955
$\|\Phi_V\|$	50686	107092	9985
$\langle \Phi \rangle$	88	60	179
$\|\mathcal{L}\|$	772788	1280969	172044
$\langle \mathcal{L}_\Phi \rangle$	5	5	11
$\langle \mathcal{E} \rangle$	0.05	0.25	0.19

[3]https://foursquare.com/about
[4]http://www.public.asu.edu/~hgao16/Publications.html

Considering the second part of Table 2, we observe that the *number of users with at least one check-in*, $|\Phi_V|$, is a little over 85% of total users of networks. Despite the fact that Gowalla and Brightkite have more users and check-ins than Foursquare, the *average number of check-ins per user*, $\langle\Phi\rangle$, of Foursquare users is greater than that of Gowalla and Brightkite users. However, the *average of check-ins per place*, $\langle\mathcal{L}_\Phi\rangle$, is similar for Brightkite and Gowalla, whilst for Foursquare is greater, i.e. Foursquare users made more check-ins at a specific place than Brightkite and Gowalla users. Finally, the very small *average place entropy*, $\langle\mathcal{E}\rangle = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\mathcal{E}(\ell)$, of Brightkite suggests that the location information in this LBSN is a stronger factor to facilitate the establishment of new relationships between users than for Gowalla and Foursquare users.

5.1.2. Data Processing

We preprocess the datasets to make the data suitable for our experiments. Considering that isolated nodes and locations without visits can generate noise when measuring the performance of different link prediction methods, it is necessary to apply a policy for selecting data samples containing more representative information. Therefore, for each dataset, we consider only users with at least one friend and with at least one check-in at any location.

Since our goal is to predict new friendships between users, we divided each dataset into training and test (or probe) sets while taking the time stamps information available into account. Therefore, links formed by Brightkite users who checked-in from April 2008 to January 2010 were used to construct the training set, whilst links formed by users who checked-in from February 2010 to October 2010 were used for the probe set. For Gowalla, the training set was constructed with links formed by users who checked-in from February 2009 to April 2010, and the probe set was constructed with links formed by users who checked-in from May 2010 to October 2010. Whereas, for Foursquare the training set is formed by users who checked-in from January 2011 to September 2011, whilst the probe set is formed by users that made check-ins over the October 2011 to December 2011 period. Table 3 shows the training and testing time ranges for the three datasets.

Table 3: Details of pre-processed datasets.

Dataset	**Training time range**	**Testing time range**	$\langle\|V\|\rangle$	$\langle\|\mathcal{L}\|\rangle$	$\langle\|E^T\|\rangle$	$\langle\|E^P\|\rangle$
Brightkite	2008/04 - 2010/01	2010/02 - 2010/10	4606	277515	49460	24800
Gowalla	2009/02 - 2010/04	2010/05 - 2010/10	19981	607094	232194	87619
Foursquare	2011/01 - 2011/09	2011/10 - 2011/12	7287	101546	12258	8565

Different studies have used a similar strategy for splitting data into training and probe sets, but they were not concerned about maintaining the consistency between users in both sets [6, 40, 25], which could affect the performance of link prediction methods in different ways [46]. To avoid that, we proceeded to remove all links formed by users who checked-in only during the training time range or only in the testing time range. From the links formed by users with

check-ins in both the training and testing time ranges, we chose one-third of the links formed by users at random with a higher degree than the average degree for the probe set, while the remaining links were part of the training set. Therefore, we obtained the training set $G^T(V, E^T, \mathcal{L}, \Phi^T)$ and probe set $G^P(V, E^P, \mathcal{L}, \Phi^P)$, where both sets keep the same users (V) and locations ($\mathcal{L}$) but differ in the social (E^T and E^P) and user-location (Φ^T and Φ^P) links.

Table 3 also summarizes the average number of users, $\langle|V|\rangle$, average number of different locations, $\langle|\mathcal{L}|\rangle$, average number of training social links, $\langle|E^T|\rangle$ and average number of testing social links, $\langle|E^P|\rangle$, obtained by averaging 10 independent partitions of each dataset. It is important to comment that, for the three datasets, the average number of check-ins in training set, $\langle|\Phi^P|\rangle$, is two-thirds of the total number of check-ins whilst the average number of check-ins in probe set, $\langle|\Phi^T|\rangle$, is the remainder part.

