# Where You Go Reveals Who You Know: Analyzing Social Ties from Millions of Footprints

Hsun-Ping Hsieh
National Taiwan University
Taipei, Taiwan
d98944006@csie.ntu.edu.tw

Rui Yan
Baidu Inc.,
Beijing, China
yanrui02@baidu.com

Cheng-Te Li
Academia Sinica
Taipei, Taiwan
ctli@citi.sinica.edu.tw

## ABSTRACT

This paper aims to investigate how the geographical footprints of users correlate to their social ties. While conventional wisdom told us that the more frequently two users co-locate in geography, the higher probability they are friends, we find that in real geo-social data, Gowalla and Meetup, almost all of the user pairs with friendships had *never* met geographically. In this sense, can we discover social ties among users *purely* using their geographical footprints even if they never met? To study this question, we develop a two-stage feature engineering framework. The first stage is to characterize the *direct* linkages between users through their spatial co-locations while the second is to capture the *indirect* linkages between them via a *co-location graph*. Experiments conducted on Gowalla check-in data and Meetup meeting events exhibit not only the superiority of our feature model, but also validate the predictability (with 70% accuracy) of detecting social ties solely from user footprints.

## 1. INTRODUCTION

Location-acquisition techniques and ubiquitous mobile devices, along with the social networking services, foster the emergence of geo-social media, including Facebook Place, Foursquare, Meetup, and Plancast. Geo-social media allows users to not only maintain social connections but also keep track of their geographical footprints that are reflected by either *check-in* records or *social activities* (e.g. dinner gathering, attending conferences, and seeing movies). Both check-in records and social activities consist of latitude-longitude information and time stamps. Such geographical footprints depict where people show up and have face-to-face interactions in the physical world.

In this paper, we aim to uncover the hidden relationship between users' social ties and their geographical footprints. Specifically, we would like to answer the following three questions. First, can we discover social ties *purely* based on the geographical footprints of users? That said, to understand how footprints shape social ties, no social network information can be used. Second, what is the deterministic factor among users' geographical activities that reflect their friendships? Last but not least, is it possible to detect social ties between users even if they do *not* have any geographical footprints, because in real data the fraction that user pairs with friendships had ever met in geography is very low (i.e., only 3%, see Section 2)? We think analyzing geo-social interactions can bring insights for various applications, such as geo-social recommendation, anti-terrorism strategies, and computational advertising.

**Table 1: Social & geo-info used to study social ties.**

| | Social | Direct Linkage | | | | | | | Indirect |
|---|---|---|---|---|---|---|---|---|---|
| | **Links** | **FoCL** | **CoSL** | **PoCL** | **SoLH** | **PF** | **CF** | **TF** | **Co-location Graph** |
| [2] | √ | √ | | √ | | | | √ | |
| [7] | √ | √ | √ | | √ | | | √ | |
| [10] | | √ | | | √ | | | √ | |
| [5] | √ | √ | √ | √ | √ | | | | |
| [8] | √ | √ | √ | √ | √ | | | | |
| [9] | | √ | | | | √ | √ | √ | |
| **Ours** | | √ | √ | √ | √ | √ | √ | √ | √ |

Some studies showed how social ties correlate to spatio-temporal co-occurrence [2], group events [7], check-in records [1][9], location history of users [10], and geographical events [11]. Some work leverage geographical information for link prediction [3], such as co-location [5][8]. However, most of them assume a partial social network structure has been observed and used as features when analyzing the geographical properties that affect the formation of social links, or measuring the probability of being acquainted with each other. We think such assumption will prevent us from purely understanding the relationship between footprints and social ties. Therefore, we erase the social effect by setting no social ties are known in prior, and propose a robust set of features to model the direct and indirect linkages between users. To highlight the social and geographical information we consider, comparing to existing work, we create Table 1, in which the abbreviations and the meanings are described in Section 3.

