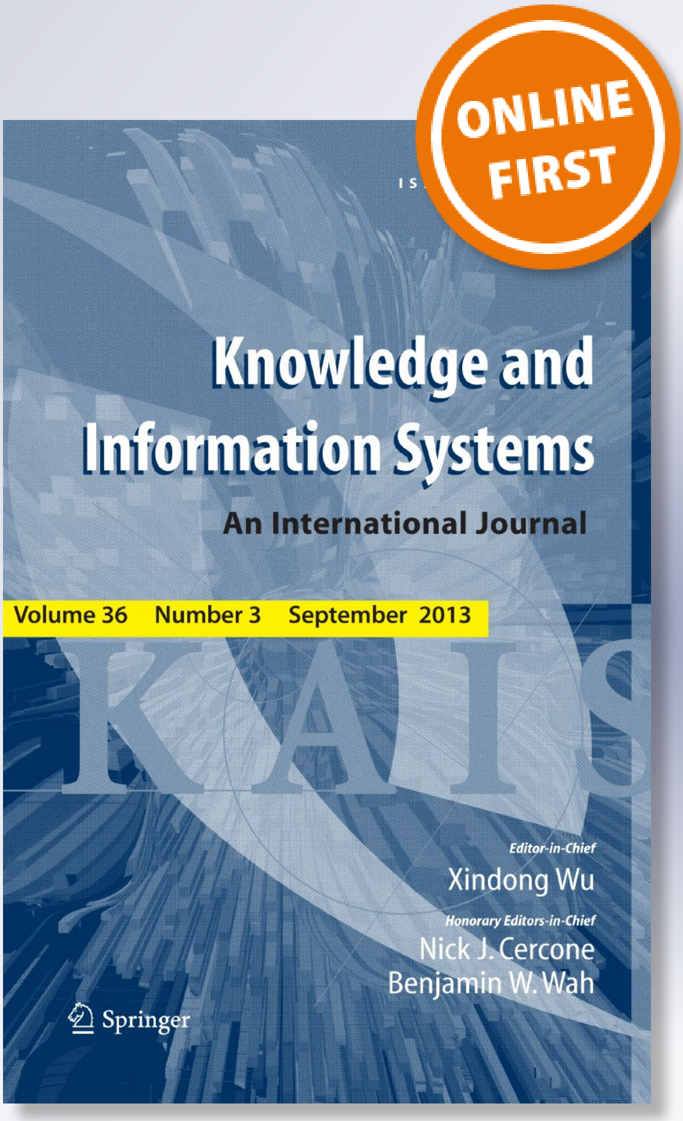
An active learning-based approach for location-aware acquaintance inference

Bo-Heng Chen, Cheng-Te Li, Kun-Ta Chuang, Jun Pang & Yang Zhang

Knowledge and Information Systems
An International Journal

ISSN 0219-1377

Knowl Inf Syst
DOI 10.1007/s10115-018-1196-8



Knowl Inf Syst
https://doi.org/10.1007/s10115-018-1196-8

REGULAR PAPER

An active learning-based approach for location-aware acquaintance inference

Bo-Heng Chen[1,2] · **Cheng-Te Li**[3,6] · **Kun-Ta Chuang**[2] · **Jun Pang**[4] · **Yang Zhang**[5]

Received: 4 March 2017 / Revised: 6 March 2018 / Accepted: 18 April 2018

Abstract With the popularity of mobile devices and various sensors, the local geographical activities of human beings can be easily accessed than ever. Yet due to the privacy concern, it is difficult to acquire the social connections among people possessed by services providers, which can benefit applications such as identifying terrorists and recommender systems. In this paper, we propose the *location-aware acquaintance inference* (LAI) problem, which aims at finding the acquaintances for any given query individual based on *solely* people's *local* geographical activities, such as geo-tagged posts in Instagram and meeting events in Meetup, within a targeted geo-spatial area. We propose to leverage the concept of *active learning* to tackle the LAI problem. We develop a novel semi-supervised model, *active learning-enhanced random walk* (ARW), which imposes the idea of active learning into

✉ Kun-Ta Chuang
ktchuang@mail.ncku.edu.tw

Bo-Heng Chen
bhchen@netdb.csie.ncku.edu.tw

Cheng-Te Li
chengte@mail.ncku.edu.tw

Jun Pang
jun.pang@uni.lu

Yang Zhang
yang.zhang@uni-saarland.de

[1] Graduate Program of Multimedia Systems and Intelligent Computing, National Cheng Kung University and Academia Sinica, Tainan, Taiwan

[2] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

[3] Department of Statistics, National Cheng Kung University, Tainan, Taiwan

[4] Faculty of Science, Technology and Communication, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[5] CISPA, Saarland Informatics Campus, Saarbrücken, Germany

[6] Institute of Data Science, National Cheng Kung University, Tainan, Taiwan

Published online: 03 May 2018

the technique of random walk with restart (RWR) in an *activity graph*. Specifically, we devise a series of *candidate selection* strategies to select unlabeled individuals for labeling and perform the different *graph refinement* mechanisms that reflect the labeling feedback to guide the RWR random surfer. Experiments conducted on Instagram and Meetup datasets exhibit the promising performance, compared with a set of state-of-the-art methods. With a series of empirical settings, ARW is demonstrated to derive satisfying results of acquaintance inference in different real scenarios.

Keywords Active learning · Acquaintance inference · Location-based social network · Check-in data

1 Introduction

With the maturity of information and communication technology, mobile devices have become popular and diverse kinds of sensors have been deployed ubiquitously. Mobile devices, such as smartphone and digital camera, allow people to record where they have visited in the form of geographical locations in online social platforms. For example, users can share *geo-tagged* postings, check-ins, and photographs on Twitter and Instagram. Similarly, the deployed sensors in the context of the Internet of Things (IoT), such as surveillance systems and various kinds of smart cards, can also depict the geographical activities of human beings in the physical world. Such geo-spatial footprints provide researchers an unprecedented opportunity to study human mobility and in particular the interaction between mobility and social relations.

This paper aims at *actively* inferring the acquaintances of any given person *solely* based on her geographical activities and interactions with other people within a certain *physical* area. Let us use two real scenarios to elaborate the motivation of this work and the proposed approach.

Homeland Security Analyzing the occurrence of suspected individuals in the geography space has been reported to be effective in discovering terrorist networks [27]. With partially identified interactions between terrorists, it is possible to construct the terrorist network through the geo-spatial analysis by human investigators [26,37]. The human investigation is *interactive*: when a person is suspected, the agency will further inspect in the corresponding *local* area to examine whether it is *positively* or *negatively* identified [27]. The positive identification will lead to a progressive construction of the terrorist network, while negative cases can also prevent the investigation from false alarms and a waste of effort. However, such a process is normally tedious and requires a tremendous amount of resources.