5.1.3. Data Limitations

Although the datasets selected contain thousands of users and links, they can be considered as relatively small compared to other online social network datasets. Notwithstanding this limitation present in the datasets analyzed in this study, we use them since they meet the requirements explained previously in Section 5.1.1 and also because they are frequently used in the state-of-the-art in order to propose a quantitive and qualitative analysis on the social and spatial factors impacting the friendships [16, 4, 18, 40]. Therefore, this work offers new light on exploiting the different information sources to improve friendship prediction in Brightkite, Gowalla and Foursquare, but our findings could be applied for other LBSNs.

Some studies of the state-of-the-art use other datasets, *e.g.* Foursquare [25, 19], Facebook [27], Twitter [19], Second Life [41, 42], and other LBSNs. But we cannot use them for two main reasons: i) generally they are not publicly available, and ii) they do not respect the requirements detailed in Section 5.1.1.

5.2. Experimental Setup

For each of the 10 independent partitions of each dataset obtained as explained in Section 5.1.2, we considered 10 executions of each link prediction method presented in Section 3 and our proposals described in Section 4. We then applied different performance measures to the prediction results to determine which were the most accurate and efficient link prediction methods.

All of the evaluation tests were performed using the *Geo-LPsource* framework, which we developed and is publicly available[5]. We set the default parameters of the link prediction methods as follows: i) for Co method we considered that $\tau = 1$ day, ii) for LC method we considered that $\tau_{\mathcal{E}} = \langle\mathcal{E}\rangle$, iii) for CNNP method we considered that $\tau_d = 1500$ m., and iv) for AHD method, for a user x and being ℓ the most visited place by him, we considered that the comparison

[5] `https://github.com/jvalverr/Geo-LPsource`

value for the calculation of supremum was $v_s = \frac{|\Phi(x,\ell)|}{2}$, whilst the comparison value for the calculation of infimum was $v_i = \frac{|\Phi(x,\ell)|}{5}$.

5.3. Evaluation Results

For the three LBSNs analyzed, Table 4 summarizes the performance results for each link prediction method through different evaluation metrics. Each value in this table was obtained by averaging over 10 runs, over 10 partitions of training and testing sets, as previously detailed in Section 5.2. The values highlighted in bold correspond to the best results achieved for each evaluation metric.

Table 4: Friendship prediction results for Brightkite, Gowalla and Foursquare. Highlighted values indicate the best results for each evaluation metric considered.

		Brightkite			Gowalla			Foursquare		
Method		**IR**	**F_1**	**AUC**	**IR**	**F_1**	**AUC**	**IR**	**F_1**	**AUC**
Co		**4.934**	0.070	0.668	**14.972**	0.051	0.554	**4.488**	0.045	0.554
AAP		13.190	0.104	0.682	36.531	0.045	0.728	13.367	0.034	0.655
AAE		13.190	0.104	0.694	36.586	0.045	0.736	13.367	0.034	0.670
LC		34.000	0.055	0.629	180.945	0.011	0.542	27.844	0.017	0.470
CL		13.114	0.105	0.676	36.327	0.045	0.682	13.368	0.034	0.630
JacP		13.114	0.105	0.630	36.327	0.045	0.742	13.368	0.034	**0.708**
GeoD		35.005	0.053	0.710	180.461	0.011	**0.767**	35.710	0.018	0.705
AHD		31.689	0.056	0.685	223.714	0.011	0.681	35.782	0.018	0.656
CNP		31.180	0.060	**0.761**	66.484	0.029	0.687	23.277	0.027	0.608
TPOP		13.441	0.105	0.673	25.383	0.057	0.665	12.588	0.036	0.594
ChO		13.079	0.104	0.608	31.197	0.050	0.714	13.292	0.034	0.671
ChA		13.173	0.104	0.676	36.460	0.045	0.736	13.367	0.034	0.667
FAW		9.678	**0.113**	0.740	15.821	**0.069**	0.718	7.764	**0.046**	0.642
CNNP		9.387	0.048	0.552	18.868	0.046	0.620	4.920	0.039	0.569
NDA		22.496	0.076	0.700	47.540	0.037	0.720	15.325	0.024	0.624

From Table 4, imbalance ratio and F-measure results were calculated considering the whole list of predicted links obtained by each evaluated link prediction method. On the other hand, the AUC results were calculated from a list of $n = 5000$ pairs of wrongly and right predicted links chosen randomly and independently. Due to the number of link prediction methods studied and the different ways they were evaluated, we performed a set of analyses to determine which were the best friendship prediction methods for LBSNs.

5.3.1. Reducing the Prediction Space Size

The prediction space size is related to the size of the set of predicted links, L_p. Most existing link prediction methods prioritize an increase in the number of correctly predicted links even at the cost of a huge amount of wrong predictions. This generates a extremely skewed distribution of classes in the prediction space, which in turn impairs the performance of any link prediction

method [23]. Therefore, efforts should also focus not only on reducing the number of wrong predictions but also on increasing the number of correctly predicted links relative to the total number of predictions.