It is very challenging to use solely geographical footprints of users to study social ties. The reason is three-fold. First, assume Alice and Bob do not know each other but they live in the same area and have similar moving behaviors, how to avoid treating them as friends? Second, on the contrary, assume Claire and David are friends but they *never* have interactions in geography, how to reveal their social connection? Third, different individuals have various geographical behaviors due to their personalities. One may be very active on producing her footprints (lead to a rich set records of geographical interactions with others) while another could be very down-to-earth and have very rare encountering infor-


*CIKM* '15, Oct 19-23, 2015, Melbourne, Australia

http://dx.doi.org/10.1145/2806416.2806653

mation with friends. How to represent their possible spatial behaviors that characterize social ties?

To answer the three questions, we develop a novel feature engineering framework. We represent the geographical footprints of two users considering various degree of importance on their co-occurrence from the *personal*, *collective*, and *temporal* aspects, which will be validated to deal with some of the aforementioned challenges. Besides the co-location features that depict the possibility of *direct* linkage between nodes, to model the the potential *indirect* friendships among users, we further devise a series of structural features based on a *Co-location Graph* that is constructed from co-location features of users. In the end, we conduct a systematic experiment on two well-known geo-social media datasets, GOWALLA and MEETUP, to evaluate the proposed feature engineering. framework.

# 2. DEFINITION AND DATASETS

**Definition 1: Social Network.** A social network is a graph $G = (V, E)$, where $V$ is the set of nodes and $E$ is the set of edges. Each node $v \in V$ is associated with a list of geographical footprints $a_v$.

**Definition 2: Geographical Footprint.** The geographical footprint of node $v$ is a sequence of locations and timestamps, $a_v = \langle (l_1^v, t_1^v), (l_2^v, t_2^v), ..., (l_n^v, t_n^v) \rangle$, where each location $l_i \in L_v$ is a geographic coordinate (longitude and latitude), and $L_v$ is the set of locations visited by node $v$.

**Definition 3: Co-location.** The geo-footprint $(l_i^u, t_i^u) \in a_u$ and $(l_j^v, t_j^v) \in a_v$ of nodes $u$ and $v$ can form a co-location $c = (l_k, t_k)$ if they satisfy both a spatial condition $dist(l_i^u, l_j^v) < \delta$ and a temporal condition $|t_i^u - t_j^v| < \tau$, where $\delta$ and $\tau$ are the distance and time thresholds respectively, and $l_k$ is the average geographic position of $l_i^u$ and $l_j^v$ and $t_k$ is the average time of $t_i^u$ and $t_j^v$. We also denote the set of co-locations of nodes $u$ and $v$ as $C_{uv} = \{c_1, c_2, ...\}$. Note that the settings of $\delta$ and $\tau$ empirically depends on the various applications, and this study sets $\delta = 20$ (meters) and $\tau = 1$ (hour).

**Datasets.** We use two datasets in this study, and their statistics are given in Table 2. GOWALLA[1] data, which is crawled from Sep 2011 to Nov 2011, is an online location-based social networking service that allow users to "check-in" their geographical locations and share with their friends. MEETUP[2] is crawled from an event-based social network *meetup.com* from July 2013 to Oct 2013, in which users belong to multiple online social groups, and they can publish and participate geographical events via RSVP ("yes", "no", or "maybe"). Each node represents a user, and two users are connected by an edge if they appear in the same social group at least three times. Note that we have the time of each check-in in Gowalla data, but the Meetup data does not provide event time information. Table 3 reports the numbers of node pairs related to "friends" vs. "non-friends" and "co-location" vs. "no co-location", where "co-location" indicates that two users had ever meet in terms of geography and time. We can find that friends without co-locations take the major fraction, which reflects that users who are friends *nearly never* meet in the physical world (98% for Gowalla and 97% for Meetup).

# 3. FEATURE ENGINEERING FRAMEWORK

The feature engineering has two stages. We first depict the *direct* linkages between users through their *co-location features*. By first introducing four basic features, which represent the degree of co-location between users. Then by modeling the *personal* importance and the *collective* characteristics of each location and the temporal factor between co-locations, basic features are integrated into personal (PF), collective (CF), and temporal features (TF). The second stage targets at capturing the *indirect* linkages between users who have no co-locations. We construct a *co-location graph* from co-location features of node pairs and develop three categories of structural features: graph-based personal (GPF), collective (GCF), and temporal features (GTF).