Geo-Social Recommendation It is evident that social relationships can help to improve the performance of location-aware recommendation systems [17]. In the literature, researchers have integrated social relation into collaborative filtering-based approaches [15,49] to gain effective recommendation. Most of these studies presume that the social information is available in online platforms, while due to the privacy concern, the recommenders normally have no access to the social relations among users, which are only possessed by the service providers. Hence, the real-world location recommenders can either solely exploit users' visiting histories as input or simultaneously infer friendships and use it in recommending locations [33]. In practice, users can provide *feedback* to the recommenders by implicitly returning which of the recommended users are true friends so that the satisfaction can get improved. With the inference of social acquaintances with users' interactive feedback in this work, we have a better location-aware recommender when social information is unavailable.

Given the geographical activities of the population of a targeted geo-spatial area, and a given individual u, our goal is to find u's acquaintances in an interactive manner. We term such a task as *location-aware acquaintance inference* (LAI) problem. In this paper, the geographical activities refer to where, who and when people co-locate and meet. For example, as evaluated in this work, the geo-tagged posts in Instagram and the meeting events in Meetup are considered as the so-called geographical activities. In addition, both individuals and geographical activities may be associated with a set of *attributes*, referred to be as *profile attributes* and *activity attributes*. In this paper, profile attributes are users' interests, attributes or tags, and activity attributes are the tags or the categories of the encountering locations.

In the literature, the most relevant problems are *link prediction* [1,20,23] and *social tie inference* [6,34,43]. Essentially, the prediction of future connections in the link prediction task is derived solely based on the provided partial graph, but in our work, the model is considered without any social link given beforehand. In addition, social tie inference is proposed to infer the strong connection which is not tagged in the given network. Similarly, methodologies of social tie inference also assume that there are a collection of observed social ties. Few recent studies [31,44] started exploring the inference of social ties without existing link observations based on check-in data. However, they need the entire *global* geo-spatial footprints of people, which is impractical in real scenarios as aforementioned. Besides, they do not allow interactive/active inference, which are the goals of this paper.

It is challenging to infer the acquaintances of a given individual using solely the geographical activities of users within a local geo-spatial area. The reasons are threefold. The first is the *encountering uncertainty* issue. Non-friends could meet, while friends may not encounter geographically. Suppose that Alice and Bob are not friends, but they have similar location visiting histories, how can we avoid treating them as acquaintances? On the contrary, if Claire and David are friends, but they never share geo-spatial records together, how can we find evidence to show their social connection? Second, the assumption of *no* social tie observation makes supervised learning inference model infeasible. On the other hand, unsupervised learning approaches based on solely geo-spatial footprints of users lead to unsatisfying performance [31,44]. Third, the *global* geo-spatial footprints of users (i.e., around the world) have shown to be promising on acquaintance inference [44], but what we have are the collection of *local* footprints, i.e., users' geographical activities of a certain area (e.g., within a city). Using purely local footprints makes it difficult to differentiate acquaintance and non-acquaintance user pairs. We provide the observations on real data in Fig. 1, which depicts the cumulative distribution functions (CDF) of the global-local score proposed by PGT [44] for all the acquaintance and non-acquaintance user pairs of three cities in the Instagram dataset. (The detail data description is discussed in Sect. 5.) The nearly overlapping between two curves indicates the difficulty in separating the acquaintance and non-acquaintance user pairs. Without discriminative information from global footprints, the extension of previous models for acquaintance inference will clearly suffer from the insufficient quality.

To address these challenges, in this paper, we propose a semi-supervised learning method, namely *active learning-enhanced random walk* (ARW), for location-aware acquaintance inference. ARW consists of three parts. First, ARW is based on the technique of *random walk with restart* (RWR). We exploit RWR to realize the intuitions about the correlation between social acquaintances and geographical activities: two persons tend to be acquainted with each other if they involve in more common activities, possess more common profile attributes and activity attributes and visit more common locations. We propose to construct an *activity graph* to represent the interactions between individuals and geographical activities

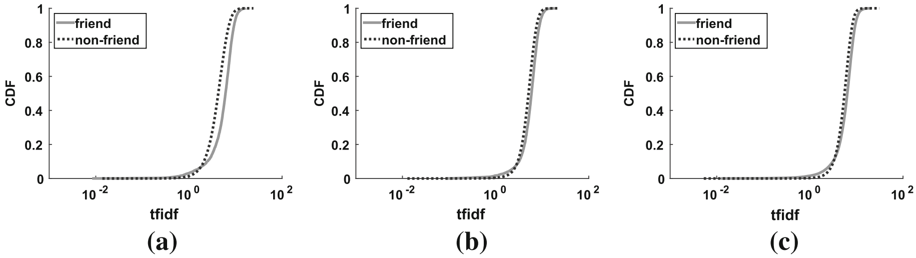


Fig. 1 Relationship differentiation observed in Instagram. **a** San Francisco, **b** London, **c** New York

so that RWR can be performed to realize aforementioned intuitions. Second, we leverage the concept of *active learning* to tackle these discussed challenges and to enable the interactive acquaintance inference. In active learning, investigators or domain experts are allowed to provide labels for a *small* set of selected instances. Motivated by this concept, we propose a series of methods to select individuals to be labeled round by round. The labeling will answer whether or not the selected individual is acquainted with the queried individual based on some external investigations. Third, while the labeling feedback is returned at each round, we propose to iteratively adjust the RWR mechanism so that the new acquaintance (positive feedback) or non-acquaintance (negative feedback) information can be incorporated. The adjustment is designed to boost the possibility that the random surfer reaches the true acquaintance, and to prevent the random surfer from arriving at more non-acquaintances. We propose the mechanisms of *graph refinement* to incorporate the positive and negative feedback, respectively.

It is worthwhile noticing that one may question how can an acquaintance inference method distinguish the terrorists a target is connected to from the guys who sell coffee to the target terrorist every day. In fact, everyone has a chance to be the acquaintance of another person, and thus, we do not explicitly make such a distinction in this work. Nevertheless, if the goal is to identify the acquainted terrorists of a given target, the labeling feedback via active learning will help lower down the possibility the coffee salesclerk being inferred as the acquaintance. The guy who sells coffee to the terrorist every day may be the bridge for terrorists' communication, and can be captured by the activity graph to boost the performance. In short, with active learning, we can prevent the investigation from false alarms and a waste of effort.

We summarize the contributions of this work as the following.

- While it is practical and challenging to infer social acquaintances without observed social ties based on solely geographical activities, this work is the first attempt to perform acquaintance inference with active learning.
- We formulate the location-aware acquaintance inference (LAI) problem. The LAI problem is technically re-formulated as how to select a small set of individuals such that their labeling can lead to the best inference performance in the setting of active learning.
- We develop the ARW model, which seamlessly integrates random walk with active learning. While a heterogeneous information graph is designed as the fundamental of ARW, ARW can be regarded as a generalized framework for link recommendation tasks in various applications.
- The experimental studies show that the proposed ARW model can deliver promising results in terms of accuracy as compared with several state-of-the-art methods. A robust

set of empirical settings are examined, and the results demonstrate the usefulness of ARW.

The remainder of this paper is organized as follows. We first review the relevant studies in Sect. 2 and then give the problem formulation in Sect. 3. In Sect. 4, we present the proposed ARW framework. The experimental results are exhibited and discussed in Sect. 5. Last, we conclude this paper in Sect. 6.