Previous studies showed that the prediction space size of methods based only on the network topology is around $10^{11} \sim 10^{12}$ links for Brightkite and Gowalla. However, by using methods based on location information, the prediction space can be reduced by about 15-fold or more [6, 23]. Based on that and to determine if reduction of the prediction space is related to different information sources, in Figure 2 we report the average prediction space size of the different link prediction methods analyzed in this study.

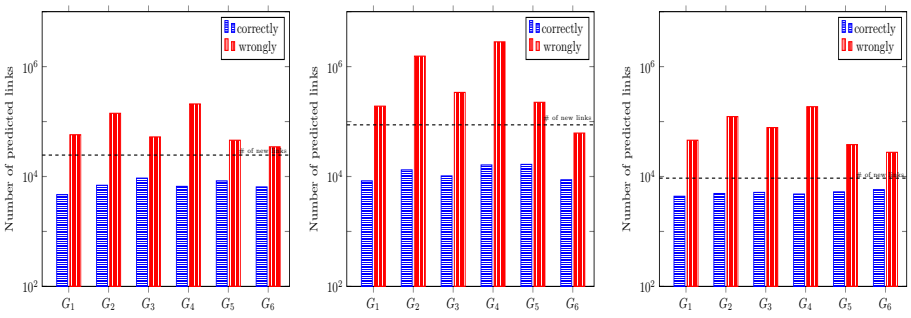


Figure 2: Number of correctly and wrongly predicted links for methods based on check-in frequency (G_1), information gain (G_2), distinct visitations at places (G_3), geographical distance (G_4), social strength (G_5) and our proposals (G_6) for (a) Brightkite, (b) Gowalla and (c) Foursquare. The dashed horizontal lines indicate the number of truly new links (links into the probe set) for each dataset. Results averaged over the 10 analyzed partitions and plotted in log 10 scale.

Figure 2 shows that for the analyzed networks, methods based on check-in frequency, information gain, distinct visitations at places and geographical distance, followed the traditional logic of obtaining a high number of right predictions at the cost of a much higher number of wrong predictions [47]. On the other hand, methods based on social strength led to a considerably lower number of wrong predictions at the cost of a small decrease in the number of correctly predicted links relative to the results obtained by the first cited methods, which is important in a real scenario [6]. Our proposals followed a similar scheme as methods based on social strength, leading to less wrong predictions.

This fact is clearly shown by the IR results in Table 4 where, besides highlighting that Co method generally had a better IR performance, we observed that some methods based on check-in frequency, information gain, distinct visitations at places and geographical distance had an IR higher than most methods based on social strength and our proposals. Therefore, Co was the method with the overall best IR performance, whilst GeoD and AHD were the worst ones.

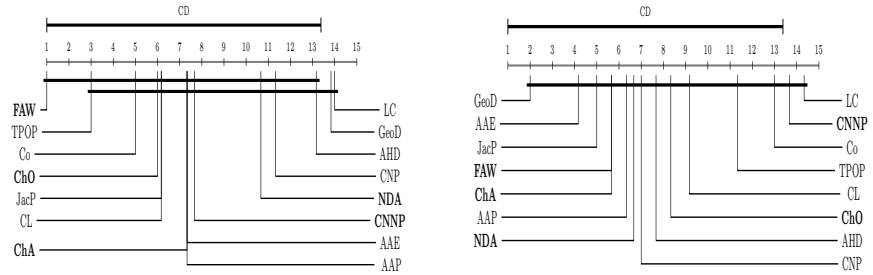
Considering only our proposals, we found that FAW and CNNP performed better in IR. These two methods have social components, which help to significantly reduce the prediction space size. The worst IR performance of our proposals was obtained by NDA, which is based on geographical distance, thus confirming that this type of information source generates a large prediction space.

5.3.2. Measuring the Accuracy

Since the IR results shown that some methods obtained a considerable number of correctly predicted links whilst others obtained an absurdly large number of wrongly predicted links, we adopted the f-measure (F_1) to evaluate the performance of prediction methods in terms of relevant predicted links. Therefore, we observe that FAW method, which is one of our proposals, had the best f-measure performance in the three analyzed LBSNs.

To facilitate the analysis of all link prediction methods, based on Table 4 we ranked the average F_1 results obtained by all the link prediction methods in the three analyzed networks, and then we applied the Friedman and Nemenyi post-hoc tests [48]. Therefore, the F-statistics with 14 and 28 degrees of freedom and at the 95 percentile was 2.06. According to the Friedman test using F-statistics, the null-hypothesis that the link prediction methods behave similarly when compared with respect to their F_1 performance should be rejected.