[1]https://snap.stanford.edu/data/loc-gowalla.html
[2]http://lsna2012.net76.net/ebsn/

**Table 2: Data Statistics.**

| | Gowalla | Meetup |
|---|---|---|
| **#Locations** | 1,280,969 | 5,183,840 (events) |
| **#GeoActivities** | 6,442,890 (checkins) | 42,733,136 (RSVPs) |
| **#Nodes** $\|V\|$ | 196,591 | 1,013,453 |
| **#Edges** $\|E\|$ | 1,900,654 | 34,410,754 |

**Table 3: Statistics of node pairs.**

| | Gowalla | Meetup |
|---|---|---|
| **Friends & Co-location** | 40,694 | 892,117 |
| **Friends & No Co-location** | 1,859,960 | 33,518,637 |
| **Non-Friends & Co-location** | 482,568 | 10,173,539 |
| **Total** | 2,342,528 | 44,584,293 |

## 3.1 Co-Location Features

### 3.1.1 Basic Features

Basic features model the affinity of geographical interactions between users. The idea is two-fold. First, two nodes with higher closeness, frequency, probability, and similarity for their co-locations are expected to have a higher possibility to be acquainted with each other. Second, if the locations of their meetings are more meaningful or important to both nodes, such co-location should be assigned a higher weight which indicates they have higher potential to be friends.

- *Frequency of Co-location.* Two users possess higher meeting frequency tend to have the friendship. We can use the number of their co-locations as the feature value: $FoCL(u, v) = |C_{uv}|$.
- *Closeness of Significant Locations.* If the locations *significantly* visited by two nodes are closer to one another, they have a better likelihood of being friends. We denote the *location significance* of a location $l$ for a node $u$ as $LS(u, l)$, which will be defined under different personal and collective factors in the following. Let the most significant location of node $u$ as $l^\star(u) = \arg\max_{l \in L_u} LS(u, l)$. The feature of closeness of significant locations is defined as: $CoSL(u, v) = \text{dist}(l^\star(u), l^\star(v))$, standing for the geographical distance between their most significant locations.
- *Probability of Co-location.* The co-location probability measures the possibility of two nodes visiting the same location (not necessarily at the same time). We define the co-location probability as: $PoCL(u, v) = \sum_{l \in L_u \cap L_v} LS(u, l) \times LS(v, l)$.
- *Similarity of Location History.* If the location histories of two nodes are more similar to each other, in terms of the locations themselves and the significance, they tend to be friends. We define the similarity as: $SoLH(u, v) = \sum_{l \in L_u \cap L_v} \frac{LS(u,l) \times LS(v,l)}{\|LS(u,l)\| \times \|LS(v,l)\|}$.

### 3.1.2 Personal Factor

People have different probabilities to visit a particular location. One may prefer to visit a place than another places.

Hence, it is important to consider the personal location history into the measure of location significance $LS(u,l)$. The most intuitive and effective way to model the personal factor is the probability that a node $u$ visits a location $l$ among all of its visits (at any locations), which is defined as $p_{per}(u,l) = \frac{freq(u,l)}{\sum_{l' \in L_u} freq(u,l')}$, where $freq(u,l)$ is the times of visiting of node $u$ at location $l$.

In addition to the visiting probability at a specific location, a more general way to model the personal factor is consider the probability of traveling around a *region*, because sometimes people prefer to visit a certain district containing multiple locations. For a node $u$, we use a density function with respect to the geographical distances between a certain location $l$ and all the locations visited by $u$, given by:

$$d(u,l) = \sum_{l' \in L_u} \frac{\exp(-\rho_d \cdot \mathrm{dist}(l,l'))}{freq(u,l)}, \qquad (1)$$

where $\rho_d$ is a parameter that implicitly determines the area of a region. A larger $\rho_d$ may refer to a certain suburb district (e.g. an amusement park or zoo) while a smaller $\rho_d$ may be suitable for an urban district. We choose a small $\rho_d = 1.5$ since the data we use for experiments is an urban area.