2 Related work

Some existing studies are related to our work such as *inference and analysis of social ties, link prediction, and relationship between social ties and mobility*. We create Table 1 to summarize the past relevant studies, and to distinguish our work from them in the research line of acquaintance inference. It can be observed that our work is the first attempt to perform acquaintance inference based on geo-activities with active learning. The relevant studies are briefly described as follows.

Inference and Analysis of Social Ties Existing studies have investigated how to construct and infer the link structures of social networks based on a variety of data sources, including users' self-reports [45], email communication [8], call data records [11] and contagion spread [29]. Recent studies attempt to characterize and uncover the hidden connections between social ties and various kinds of spatial footprints of users, including spatiotemporal co-occurrence events [9], group events [42], check-in records [7,44] and location histories of users [46], in which various supervised learning methods are proposed for the inference of social ties. Some papers have also explored how friends in online social networks affect users' offline geographical activities [22,47] and their offline check-ins [7]. However, the acquaintance inference using **purely** geographical activities of users under an interactive setting has not been investigated yet. We aim to exploit the concept of active learning, together with an unsupervised mechanism, to fulfill the task of interactive acquaintance inference.

Link Prediction is to predict social ties in the future, given the current snapshot of a social network [20]. There are numerous studies that propose various kinds of social network features and supervised learning methods (e.g., please refer to the survey paper [23]) to predict future social links. Some recent works further leverage the geographical information of users to boost the performance of link prediction, including co-location [34,43] and call data records [28]. Our problem setting is different from the conventional link prediction problem in two aspects. First, we aim to infer the acquaintance links for a given user without any observed social ties. Specifically, what we have for the inference is the set of local geographical activities for each user. Such setting makes supervised learning infeasible. Second, what we aim to present is an interactive inference mechanism, in which the third parties or the investigators can involve in the inference process by providing feedback. Although the technique of active learning has been used for link prediction [5] and link-type inference [50], their proposed methods highly rely on sufficient volume of training instances to build effective supervised learning models.

Relationship between Social Ties and Mobility A number of existing studies have investigated the relationships between social connections and their visited locations [6,10,14,31,44]. Cheng et al. predict whether two individuals are friends based on their mobility information [6]. They simultaneously consider the co-occurrences and their visiting time intervals. Cranshaw [10] et al. introduce a set of location-based features such as location entropy for

Table 1 Comparison with relevant studies

	Geo-activities	Labeled data	Approach	Active learning
[1]		✓	Supervised	
[50]		✓	Supervised	✓
[6,34,43]	✓	✓	Supervised	
[23]		✓	Supervised	✓
[4]		✓	Supervised	
[46]	✓	✓	Supervised	
[31,44]	✓		Unsupervised	
Ours	✓		Semi-supervised	✓

analyzing the social context of a geographical region and devise a model to predict friendships between users using their location trails. EBM [31] not only infers social connections but also estimates the strength of social connections by analyzing people's co-occurrences using check-in data. PGT [44] is a unified framework that combines personal, global and temporal factors to measure the mobility-based social relationships between users. In addition, Hsieh et al. [14] further develop a two-stage feature engineering by identifying the direct and indirect linkages between users according to a devised check-in co-location graph.

Random Walk-based Recommendation Random walk-based mechanisms had been widely used for various recommendation tasks, such as user recommendation [2,18], activity recommendation [21] and location recommendation [3,48]. For user recommendation, Bagci [2] et al. consider social relations, personal preferences and visited locations of users to provide the personalized friend recommendation by executing random walks in a hybrid graph. Li [18] et al. modify the random walk algorithm to jointly learn how co-authorships and temporal factors affect collaborator recommendation. On the other hand, Liu [21] et al. construct a hybrid network to represent multiple types of entities in an event-based social network and run a modified random walk method for event recommendation. In addition, Ying [48] et al. and CLoRW [3] further devise the extended random walk algorithms to model social ties, user preferences, check-in historical data and location popularity for POI recommendation.

It is worthwhile to mention that research on acquaintance inference is similar to research on user recommendation. Both tasks are to predict or to find the potential individuals who have the highest probabilities to connect with a given user in the context of social networks. However, their purposes are orthogonal. The acquaintance inference aims at inferring the acquaintances of any given person based on her historical data and interactions with other people. Hence, the historical records of users are used to infer which two users are acquainted with each other. In empirical studies, researchers generally exploit social relationships in the current time stamp to validate the inferring results [31,44]. On the other hand, user recommendation, which is also termed "link prediction" [1,23,34], aims at exploiting the historical data to predict whether two users will be connected with each other in the future. As such, in the empirical studies, the recommendation results are generally validated by the testbed consisting of future social links [2,18].

Last, it is also important to emphasize that we aim at estimating the structural "proximity" between two users, instead of estimating the "similarity" of two users, in a heterogeneous information network. Users possessing higher relational structural similarity (e.g., measured by PathSim [39] and SimRank [16]) to one another tends to reflect that they play similar roles in the network, rather than exhibiting the potential connections between them. In other words, similarity measure based on the network cannot quantify the degree of being acquainted with each other. The pathways via direct and indirect visited locations, attributes and users can depict the possibility that two users are friends. If two users with more, shorter and denser pathways toward one another, they have higher potential to be friends (and thus derive higher proximity score). The random walk with restart (RWR) [41] had been validated to be an effective measure to estimate the proximity between nodes in a network [20,30,38,40].

3 Problem statement

Here, we formally describe *geographical activities*, *user profile* and *social acquaintances*. We use $U = \{u_1, \ldots, u_n\}$ to represent the entire collection of users where n is the number of users. The geographical activities of a user u_i are defined as $\mathcal{A}_i = \langle a_{i,1}, a_{i,2}, \ldots, a_{i,m} \rangle$ where $a_{i,j}$ denotes user u_i's j-th geographical activity. Each activity $a_{i,j}$ is a tuple in the form of $a_{i,j} = (place, event, time)$, where $place$ can be either a geographical position or the name of the place, $event$ indicates the action that u_i performs, and $time$ is the time stamp that u_i participates in the activity. Each $place$ and each $event$ can be associated with a set of $attributes$ that depict its semantics, in which place attributes can be venue categories, while event attributes can be the set of associated tags. The implementation of activities depends on our datasets. For example, in Instagram, "u_i visits Time Square at 2 pm, August 1st, 2016" can be represented as ("Time Square," check-in, 2016-08-01 14:00), along with the place attribute {Plaza, Landmark}. In Meetup, "u_i attends a data-mining study group at Starbucks (750 7th Ave) at 8 pm, August 15th, 2016" is represented as ("Starbucks (750 7th Ave)," Meetup, 2016-08-15 20:00), together with the set of attributes {coffee shop, data-mining study}. $\mathcal{A} = \{\mathcal{A}_i \mid u_i \in U\}$ is used to represent all users' activities.