Figure 3(a) shows the Nemenyi test results for the 15 analyzed link prediction methods considering the F_1 ranking. The critical difference (CD) value for comparing the mean-ranking of two different methods at the 95 percentile was 12.38, as shown on the top of the diagram. The method names are shown on the axis of the diagram, with our proposals highlighted in bold. The lowest (best) ranks are on the left side of the axis. Methods connected by a bold line in the diagram have no statistical significant difference, so the Nemenyi test indicated that FAW has statistical significant difference with LC and GeoD.



(a) (b)

Figure 3: Nemenyi post-hoc test diagrams obtained from (a) f-measure and (b) AUC results showed in Table 4. Our proposals are highlighted in bold.

Figure 3(a) indicates that methods based on social strength, such as FAW

and TPOP, performed better than the others since occupied the first and second position, respectively. Co and ChO are in third and fourth position, respectively, whilst JacP and CL tied for the fifth position. After these methods, and a little further away, we have that ChA, AAP and AAE tied for the sixth position, CNNP and NDA are in seventh and eight position, respectively. CNP is ninth, AHD is tenth , GeoD is eleventh and LC is twelfth. Therefore, we observe that two of our proposals, FAW and ChO, are in the top-5. Moreover, methods based on information gain, such as LC, and methods based on geographical distance, such as GeoD and AHD, were at the end of the ranking.

5.3.3. Analyzing the Predictive Power

Table 4 also shows the prediction results obtained for AUC. From these results, we observed that CNP, GeoD and JacP outperformed all the other link prediction methods in Brightkite, Gowalla and Foursquare, respectively. In addition, we found that all the link prediction methods performed better than pure chance, except for LC in Foursquare.

Furthermore, to gain further insight into the real prediction power of evaluated link prediction methods, we followed the same scheme used previously for F_1 analysis. Therefore, we ranked the average results of AUC obtained by all the link prediction methods, and then we applied Friedman and Nemenyi post-hoc tests. Similarly that for F_1 analysis, the critical value of the F-statistics with 14 and 28 degrees of freedom and at the 95 percentile was 2.06. However, unlike the F_1 analysis, this time the Friedman test suggested that the null-hypothesis that the link prediction methods behave similarly when compared by their AUC performance should not be rejected.

Figure 3(b) shows the Nemenyi test results for the evaluated methods ranked by AUC. The diagram indicates that the CD value calculated at the 95 percentile was 12.38. This test also showed that the link prediction methods have no statistical significant difference, so they are connected by a bold line.

Figure 3(b) indicates that, differently from F_1 analysis, this time the methods based on geographical distance and information gain are in the firsts positions. Thus, GeoD and AAE are in first and second position, respectively. JacP is third whilst FAW and ChA tied for the fourth position and AAP is fifth. The rest of the ranking was in the following order: NDA, CNP, AHD, ChO, CL, TPOP, Co, CNNP and LC. In this ranking, we also have two of our proposals in the top-5. FAW and ChA. To our surprise, LC continues in last position and some methods that have performed well in the F_1 ranking, such as TPOP, Co and CL, this time were in compromising positions.

5.3.4. Obtaining the Top-5 Friendship Prediction Methods

Since some link prediction methods performed better in the prediction space analysis whilst other ones did in the prediction power analysis, we analyzed the F_1 and AUC results at the same time. Therefore, from Table 4 we ranked the average F_1 and AUC results obtained by all the link prediction methods, and then applied Friedman and Nemenyi post-hoc tests to them. The critical F-statistic value with 14 and 70 degrees of freedom and at the 95 percentile

was 1.84. Based on this F-statistic, the Friedman test suggested that the null-hypothesis that the methods behave similarly when compared according to their F_1 and AUC performances should be rejected.

Figure 4 shows the Nemenyi test results for the analyzed methods in our final ranking. The diagram in Figure 4 indicates that the CD value at the 95 percentile is 8.76. From diagram in Figure 4, we observe that FAW has statistical significant difference with LC.