We can treat the probability $p_{per}(u,l)$ and the density $d(u,l)$ as the weight of $FoCL$, i.e., $w_{uv}^{p_{per}} = \arg\max_{l \in L_u \cap L_v} - \log(p_{per}(u,l) \cdot p_{per}(v,l))$ and $w_{uv}^{d} = \arg\max_{l \in L_u \cap L_v} - \log(d(u,l) \cdot d(v,l))$, and derive two personal features: $w_{uv}^{p_{per}} \times FoCL(u,v)$ and $w_{uv}^{d} \times FoCL(u,v)$. In addition, both $p_{per}(u,l)$ and $d(u,l)$ are considered as the measure location significance $LS(u,l)$. Therefore, there are 3 basic (i.e., $CoLS$, $PoCL$ and $SoLH$) $\times$ 2 personal (i.e., $p_{per}(u,l)$ and $d(u,l)$) = 6 features. Totally we have 8 personal features.

### 3.1.3 Collective Factor

The collective factor of a location reflects the visiting behaviors of the general public (i.e., all the users in the data). Some locations are frequently visited by a number of users such as famous landmarks and transportation sites while other locations are more specific to only a few persons, such as a private house. Those people co-locating at a *public* place tends to be strangers who encounter by coincidence. In contrast, co-locating at a *private* place has higher potential to be friends. We take advantage of *location entropy* [4] to model the collective factor of a location. We first define the probability that a node $u$ visits a location $l$ among all the visits by all the nodes in the data as: $p_{col}(u,l) = \frac{freq(u,l)}{\sum_{u \in V} \sum_{l' \in L_u} freq(u,l')}$. Then the entropy of a location $l$ can be defined as:

$$e(l) = - \sum_{\{u : u \in V \text{and} p_{col}(u,l) \neq 0\}} p_{col}(u,l) \cdot \log(p_{col}(u,l)). \qquad (2)$$

A high entropy $e(l)$ means location $l$ is a public place at which many people have ever visited while a low $e(l)$ implies $l$ tends to be a private place that is visited by few users.

Similar to the personal features, we can consider the location entropy $e(l)$ as the weight of $FoCL$, and derive a collective feature: $\sum_{l \in L_u \cap L_v} \exp(-e(l))$. In addition, $e(l)$ can be regarded as the location significance $LS(u,l)$, and thus we have three additional collective features (i.e., $CoLS$, $PoCL$ and $SoLH$). Totally we have 4 collective features.

### 3.1.4 Temporal Factor

Considering consecutive co-locations of two users, the *time gaps* of friends are different from those of strangers who meet occasionally. Strangers have higher possibility (but not absolutely) to co-locate with each other up to several times within a short time period (e.g. several times in a day), such as traveling by the same train or doing daily routines to work there and back. As for friends, on the contrary, the time gaps of their co-locations have much higher potential to be longer (e.g. once after several days or weeks). Therefore, we think the time gap is an indispensable factor to characterize the co-locations between two persons. We define the time gap between two co-locations $l_{k-1}$ and $l_k$ as: $g(l_{k-1}, l_k) = \exp(-\rho_t \cdot (t_{l_k} - t_{l_{k-1}}))$, where $\rho_t$ is a parameter that controls the extent of correlation between time gap and friendship, and should be determined by the types of co-location events. Similar to $\rho_d$, we empirically choose $\rho_t = 0.3$.

Similar to the personal and collective features, $g(l_{k-1}, l_k)$ is treated as the weight of $FoCL$: $w_{uv}^{t}(l_k) = 1 - g(l_{k-1}, l_k)$. Thus we can have a temporal feature: $\sum_{l_k \leftarrow \mathrm{sort}(L_u \cap L_v)} w_{uv}^{t}(l_k)$, where sort() orders the co-locations by time, "$\leftarrow$" gets the co-location by time, and $k > 1$.