The user profile of user u_i is defined by a tuple in the form: $\mathcal{D}_i = (uid, C)$, where uid is the identity of u_i (e.g., name), and C is the set of personal attributes or tags, such as gender, age, interests and hometown. For example, the user profile of u_i can be ("Tom," {male, 20, movie, music, basketball}). $\mathcal{D}$ contains all users' profiles in the dataset. Finally, the social acquaintance of user u_i is defined as: $F_i = \Gamma^G(u_i)$, where G is users' underlying social network that is supposed to be unobserved, and $\Gamma^G(u_i)$ returns the set of neighbors of user u_i in G.

Definition 1 *Location-aware acquaintance inference problem (LAI).* Given a query user $u_q \in U$, and the entire collection of users U, along with their geographical activities $\mathcal{A}$ and user profiles $\mathcal{D}$, the LAI problem is to find the set of social acquaintances F_q from U for user u_q. Here all of the geographical activities are presumed to be observed within a certain geo-spatial region.

Motivated by the real scenarios of Homeland Security and Geo-social Recommendation in Sect. 1, we propose to take advantage of the concept of *active learning* for dealing with these challenges so that the LAI problem can be successfully solved. Recall that the idea of active learning is to allow some external investigation to provide labeling feedback on a small set of selected instances. Under the setting of supervised learning, different kinds of methods are proposed to pick a few instances so that the predictor can be better trained. Here we formally apply the concept of active learning to solve the LAI problem under a semi-supervised setting. The labeling feedback returned by the external investigation is whether a picked user is acquainted (positive feedback) or unacquainted (negative feedback) with the query user. Given an unsupervised learning model $\mathcal{M}$ to tackle LAI, the problem can be alternatively formulated as follows.

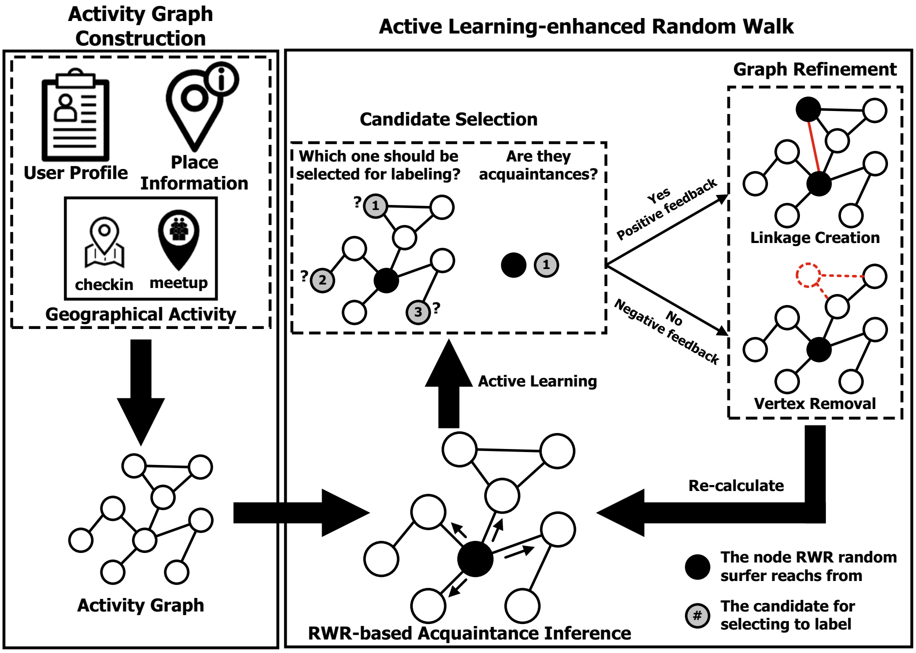


Fig. 2 The ARW approach overview

Definition 2 *Active location-aware acquaintance inference (A-LAI).* Given a query user u_q, the entire collection of users U, along with their geographical activities $\mathcal{A}$ in a local geo-spatial region and user profiles $\mathcal{D}$, an unsupervised learning model $\mathcal{M}$ and a number b of labeling budget, the A-LAI problem is to select a set of b users (from U) to be labeled such that the set of inferred social acquaintances F_i by $\mathcal{M}$ can get better performance.

4 Methodology

In this section, we present the proposed *ARW* framework to tackle LAI problem, as shown in Fig. 2. ARW consists of two major components: (a) *activity graph construction*, (b) *active learning-enhanced random walk*. We first give an overview of the components and then discuss the technical details in the following sections. In the first component, user profile, place information and geographical activity are exploited to construct a heterogeneous graph. The relationships among people, attributes, activities and places are represented by a proposed activity graph, as shown in left side of Fig. 2. In the second component, as presented in the right side of Fig. 2, equipped with the activity graph, the random walk mechanism can be used for acquaintance inference. The query user has higher potential to acquaint with the user who has higher score obtained from the random walk mechanism. In addition, we propose the active learning-enhanced random walk to enable the interactive acquaintance inference. The active learning-enhanced random walk is designed to select one candidate user for labeling at each time. According to the labeling feedback, **linkage creation** and **vertex removal** will be performed in the process of graph refinement so that the positive feedback and negative feedbacks can be incorporated into the activity graph and the random walk can reflect the

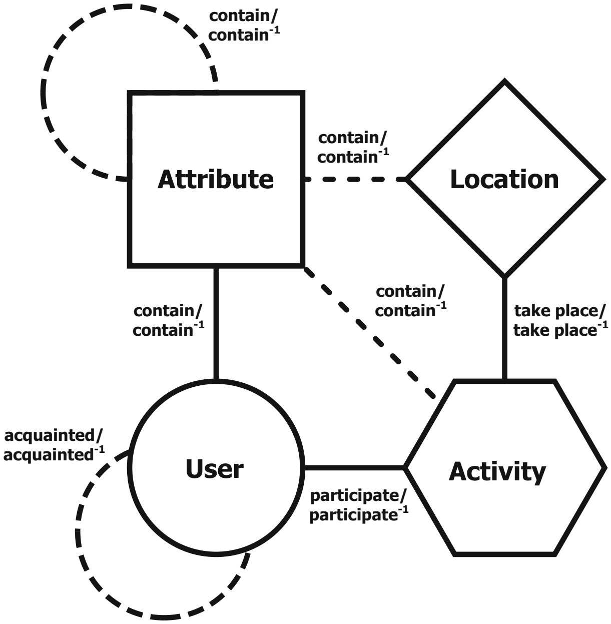


Fig. 3 The generalized activity schema

labeling outcomes. In **linkage creation**, the red dotted line will be added to the graph for linking two black nodes. Then, two black nodes will be the starting node in the random walk mechanism. In **vertex removal**, the red dotted circle and red dotted line will be removed. When the graph refinement is completed, the scores are re-calculated by the random walk mechanism. The user possessing the highest score is considered as an inferred acquaintance for the query user.