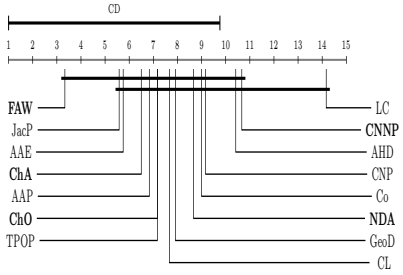
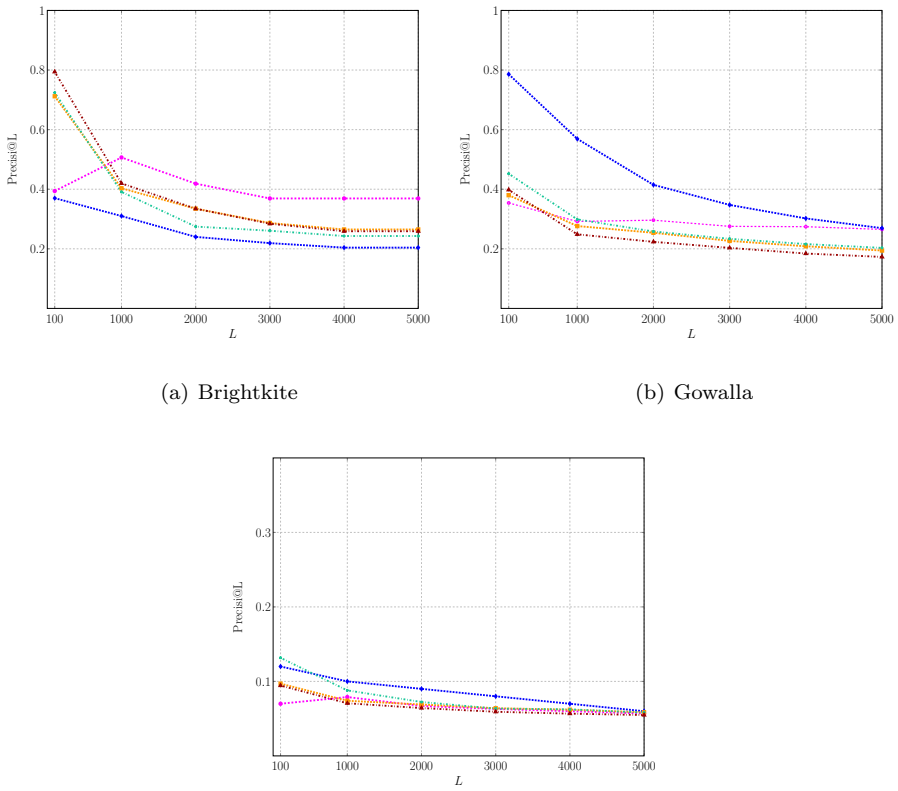


Figure 4: Nemenyi post-hoc test diagram obtained over the F_1 and AUC average ranks showed in Table 4. Diagram shows the final ranking of link prediction methods considering both the optimal reduction of prediction space size and high prediction power. Our proposals are highlighted in bold.

Figure 4 indicates that FAW is in first position, JacP is second, AAE is third, ChA is fourth and AAP is fifth. ChO and TPOP tied for the sixth position. The rest of the ranking was in the following order: CL, GeoD, NDA, Co, CNP, AHD, CNNP and LC. Therefore, two of our proposals, FAW and ChA, are in the top-5 of the final ranking. LC definitively has the worst performance. Note that the methods in the top-5 belong to the different information sources identified in this study, so we have a method based on social strength (FAW), a method based on distinct visitations at places (JacP), a method based on information gain (AAE) and two methods based on check-in frequency (ChA and AAP). The only one missing in the top-5 of the final ranking is some method based on geographical distance.

For recommending to users some links as possible new friendships, we can just select the links with the highest scores [8, 9, 6]. Furthermore, whereas for recommendation task is not enough only a method with good prediction performance, also it is necessary that from a limited portion of the total predicted links it generates a high number of right predictions, good enough to be showed to users as appropriate friendship suggestions [15, 49]. Therefore, to assess the performance of top-5 methods from the final ranking through limited segments of the total list of predicted links, we analyzed them by precisi@L. Figure 5 shows the different precisi@L performances for the top-5 methods of our final ranking. These precisi@L results were calculated for different L values and for each analyzed LBSN.

Figure 5 indicates that most of the evaluated methods performed best when $L = 100$, i.e. they are able to make a few accurate predictions. When link



(a) Brightkite

(b) Gowalla

(c) Foursquare

Figure 5: Precisi@L performance for the top-5 methods of the final ranking considering different L values for (a) Brightkite, (b) Gowalla and (c) Foursquare.

prediction methods have to make more than a thousand predictions, i.e when $L > 1000$, their prediction abilities decrease considerably. Moreover, Figure 5 shows that the evaluated methods have a similar behavior in the three analyzed LBSNs. Thus, ChA, AAP and AAE performed similarly with a slight superiority of ChA. Moreover, JacP and FAW showed similar performance with a slight superiority of JacP.