**Table 4: The list of graph features computed on $\mathbb{G}$.**

| GF | Formula |
|---|---|
| Sum | $\sum_{z \in \Gamma_u} s(u,z) + \sum_{y \in \Gamma_v} s(v,y)$ |
| pref-att | $\sum_{z \in \Gamma_u} s(u,z) \times \sum_{y \in \Gamma_v} s(v,y)$ |
| ComNbr | $\sum_{z \in \Gamma_u \cap \Gamma_v} s(u,z) \times s(v,z)$ |
| Jaccard | $\frac{\sum_{z \in \Gamma_u \cap \Gamma_v} s(u,z) \times s(v,z)}{\sum_{z \in \Gamma_u \cup \Gamma_v} s(u,z) \times s(v,x)}$ |
| Ada-Adar | $\sum_{z \in \Gamma_u \cap \Gamma_v} \frac{1}{\log(\sum_{y \in \Gamma_z} s(z,y))}$ |
| Katz | $\sum_{len=1}^{\infty} \sum_{p \in path_{u,v}^{\langle len \rangle}} \prod_{(y,z) \in p} s(y,z)$ |
| RWR | $\frac{RWR(u,v) + RWR(v,u)}{2}$ |

## 3.2 Graph Features

The co-location features aim to characterize the relationships between users using their geographical footprints. However, not all of the users have *rich* and plentiful spatio-temporal footprints. Some may be very enthusiastic to leave their geographical activities in a digital form (e.g. via doing check-in actions or responding to R.S.V.P. of social events) while other people might have the geographical interactions with only few of their friends. In addition, users can have no *direct* co-locations with their friends (as exhibited in Table 2), and thus using purely co-location features might not be able to capture all of their social relationships. Therefore, we propose to construct a *co-location graph* from co-location features, and extract a series of graph features that refine the predictability via *transitivity* of co-location relationships. The idea is similar to link prediction [3]: if two individuals have more common co-locating persons with similar co-location features, they are more likely to be friends. In the following, we define the co-location graph and the graph features that are extended from link prediction [3].

**Definition 4: Co-location Graph.** A co-location graph is a weighted graph $\mathbb{G} = (V, \mathbb{E})$, where $V$ is the set of nodes (the same as $V$ in the network $G$), and $\mathbb{E}$ is a set of edges that represent the *co-location relationships* between nodes. Each edge $(u,v) \in \mathbb{E}$ is associated with a weight $s_{uv} = [0,1]$ that stands for the strength of co-location relationship, which is defined as a normalized feature score of nodes $u$ and $v$: $s_{uv} = \frac{\log(fs(u,v)+1)}{\log(\arg\max_{u,v \in V} fs(u,v)+1)}$, where $fs(u,v)$ is the feature score of one of the co-location features described above. Note that, to avoid too many edges being constructed, we use a threshold parameter $\mu$ to filter out edges with lower $s_{uv}$. We set $\mu = 0.2$ by default.

By extending link prediction features [3], Table 4 lists the graph features (GF) based on the co-location strength values on edges in the co-location graph $\mathbb{G}$, in which $\Gamma_u$ is the set of neighbors of node $u$ in $\mathbb{G}$, $path_{u,v}^{\langle len \rangle}$ is the set of paths of length $len$ between nodes $u$ and $v$, and $RWR(u,v)$ [6] restarts from $u$. We construct a location graph for each co-location feature, Graph-based Personal/Collective/Temporal Features (GPF/GCF/GTF), and the list of graph features is computed for each $\mathbb{G}$.

# 4. EXPERIMENTS

The experiment consists of two parts: *feature-based user pair ranking* and *social tie detection*. The former is an unsupervised method to present the effectiveness of various feature sets while the latter uses a supervised learning method to determine whether a given user pair has a friendship.