4.1 Activity graph construction

Definition 3 *Activity Graph* An activity graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is an undirected graph, where $\mathcal{V}$ and $\mathcal{E}$ are sets of nodes and edges, respectively. There are two mapping functions in light of *activity schema*, whose construction is based on dataset. One is node-type mapping function $\eta : \mathcal{V} \to \mathcal{T}$. The other is edge-type mapping function $\xi : \mathcal{E} \to \mathcal{R}$. Each node $v \in \mathcal{V}$ belongs to one particular node type $\eta(v) \in \mathcal{T}$, and each edge $e \in \mathcal{E}$ belongs to a particular relation $\xi(e) \in \mathcal{R}$. On top of that, due to the types of nodes $|\mathcal{T}| > 1$ and the types of relations $|\mathcal{R}| > 1$, activity graph can be referred as heterogeneous information network as well.

Definition 4 *Activity Schema* The activity schema $S_{\mathcal{H}}(\mathcal{T}, \mathcal{R})$ is a meta-template for an activity graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, along with a node-type mapping function $\eta : \mathcal{V} \to \mathcal{T}$, and an edge-type mapping function $\xi : \mathcal{E} \to \mathcal{R}$, where $\mathcal{T}$ is the set of node types and $\mathcal{R}$ is the set of relations between node types.

The activity schema represents the relationships between different types of nodes. An activity graph $\mathcal{H}$ is constructed based on the corresponding activity schema $S_{\mathcal{H}}$. A *generalized* activity schema is shown in Fig. 3, in which there are four kinds of entities representing `User`, `Activity`, `Location` and `Attribute`. In addition, seven different relations depict how four entities interact with one another: `User` *participates in* `Activity`, `Activity`

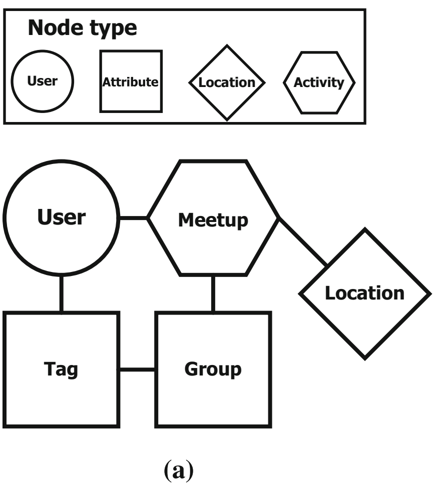


Fig. 4 Two different activity schemas. **a** Meetup, **b** Instagram

takes place in Location, User, Activity and Location *contain* Attribute, and Attribute also can *contain* Attribute. Note that the real activity schema depends on the dataset, that says, various datasets may have the different subsets of such seven relations in their schemas. In the schema, the solid lines represent that these relations would exist in each of specific activity schema. The dashed lines represent that the existence of these relations depends on the dataset. Moreover, the relations between User nodes will be added when the labeling feedback is positive in the process of active learning-enhanced random walk, which is discussed in Sect. 4.3.

We give two examples of the activity schemas using Instagram and Meetup datasets, as shown in Figure 4. A **Meetup schema** is shown in Figure 4(a), which contains five entities. The correspondence is User: "User," Attribute: {"Tag," "Group"}, Activity: "Meetup" and Location: "Location." Since a group can contain a set of tags, there is a relation between "Tag" and "Group." An **Instagram schema** is presented in Fig. 4b, in which there are seven entities. The correspondence is User: "User," Attribute: {"Gender," "Age," "Race," "Location Category"}, Activity: "Check-in," and Location: "Location."

Note that the activity graph is an unweighted graph. It is true that we can consider weights for edges in the heterogeneous information network. The reason we use unweighted edges is threefold. First, we aim at concentrating our technical contribution on designing the active learning process via heterogeneous network structure, instead of accurately estimating how different entity types interact with one another. Second, it is difficult to define a general but suitable edge weighting mechanism for every pair of entity types and have them well normalized, because the definition of edge weighting is usually data-dependent. Third, even a certain edge weighting method is applied, it will be quite time-consuming to re-calculate all edge weights in each round of active learning.

4.2 RWR-based acquaintance inference

Equipped with the activity graph, we propose to exploit one of the well-known graph-based node ranking algorithms, *random walk with restart* (RWR) [41], to be the fundamental of acquaintance inference. Given an activity graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and a query user node $u_q \in \mathcal{V}$, RWR will produce an *proximity* score p_{v_i} for every node $u_i \in \mathcal{V}$. The proximity score estimates the probability that a random surfer reaches u_i starting from u_q in $\mathcal{H}$. We consider that the higher proximity score u_i obtains, the higher potential that u_i is acquainted with u_q. User u_i can get higher proximity score p_{v_i} if u_i and u_q (a) have more common neighbors, (b) have shorter graph distance to reach each other and (c) have more shorter paths connecting one another in the graph. These facts correspondingly reflect the acquaintanceship between two users in the activity graph: involving in more common activities, possessing more common profile attributes and activity attributes and visiting more common locations. Hence, RWR is employed as the basis of our approach.

Given an activity graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and a query user node $u_q \in \mathcal{V}$, the computation of proximity scores for nodes can be formulated by $\vec{p}_{v_i} = (1 - \alpha)\tilde{M}\vec{p}_{v_i} + \alpha\vec{e}_{v_i}$, where $\tilde{M} \in \mathbb{R}^{n \times n}$ is the transition matrix of $\mathcal{H}$. Each element in $\tilde{M}$, denoted by $m_{i,j}$, is defined by $m_{i,j} = \frac{1}{d(v_i)}$, where $d(v_i)$ be the degree of v_i if there is an edge connecting u_i and u_j; otherwise, $m_{i,j} = 0$. In addition, $\vec{e}_{v_q}$ is a unit vector of the starting indicator of node v_q, and $\alpha \in [0, 1]$ is the restart probability in RWR (generally, $\alpha = 0.15$).

Algorithm 1 RWR-based Acquaintance Inference

```
Require: $\mathcal{H} = (\mathcal{V}, \mathcal{E})$: the activity graph; $u_q$: the query user node; $\alpha$: the restart probability; $N$: the number
    of acquaintances to be returned
Ensure: $F_q$: the acquaintance set of $u_q$
1: Construct the transition matrix $\tilde{M}_{\mathcal{H}}$;
2: Initialize $\vec{p}_{u_q} \leftarrow \vec{e}_{u_q}$;
3: while $\vec{p}_{u_q}$ has not converged do
4:     $\vec{p}_{u_q} \leftarrow (1 - \alpha)\tilde{M}_{\mathcal{H}}\vec{p}_{u_q} + \alpha\vec{e}_{u_q}$;
5: for $i$=1 to $N$ do
6:     $u_i^\star \leftarrow \arg\max^{(i)}_{u_i \in U \setminus \{u_q\} \wedge u_i \in \mathcal{U}} \vec{p}_{u_q}(u_i)$;
7:     $F_q \leftarrow F_q \cup \{u_i^\star\}$;
8: return $F_q$;
```

We present the RWR-based acquaintance inference method in Algorithm 1. First, we initialize the proximity vector and transition matrix (lines 1–2). Then RWR iteratively derives the proximity score of every node u_i (lines 3–4) with respect to the query user u_q. Since we are finding u_q's acquaintances, we examine the proximity scores of all the `User` nodes and return the top-N users who possess highest proximity scores (lines 5–7).