Analyzing the precis@n performance of methods in each analyzed network, Figure 5(a) shows that in Brightkite, AAP outperformed all the other evaluated methods when $L = 100$. Thereafter, our proposal FAW performed better than the rest of methods for the rest of L values. JacP outperformed poorly. Figure 5(b) shows that in Gowalla, the methods JacP achieved the best performance for all of the L values. One of our proposals, i.e. ChA, ranks second when

$L = 100$, to remain in third position for the rest of L values. When $L = 1000$, other of our proposals, i.e. FAW, achieve the second position and it holds that position for the rest of L values. Finally, Figure 5(c) shows that methods in this network achieved very low precisi@L values (less than 0.2). However, in Foursquare, ChA outperformed all the methods when $L = 100$ but it is overcome by JacP, which keeps the second position for the rest of L values. Our proposal FAW performed poorly when $L = 100$ but it achieves the third position when $L = 1000$ and maintain this position since it slightly exceeds AAE and AAP.

6. Conclusion

In last years, a variety of online services which provide users with easy ways to share their geo-spatial locations and location-related content have become popular. These services, called LBSNs, constitute a new type of social network and give rise to new opportunities and challenges with regard to different social network issues, such as location recommendation [27, 14, 15], user identification [36, 37], discovery of local experts [33, 34, 35], and discovery of travel companions [32]. Motivated by the important role that LBSNs are playing for millions of users, we conducted a survey of recent related research on friendship prediction and recommendation.

Although there is abundant methods to tackle the friendship prediction problem in the LBSN domain, there is a lack of well organised and clearly explained taxonomy that helps the best use of current literature. Therefore, our first contribution in this work was related to proposes a taxonomy for friendship prediction methods for LBSNs based on five information sources identified: frequency of check-ins, information gain, distinct visitations at places, geographical distance and social strength.

Based on the taxonomy proposed, we identified some gaps in existing friendship prediction methods and proposed five new ones: check-in observation (ChO), check-in allocation (ChA), friendship allocation within common places (FAW), common neighbors of nearby places (CNNP) and nearby distance allocation (NDA). These new friendship prediction methods are exclusive to perform friendship prediction task in the LBSN context and constituted our second contribution.

Due to the fact that we aimed objectively quantify the predictive power of friendship prediction methods in LBSNs as well as determine how good they work in the context of recommender systems, our third contribution is related to the identification of the top-5 friendship prediction methods that better perform in the LBSN context. For this purpose, we performed an exhaustive evaluation process on snapshots of three well known real-world LBSNs.

Based on our results, we empirically demonstrate that some friendship prediction methods for LBSNs could be ranked as the best for some evaluation measure but could perform poorly for other ones. Thus, we stressed the importance of choosing the appropriate measure according to the objective pursued in

the friendship prediction task. For instance, in general, some friendship prediction methods performed better with regard to the F-measure than with AUC, so if in any real-world application it is necessary to focus on minimizing the number of wrong predictions, the best option is to consider methods that work well based on the F-measure. However, if the focus is to obtain a high number of right predictions, but with a high chance that these predictions represent strong connections, then the best option could be to consider methods that work well based on AUC.

Nevertheless, in a real-world scenario likely will be necessary to balance both the F-measure and AUC performance of methods. Thus, we finally identified the top-5 friendship prediction methods that performed in a balanced way for different metrics. Moreover, in this top-5 are two of our proposals, FAW in the first position and ChA in the fourth.

Other observation based on our results is related to the fact that the use of a variety of information sources does not guarantee the best performance of a method. For instance, NDA method, which is one of our proposals, is the only one that uses all the information sources identified, but it appears in the ninth position of our final ranking. Finally, we also observe that methods based purely on check-in information or place information performed worse than methods combining these information sources with social information. Therefore, we have empirical foundation to support the argument that the best way to cope with friendship prediction problem in the LBSN context is by combining social strength with location information.

The future directions of our work will focus on location prediction, which will be used to recommend places that users could visit. For that, we hope that the location information sources identified in this work can also be used in the location prediction task.

Acknowledgments

This research was partially supported by Brazilian agencies FAPESP (grants 2015/14228-9 and 2013/12191-5), CNPq (grant 302645/2015-2), and by the French SONGES project (Occitanie and FEDER).

References

[1] A.-L. Barabási, Network Science, Cambridge University Press, 2016.

[2] Y. Zheng, X. Zhou, Computing with Spatial Trajectories, 1st Edition, Springer, 2011, Ch. 8.

[3] Z. Yu, C. Wang, J. Bu, X. Wang, Y. Wu, C. Chen, Friend recommendation with content spread enhancement in social networks, Information Sciences 309 (2015) 102–118.