**Feature-based User Pair Ranking**. We rank all pairs of nodes in the social network using each feature set, select *top* $N$ node pairs, and compute the ratio of node pairs with social ties (termed Accuracy@$N$) by varying $N$. Note that to we simply use geometric mean to combine the feature values of each feature category (i.e., PF, CF, and TF). While the MEETUP data do not have timestamps for user meeting events, we do not report the results of TF and GTF for the MEETUP part. In addition to all user pairs, to understand whether our feature engineering model can identify users who have friendships but no geographical interactions, we also report the results of user pairs who do not have any co-location, termed by "No Co-location (NCL) user pairs." The results are shown in Table 5. We can find that for lower $N$ ($= 50$, 100), our feature model is able to near perfectly distinguish friends from non-friends user pairs. However, as $N$ increases, the distinguish ability decreases, especially being drastic for conventional features PF, CF, and TF. The accuracy values of our co-location graph-based features GPF, GCF, and GTF are much more stable, and using *all* the features can achieve the best accuracy. Moreover, by looking into the results of "no co-location user pairs," the superiority of GPF, GCF, and GTF is more clear. Since PF, CF, and GTF highly depend on co-locations of users, their accuracy values are generally very low under this case. Such results indicate that the proposed graph-based features are capable of identifying friends who have no geographical meetings. Nevertheless, the accuracy values of $N > 1000$ are still very low. Hence we resort to the supervised learning to detect whether any given user pair has a social tie in the following.

**Social Tie Detection.** We divide node pairs in the social network into $R\%$ training instances and $1 - R\%$ testing instances ($R = 20, 80$). The *Area under the ROC Curve* (AUC) are used as evaluation metrics. Node pairs with social ties (ground truth) are treated as *positive* instances. We use Random Forest, Support Vector Machine, and Logistic Regression as the supervised learning methods. Since different learning methods exhibit very similar results, here we report the results of Random Forest due to the page limit. The results are shown in Table 6. We can find three things. First, the GPF, GCG, and GTF still generally outperform PF, CF, and TF, especially on "no co-location user pairs." Second, using only geographical footprints (no social network information are employed) can achieve up to 70% AUC. This result tells us that social ties are able to be successfully detected through indirect linkages in the proposed co-location graph. Third, even using 20% pairs as the training set can lead to AUC around 60%.

**Table 5: Results of feature-based user pair ranking.**

| | Top-$N$ Ranked User Pairs | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gowalla | | | | | | | Meetup | | | | |
| | PF | CF | TF | GPF | GCF | GTF | All | PF | CF | GPF | GCF | All |
| 50 | **0.90** | **0.97** | **0.98** | **0.96** | **0.98** | **0.98** | **0.99** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| 100 | **0.95** | **0.88** | **0.96** | **0.97** | **0.93** | **0.95** | **0.98** | **0.98** | **0.98** | **1.00** | **0.99** | **1.00** |
| 500 | 0.70 | 0.62 | 0.71 | **0.85** | 0.77 | 0.79 | **0.88** | **0.96** | **0.96** | **0.98** | **0.96** | **0.98** |
| 1000 | 0.43 | 0.38 | 0.43 | 0.64 | 0.56 | 0.55 | 0.68 | 0.64 | 0.60 | 0.78 | 0.73 | **0.81** |
| 5000 | 0.12 | 0.11 | 0.10 | 0.51 | 0.47 | 0.44 | 0.55 | 0.31 | 0.29 | 0.62 | 0.59 | 0.65 |
| 10000 | 0.07 | 0.06 | 0.06 | 0.33 | 0.28 | 0.27 | 0.35 | 0.26 | 0.26 | 0.39 | 0.33 | 0.43 |
| | Top-$N$ Ranked "No Co-Location" User Pairs | | | | | | | | | | | |
| | Gowalla | | | | | | | Meetup | | | | |
| | PF | CF | TF | GPF | GCF | GTF | All | PF | CF | GPF | GCF | All |
| 50 | 0.000 | 0.000 | 0.000 | **0.62** | **0.58** | **0.60** | **0.69** | 0.000 | 0.000 | 0.46 | 0.44 | 0.49 |
| 100 | 0.020 | 0.010 | 0.010 | **0.53** | 0.47 | **0.51** | **0.61** | 0.000 | 0.000 | 0.38 | 0.37 | 0.41 |
| 500 | 0.008 | 0.006 | 0.005 | 0.44 | 0.41 | 0.39 | 0.46 | 0.003 | 0.002 | 0.32 | 0.32 | 0.35 |
| 1000 | 0.007 | 0.004 | 0.006 | 0.23 | 0.21 | 0.22 | 0.29 | 0.005 | 0.004 | 0.18 | 0.17 | 0.20 |
| 5000 | 0.005 | 0.004 | 0.005 | 0.16 | 0.15 | 0.11 | 0.19 | 0.003 | 0.003 | 0.11 | 0.10 | 0.14 |
| 10000 | 0.007 | 0.007 | 0.006 | 0.13 | 0.11 | 0.12 | 0.15 | 0.004 | 0.003 | 0.09 | 0.08 | 0.13 |