4.3 Active learning-enhanced random walk

Inspired by the interactive exploration of social acquaintances in the real world, in which some external investigation can provide feedback during the exploration, we aim at implementing this idea in our framework using the concept of active learning. In active learning [35], a supervised learning model usually employs a certain query strategy to select a small set of instances, which are sent to some external investigation for labeling. The external agency can provide feedback by answering the label of each picked instance. The key point lies in how to select the most effective data instances to be labeled such that the performance of the supervised model can be significantly boosted.

To impose active learning into the RWR-based acquaintance inference, we consider all the user nodes $u_i \in U$ ($u_i \neq u_q$) as the candidate acquaintances of the query user u_q. The active learning-enhanced random walk is designed to be an iterative process. Each iteration will output an inferred acquaintance. During each iteration, there are three phases: (1) proximity calculation, (2) candidate selection and (3) graph refinement, which are elaborated in the following.

Proximity Calculation measures the potential of users being the acquaintance of u_q by performing the RWR procedure, as aforementioned. The proximity scores for all the nodes, i.e., $\vec{p}_{v_q}$, can be derived. The second and the third phases are proposed due to the adoption of active learning.

Candidate Selection corresponds to selecting instances for labeling in the setting of active learning. We are given a budget b of candidates, which are selected from the set of unlabeled users for labeling. Each iteration consumes one unit of budgets, i.e., selecting one candidate for labeling. LAI can be regarded as a kind of *binary* inference since we aim at classifying whether or not a user u_i and the query user u_q are acquainted with each other. Therefore, we have two labeling outcomes: *positive feedback* indicates that the selected candidate u_i is acquainted with u_q, and *negative feedback* refers to that the u_i is unacquainted with u_q. Consequently, at this phase, we need to develop strategies to select effective candidates such that their labeling can significantly benefit the inference performance.

Graph Refinement aims at adjusting the structure of activity graph so that both positive and negative labeling feedback can be incorporated and reflected in the next iteration of inference. Specifically, the adjustment is designed to boost the possibility that the RWR random surfer reaches more acquaintances, and to prevent the random surfer from arriving at more non-acquaintances. Selected candidate users with positive feedback are the references to access more acquaintances, while those with negative feedback are used to avoid further exploring irrelevant neighborhoods in the activity graph. Note that the conventional active learning automatically re-trains the learning model under the supervised setting and thus does not need the phase of graph refinement. Graph refinement is necessary since our approach is based on the RWR graph-based unsupervised approach.

Note that for each query, a general activity graph is constructed, rather than a local graph. In the active learning phase, the general graph will be adjusted according to the feedback. If the feedback is positive, the new link will be created to connect the selected candidate u_c with the query user u_q in the activity graph. On the contrary, if the feedback is negative, the selected candidate u_c can be turned off in the activity graph, i.e., u_c and its incident edges will not be considered in the random walk. In short, the graph construction is done in the level of entire network because we never know which nodes will be selected for active learning and which edges will be affected based on the feedback.

4.3.1 Candidate selection

The active learning-enhanced random walk is designed to select one candidate user for labeling in each of its iterations. We are allowed to select only b candidates in total, where b is supposed to be small (e.g., $b = 10$) due to the constraints of resources and manpower in the real world. We propose to develop the selection strategies based on two general criteria, *informativeness* and *uncertainty*. If a user u_c whose labeling can bring more information about the acquaintances or non-acquaintances of the query user, u_c can be treated as a good candidate. If the labeling of u_c can lower down the uncertainty of more users, u_c can be a good choice as well. We first present three heuristic methods and baselines for candidate selection in the following.

Random We randomly pick one unlabeled user node for feedback. Each user has the same probability of being chosen in each iteration. Such random strategy is served as the baseline.

Proximity Intuitively, a user u_i with the highest RWR proximity score possesses the highest potential to be the acquaintance of the query user. If u_i has the positive feedback, we acquire the concrete knowledge about u_q's acquaintanceship with others. If the labeling feedback of u_i is negative, we can immediately exclude u_i and its neighborhood so that it will not be ranked at top positions in the follow-up iterations. Therefore, in each iteration, we select the unlabeled user with the highest proximity score to be the labeling candidate, formulated by: $u_c^\star = \arg\max_{u_c} \vec{p}_{u_q}(u_c)$.

Uncertainty The RWR proximity scores reflect the potential of acquaintanceship. While users with lower proximity scores are less possible to be acquainted with u_q, those with medium proximity scores are supposed to be uncertain ones. We believe that labeling the most uncertain user can reduce the uncertainty and boost the confidence of inferred acquaintances. Hence, in each iteration, we select the user with the median value of proximity as the candidate. The uncertainty strategy can be represented by $\{u_c^\star | median(\vec{p}_{u_q}) = \vec{p}_{u_q}(u_c^\star)\}$.

In addition to the three heuristic methods, we propose to measure the *Informativeness Reward* of each unlabeled user u_c, denoted by $IR_{u_q}(u_c)$ based on the derived RWR proximity scores with respect to the query user u_q. The basic idea is to estimate how much information can a candidate u_c bring if it gets labeled. The user $u_c^\star$ whose labeling can lead to higher information reward will be selected to perform labeling by external investigation. This can be generally represented by: $u_c^\star = \arg\max_{u_c} IR_{u_q}(u_c)$.

Since our LAI task is to find a proportion of acquaintances for the query user, we measure the information reward by concerning only the top k unlabeled users. Let $\kappa(S)$ be the set of users whose RWR proximity scores (restarting from a given user set S) locate at the top k positions among $U \setminus S$, given by

$$\kappa(S) = \left\{ \arg\max^{(1,\ldots,k)}_{u_i \in U \setminus S} \vec{p}_S(u_i) \right\}, \tag{1}$$

where $\vec{p}_S(u_i)$ is the proximity score of user u_i restarting from a user set S in RWR. Recall that the labeling feedback can be positive or negative, which can lead to increment or decrement of the top k proximity scores. In other words, including a candidate u_c might make some users be removed from $\kappa(S)$ and add more other users into $\kappa(S \cup \{u_c\})$. Besides, if the labeling feedback of u_c is positive, we expect including u_c into the set of inferred acquaintances can lead to more increment of the top k proximity scores, which refers to boost the confidence of inference. On the contrary, if u_c's labeling feedback is negative, adding u_c should lead to more decrement of the top k proximity scores, which lowers down the potential that non-acquaintances are reported. Consequently, we propose to define the Informativeness Reward of positive and negative cases separately, which lead to two different strategies of candidate selection.