[4] E. Cho, S. A. Myers, J. Leskovec, Friendship and mobility: User movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011, ACM, 2011, pp. 1082–1090.

[5] D. H. Zhu, Y. P. Chang, J. J. Luo, X. Li, Understanding the adoption of location-based recommendation agents among active users of social networking sites, Information Processing & Management 50 (5) (2014) 675–682.

[6] J. Valverde-Rebaza, M. Roche, P. Poncelet, A. Lopes, Exploiting social and mobility patterns for friendship prediction in location-based social networks, in: The 23rd International Conference on Pattern Recognition, ICPR 2016, IEEE, 2016, pp. 2526–2531.

[7] O. Ozdikis, H. Ouztzn, P. Karagoz, Evidential estimation of event locations in microblogs using the dempstershafer theory, Information Processing & Management 52 (6) (2016) 1227–1246.

[8] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, Journal of the American Society for Information Science and Technology 58 (7) (2007) 1019–1031.

[9] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A: Statistical Mechanics and its Applications 390 (6) (2011) 1150–1170.

[10] V. Martínez, F. Berzal, J.-C. Cubero, A survey of link prediction in complex networks, ACM Computing Surveys 49 (4) (2016) 69:1–69:33.

[11] J. Wu, G. Zhang, Y. Ren, A balanced modularity maximization link prediction model in social networks, Information Processing & Management 53 (1) (2017) 295–307.

[12] J. Bao, Y. Zheng, D. Wilkie, M. Mokbel, Recommendations in location-based social networks: A survey, Geoinformatica 19 (3) (2015) 525–565.

[13] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, W.-Y. Ma, Mining user similarity based on location history, in: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '08, ACM, 2008, pp. 34:1–34:10.

[14] J. Wang, R. Tan, R.-P. Zhang, F. You, A recommender system research based on location-based social networks, in: G. Meiselwitz (Ed.), Proceedings of the 8th International Conference on Social Computing and Social Media, SCSM 2016, Springer, 2016, pp. 81–90.

[15] R. Pálovics, P. Szalai, J. Pap, E. Frigó, L. Kocsis, A. A. Benczúr, Location-aware online learning for top-k recommendation, Pervasive and Mobile Computing 38 (2017) 490–504, special Issue IEEE International Conference on Pervasive Computing and Communications (PerCom) 2016.

[16] M. Allamanis, S. Scellato, C. Mascolo, Evolution of a location-based online social network: Analysis and models, in: Proceedings of the 2012 Internet Measurement Conference, IMC '12, ACM, 2012, pp. 145–158.

[17] M. Narayanan, A. K. Cherukuri, A study and analysis of recommendation systems for location-based social network (LBSN) with big data, IIMB Management Review 28 (1) (2016) 25–30.

[18] O. Mengshoel, R. Desail, A. Chen, B. Tran, Will we connect again? machine learning for link prediction in mobile social networks, in: Eleventh Workshop on Mining and Learning with Graphs, ACM, 2013.

[19] J. Zhang, X. Kong, P. S. Yu, Transferring heterogeneous links across location-based social networks, in: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14, ACM, 2014, pp. 303–312.

[20] Q. Liu, S. Tang, X. Zhang, X. Zhao, B. Y. Zhao, H. Zheng, Network growth and link prediction through an empirical lens, in: Proceedings of the Proceedings of the 16th ACM SIGCOMM Internet Measurement Conference, IMC 2016, 2016, pp. 1–15.

[21] J. Valverde-Rebaza, A. Lopes, Exploiting behaviors of communities of Twitter users for link prediction, Social Network Analysis and Mining 3 (4) (2013) 1063–1074.

[22] A. Shahmohammadi, E. Khadangi, A. Bagheri, Presenting new collaborative link prediction methods for activity recommendation in Facebook, Neurocomputing 210 (2016) 217–226.

[23] S. Scellato, A. Noulas, C. Mascolo, Exploiting place features in link prediction on location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, ACM, 2011, pp. 1046–1054.

[24] H. Pham, C. Shahabi, Y. Liu, Ebm: An entropy-based model to infer social strength from spatiotemporal data, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD '13, ACM, 2013, pp. 265–276.

[25] H. Luo, B. Guo, Zhiwenyu, Z. Wang, Y. Feng, Friendship Prediction Based on the Fusion of Topology and Geographical Features in LBSN, in: 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, HPCC-EUC 2013, IEEE, 2013, pp. 2224–2230.