**Table 6: Results of AUC for social tie detection.**

| | **80% Training, 20% Testing** | | | | **20% Training, 80% Testing** | | | |
|---|---|---|---|---|---|---|---|---|
| | User Pairs | | **NCL User Pairs** | | User Pairs | | **NCL User Pairs** | |
| | Gowalla | Meetup | Gowalla | Meetup | Gowalla | Meetup | Gowalla | Meetup |
| **PF** | 0.66 | **0.73** | 0.20 | 0.02 | 0.48 | **0.64** | 0.16 | 0.02 |
| **CF** | 0.62 | **0.74** | 0.16 | 0.02 | 0.43 | **0.64** | 0.15 | 0.02 |
| **TF** | 0.65 | N/A | 0.14 | N/A | 0.47 | N/A | 0.11 | N/A |
| **GPF** | **0.73** | **0.75** | 0.66 | 0.68 | **0.66** | **0.67** | 0.52 | **0.60** |
| **GCF** | **0.70** | **0.76** | 0.67 | 0.69 | **0.64** | **0.68** | 0.57 | **0.60** |
| **GTF** | **0.71** | N/A | 0.65 | N/A | **0.66** | N/A | 0.47 | N/A |
| **All** | **0.74** | **0.77** | 0.68 | 0.69 | **0.68** | **0.68** | 0.59 | **0.63** |

# 5. CONCLUSIONS

This paper analyzes social ties using solely geographical footprints of users in GOWALLA and MEETUP data, considering the observation that friends are quite rare to meet in the physical world. We develop a novel feature engineering model composed by a robust set of co-location and structural features. We believe our feature model can benefit for both link prediction and recommendation tasks to not only further boost the prediction accuracy but also deal with the link sparsity problem in cold-start scenarios.

# 6. ACKNOWLEDGEMENT

This work was sponsored by Ministry of Science and Technology of Taiwan under grant 104-2221-E-001-027-MY2. This work is also supported by Multidisciplinary Health Cloud Research Program: Technology Development and Application of Big Health Data, Academia Sinica, Taipei, Taiwan under grant MP10212-0318.

# 7. REFERENCES

[1] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD*, 2011.

[2] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 2010.

[3] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

[4] H. Pham, C. Shahabi, and Y. Liu. Ebm: An entropy-based model to infer social strength from spatiotemporal data. In *SIGMOD*, 2013.

[5] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *KDD*, 2011.

[6] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *ICDM*, 2006.

[7] C. Wang, M. Ye, and W.-C. Lee. From face-to-face gathering to social structure. In *CIKM*, 2012.

[8] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *KDD*, 2011.

[9] H. Wang, Z. Li, and W.-C. Lee. Pgt: Measuring mobility relationship using personal, global and temporal factors. In *ICDM*, 2014.

[10] 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*, 5(1), 2014.

[11] P. Yin, Q. He, X. Liu, and W.-C. Lee. It takes two to tango: Exploring social tie development with both online and offline interactions. In *SDM*, 2014.