Positive Reward Let the labeling feedback of user u_c be positive, i.e., u_c is acquainted with u_q. Based on the concept of *homophily* [25], i.e., those acquainted with each other tend to connect with one another in a social network, the topology structure among the query user u_q and her acquaintances are supposed to be dense. Densely connected nodes can get higher RWR proximity scores than those with a loosely connected structure [41]. Therefore, we expect that including u_c into u_q's acquaintance set F_q can boost the top k RWR proximity scores and raise the potential acquaintances up to top positions. Let S be the current set of inferred acquaintances. We exploit such idea to define the Positive Informativeness Reward $IR^+_{u_q}(u_c)$ as:

$$IR^{+}_{u_q}(u_c) = \sum_{u_i \in \kappa(S) \cap \kappa(S \cup \{u_c\})} \left(\vec{p}_{S \cup \{u_c\}}(u_i) - \vec{p}_S(u_i)\right) + \sum_{u_i \in \kappa(S \cup \{u_c\}) \setminus \kappa(S)} \vec{p}_{S \cup \{u_c\}}(u_i). \tag{2}$$

The first summation calculates the increment of the overlapping users with top k proximity scores, and the second summation estimates the proximity scores of users who are newcomers at the top k positions.

Negative Reward The other case is if the labeling feedback of user u_c be negative. (u_c is unacquainted with u_q.) The homophily theory [25] indicates that the friends of non-acquaintances of a user are *less likely* to be the acquaintances of her. Such idea can be reflected in the structure of the social network, i.e., it is loosely connected between non-acquaintances and u_q. That says, the random surfer may need more steps via few paths to reach non-acquaintances from u_q. Consequently, the top k RWR proximity scores could be lowered down, and some acquaintances may be removed from the top positions, if we consider that u_c is non-acquaintance. Let S be the current set of inferred acquaintances. After removing u_c and its neighboring nodes, we assume $\kappa'(S)$ be the set of users whose RWR proximity scores locate at the top k positions among $U \setminus S$ and $\vec{p'}_S(u_i)$ be the proximity score of user u_i restarting from a user set S in RWR. We define the Negative Informativeness Reward $IR^{-}_{u_q}(u_c)$ as:

$$IR^{-}_{u_q}(u_c) = \sum_{u_i \in \kappa(S) \cap \kappa'(S)} \left(\vec{p}_S(u_i) - \vec{p'}_S(u_i)\right) + \sum_{u_i \in \kappa(S) \setminus \kappa'(S)} \vec{p}_S(u_i). \tag{3}$$

The first summation calculates the decrement of the overlapping users with top k proximity scores, and the second summation estimates the proximity scores of users who are removed from the top k positions.

4.3.2 Graph refinement

When a selected candidate user gets labeled by external investigation, we need to make the most of such precious and informative user so that the performance of inference can be improved as much as possible. Since the fundamental of our approach is RWR, which is a graph-based unsupervised method, we propose to adjust the activity graph. The goal is to refine the graph structure such that the random surfer can be guided to reach acquaintances with higher proximity scores and prevent from arriving at non-acquaintances. We elaborate how to refine the structure of activity graph based on the results of labeling feedback. If the selected candidate u_c gets positive feedback, we refine the graph by *linkage creation*. If u_c gets negative feedback, the graph will be refined by *vertex removal*. We elaborate the details in the following.

Linkage Creation We create a link to connect the query user u_q and the candidate u_c if u_c gets positive labeling feedback. The new linkage leads the RWR random surfer to arrive u_c directly. Thus, the proximity scores of u_c's neighbors can be significantly increased. Then based on the concept of homophily [25], in the follow-up iterations, more acquaintances of u_q can be reached with their proximity scores getting boosted.

Vertex Removal If u_c gets negative labeling feedback, we remove u_c and all of its incident edges from the activity graph. Such removal will prohibit the RWR random surfer from reaching the neighbors of u_c. The homophily effect "people with different interests tend to be unacquainted with each other" [25] is implemented by removing edges connecting u_c and its neighboring nodes of labels and activities. As a result, the proximity scores of non-acquaintances can be decreased in the follow-up iterations.

Algorithm 2 Active Learning-enhanced Random Walk

```
Require: $\mathcal{H} = (\mathcal{V}, \mathcal{E})$: the activity graph; $u_q$: the query user node; $\alpha$: the restart probability; $N$: the number
    of acquaintances to be returned; $b$: number of budgets for labeling
Ensure: $F_q$: the acquaintance set of user node $v_q$
1: $F_q \leftarrow \{u_q\}$;
2: $\vec{p}_{F_q} \leftarrow RWR_{F_q}(\mathcal{H})$;
3: for $iter = 1$ to $b$ do
4:     $u_c^\star \leftarrow \arg\max_{u_i \in U \setminus F_q} SelectionCrition_{F_q}(u_i)$; // select labeling candidate
5:     $l_{u_c} \leftarrow GetLabel(u_q, u_c^\star)$;
6:     if $l_{u_c}$ = "positive" then
7:         $\mathcal{H}.AddEdge((u_c, u_q))$; // refinement
8:     if $l_{u_c}$ = "negative" then
9:         $\mathcal{H}.RemoveVertex(u_c)$; // refinement
10:    $\vec{p}_{F_q} \leftarrow RWR_{F_q}(\mathcal{H})$;
11:    $u_a^\star \leftarrow \arg\max_{u_i \in U \setminus F_q \wedge u_i \in \mathcal{U}} \vec{p}_{F_q}(u_i)$;
12:    $F_q \leftarrow F_q \cup \{u_a^\star\}$;
13: return $F_q \setminus \{u_q\}$
```

The complete procedure of active learning-enhanced random walk is shown in Algorithm 2. We first compute the RWR proximity scores with respect to the query user u_q for all the nodes in the activity graph $\mathcal{H}$ (lines 1–2), where $RWR_{F_q}(\mathcal{H})$ returns the proximity scores of nodes by considering the node set F_q as the RWR restarting nodes in the activity graph $\mathcal{H}$. Then we iteratively perform the active learning up to b times, in which one inferred acquaintance is reported in each iteration (lines 3–12). Each iteration consists of three phases. The first phase is candidate selection (line 4), in which $SelectionCriterion_{F_q}(u_i)$ can be implemented by *Random*, *Proximity*, *Uncertainty*, *Positive Reward* and *Negative Reward*. After deriving the label of $u_c^\star$ (line 5), the second phase is to perform graph refinement based on its label values ("positive" or "negative") (lines 6–9). The last phase re-calculates the RWR proximity scores, and consider the user $u_a^\star$ possessing the highest proximity score as a new inferred acquaintance (lines 10–12).

4.4 Time complexity analysis of ARW

In this section, we analyze the time complexity of ARW. ARW consists of two major components: activity graph construction and active learning-enhanced random walk. For the first component, we take advantage of user profiles and users' geographical activity to construct the activity graph. Assume that there are vertex set $\mathcal{V}$ and $\mathcal{E}$ edge set to construct the activity graph. The time complexity of activity graph construction is $O(|\mathcal{V}| + |\mathcal{E}|)$. For the second component, the complexity of random walk depends on the number of edges in the graph. In the extreme case, nodes are fully connected to each other. The maximum number of edges is $C_2^{|\mathcal{V}|}$. Hence, the complexity of random walk is $O(|\mathcal{V}|^2)$. After executing the random walk with restart, we sort "user" vertices by their proximity scores. Let the user node set be U, the sorting complexity is $O(|U| \log U)$, where U is much less than $\mathcal{V}$. In the process of active learning, we assume that the budget for labeling is b. In each round of active learning, we need to execute the random walk after graph refinement, leading to the time cost $O(b \cdot |\mathcal{V}|^2)$ of active learning-enhanced random walk. The complexity of graph construction is much less than the complexity of active learning-enhanced random walk. Therefore, the overall complexity is $O(b \cdot |\mathcal{V}|^2)$.