[26] G. Xu-Rui, W. Li, W. Wei-Li, An algorithm for friendship prediction on location-based social networks, in: Proceedings of the 4th International Conference on Computational Social Networks, CSoNet 2015, Springer International Publishing, 2015, pp. 193–204.

[27] J. McGee, J. Caverlee, Z. Cheng, Location prediction in social media based on tie strength, in: Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, CIKM '13, ACM, 2013, pp. 459–468.

[28] Y. Zhang, J. Pang, Distance and friendship: A distance-based model for link prediction in social networks, in: Proceedings of the 17th Asia-Pacific Web Conference: Web Technologies and Applications, APWeb 2015, Springer International Publishing, 2015, pp. 55–66.

[29] S. B. Kylasa, G. Kollias, A. Grama, Social ties and checkin sites: connections and latent structures in location-based social networks, Social Network Analysis and Mining 6 (1) (2016) 95.

[30] A. E. Bayrak, F. Polat, Examining place categories for link prediction in location based social networks, in: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, 2016, pp. 976–979.

[31] X. Xiao, Y. Zheng, Q. Luo, X. Xie, Inferring social ties between users with human location history, Journal of Ambient Intelligence and Humanized Computing 5 (1) (2014) 3–19.

[32] Y. Liao, W. Lam, S. Jameel, S. Schockaert, X. Xie, Who wants to join me?: Companion recommendation in location based social networks, in: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, ICTIR '16, ACM, 2016, pp. 271–280.

[33] Z. Cheng, J. Caverlee, H. Barthwal, V. Bachani, Who is the Barbecue King of Texas?: A Geo-spatial Approach to Finding Local Experts on Twitter, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, ACM, 2014, pp. 335–344.

[34] J.-H. Liou, Y.-M. Li, Design of contextual local expert support mechanism, in: Proceedings of the 17th International Conference on Electronic Commerce 2015, ICEC '15, ACM, 2015, pp. 17:1–17:5.

[35] W. Niu, Z. Liu, J. Caverlee, On local expert discovery via geo-located crowds, queries, and candidates, ACM Transactions on Spatial Algorithms and Systems 2 (4) (2016) 14:1–14:24.

[36] L. Rossi, M. Musolesi, It's the way you check-in: Identifying users in location-based social networks, in: Proceedings of the Second ACM Conference on Online Social Networks, COSN '14, ACM, 2014, pp. 215–226.

[37] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, S. Lattanzi, Linking users across domains with location data: Theory and validation, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, Int. World Wide Web Conf. Steering Committee, 2016, pp. 707–719.

[38] G. Xu-Rui, W. Li, W. Wei-Li, Using multi-features to recommend friends on location-based social networks, Peer-to-Peer Networking and Applications 10 (2016) 1323–1330.

[39] J. Cranshaw, E. Toch, J. Hong, A. Kittur, N. Sadeh, Bridging the gap between physical location and online social networks, in: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, UbiComp '10, ACM, 2010.

[40] A. E. Bayrak, F. Polat, Contextual feature analysis to improve link prediction for location based social networks, in: Proceedings of the 8th Workshop on Social Network Mining and Analysis, SNAKDD'14, ACM, 2014, pp. 7:1–7:5.

[41] M. Steurer, C. Trattner, D. Helic, Predicting social interactions from different sources of location-based knowledge, in: The Third International Conference on Social Eco-Informatics, SOTICS 2013, IARIA, 2013, pp. 8–13.

[42] M. Steurer, C. Trattner, Acquaintance or partner predicting partnership in online and location-based social networks, in: The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013, IEEE, 2013, pp. 372–379.

[43] C.-T. Li, H.-P. Hsieh, Geo-social media analytics, in: Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, ACM, 2015, pp. 1533–1534.

[44] H. B. Goodwin, The haversine in nautical astronomy, Naval Institute Proceedings 36 (3) (1910) 735–746.

[45] H. Gao, J. Tang, H. Liu, gscorr: Modeling geo-social correlations for new check-ins on location-based social networks, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, ACM, 2012, pp. 1582–1586.

[46] Y. Yang, R. Lichtenwalter, N. V. Chawla, Evaluating link prediction methods, Knowledge and Information Systems 45 (3) (2015) 751–782.

[47] D. Wang, D. Pedreschi, C. Song, F. Giannotti, A.-L. Barabasi, Human mobility, social ties, and link prediction, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, ACM, 2011, pp. 1100–1108.

[48] J. Demšar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine Learning Research 7 (2006) 1–30.

[49] S. Ahmadian, M. Meghdadi, M. Afsharchi, A social recommendation method based on an adaptive neighbor selection mechanism, Information Processing & Management (2017) 1–19. In Press.