5 Evaluation

Our algorithms are implemented in Python. All of the experiments are conducted on a Linux machine with a 3.40 GHz Intel i7 Core and 16-GB RAM.

5.1 Datasets

Instagram Dataset Instagram is a photograph-sharing social network with a fast-growing user number. Currently, it has 400M monthly active users and generates 75M photographs every day. Similar to other social network services such as Facebook and Twitter, Instagram allows users to share their locations when publishing photographs. Unlike Twitter where only a small amount of tweets are geo-tagged, a past study [24] has shown that Instagram users are much more willing to share their locations (31 times more than Twitter users), which makes Instagram data suitable for our experiments.

We exploit Instagram's public API to collect the geo-tagged posts, which are treated as the geographical activities of users, from three major cities worldwide including San Francisco, London and New York. We first resort to Foursquare, a popular location-based social network with resourceful information about locations and their categories, to extract all locations with their IDs in each city. Since Instagram's location service is linked with Foursquare[1], we then use the obtained Foursquare's location IDs to extract the corresponding Instagram's location IDs. We query Instagram's API with such location IDs to derive all geo-tagged posts in three targeted cities. In addition, we collect all users who share geo-tagged posts in Instagram. The *followships* between users are also crawled. Two users are considered as being acquainted with one another if they follow each other. Besides, we obtain user demographics as the attributes and tags by resorting to Face++,[2] state-of-the-art deep learning-based facial recognition service, to analyze each user's profile. The output of Face++ includes a user's age, gender and race (White, Asian and Black). It is worth noticing that Face++ has been widely used to extract demographics from social media photographs [32,36]. The statistics of the obtained Instagram data is shown in Table 2.

Meetup Dataset The Meetup dataset compiled by in an existing study [22] is used for our experiments. Meetup is a social networking portal that facilitates *offline* meeting events, which are considered as the geographical activities of users, in various localities around the world. In addition, Meetup users can participate in online groups, in which each group is associated with a set of tags depicting its semantics. Over four millions of users and eight

[1] The connection is aborted on April 20, 2016 (https://www.instagram.com/developer/changelog/).

[2] http://www.faceplusplus.com/.

Table 2 Statistics of Instagram and Meetup datasets

	City	#$Users$	#Nodes	#Edges	#Geo-activities	#Labels	#Locations
Instagram	New York	148,486	4,764,312	10,863,764	6,445,374	371	44,951
	London	63,698	1,780,779	3,893,716	2,268,959	335	21,327
	San Francisco	35,102	1,291,884	2,807,799	1,600,287	334	14,517
Meetup	New York	67,240	198,483	1,500,334	727,858	58,130	10,339
	San Francisco	30,190	89,114	637,516	219,364	41,042	2,996
	London	21,369	69,446	475,179	209,829	26,171	3,694
	Cambridge	11,613	42,111	253,711	69,879	24,478	725
	Berkeley	7380	33,419	174,311	24,597	22,676	499
	Miami	4796	25,456	126,113	16,242	17,736	598
	Paris	4386	23,436	131,992	45,236	14,955	896
	Sydney	3563	16,158	79,108	15,422	10,991	239
	Melbourne	2800	13,512	64,131	12,538	9037	354
	New Orleans	1416	11,057	44,237	9101	7752	457
	Roma	1278	7328	27,411	9719	4210	404
	Oxford	615	5419	19,752	5728	3659	173

millions of user-group pairs are collected in this dataset. Such data were crawled during October 2011 and January 2012.

There are no explicit friendships between users in Meetup [12,19,22]. We follow existing studies [12,19,22] to construct the social links for the ground truth. They follow three principles to construct the friendships between users: (a) users are connected if they join the same social groups, (b) users involving in smaller groups tend to closely connected than those in larger groups, and (c) friends tend to be close to each other in geography and around 70% of Meetup online friends live within 10 miles. Here we use the same method and settings to define the friendship score $f(u_i, u_j)$ of users u_i and u_j, given by:

$$f(u_i, u_j) = \sum_{\forall g_k, u_i \in g_k \wedge u_j \in g_k \wedge dist(u_i, u_j) \leq 10\ miles} \frac{1}{|g_k|}, \tag{4}$$

where g_k is an online group in Meetup, and $dist(u_i, u_j)$ returns the geographical distance between u_i and u_j's home locations. User pairs whose $f(u_i, u_j)$ scores locate at the highest τ percentage are considered as friends. We empirically set $\tau = 10\%$ in this work. The statistics of Meetup data is also shown in Table 2.

5.2 Competitors and evaluation metric

We compare the proposed *ARW* with the following eight competitive methods.

- *RWR* [41]: This is the random walk with restart approach without active learning.
- *LD* [31]: Location diversity (LD) is a referring value for social strength between users. It captures the intuition that two users who meet at more different places tend to possess a higher probability of friendship than those meeting at fewer places.
- *WF* [31]: Weighted frequency (WF) is based on the assumption that users who meet at popular places (such as train station and city center) are less likely to be friends since these meeting events can be coincidences. The popularity of a location in WF is defined by location entropy [10].
- *EBM* [31]: The EBM model combines LD and WF together with a linear regression model. The output value is considered as the predicted social strength between users and then treated as a measure for friendship prediction.
- *EBM_AL*: While the original EBM cannot consider active learning to adjust the acquaintance score, to have a fair comparison, we extend the EBM model to incorporate with the concept of active learning, denoted by EBM_AL. We select the user pair with the highest EBM score for labeling in active learning. Then, we adjust location entropy according to the labeling feedback. The location entropy is the value between 0 and 1. Let $L_{i,j}$ be the co-occurrence location list between query user i and user j selected by active learning. The location entropy of location $l \in L_{i,j}$ will be adjusted according to the labeling feedback as follows:

$$H_l = \begin{cases} {H_l}^{\frac{1}{s}}, & \text{if user } i \text{ is acquainted with } j \\ {H_l}^{s}, & \text{if user } i \text{ is not acquainted with } j \end{cases} \tag{5}$$

  The location entropy values of locations that acquaintances co-occur will be raised. On the contrary, the location entropy of location co-occurred by non-acquaintances will be lowered down. WF score will also be adjusted according to the update of location entropy H_l. The updated range of location entropy can be changed by adjusting parameter s. We set $s = 2$ in all experiments by default.

- *PGT* [44]: Similar to EBM, the PGT model, which is the state-of-the-art method, also an unsupervised method to estimate the acquaintance based on two users' meeting events. PGT models how two users correlate with each other from personal, global and temporal perspectives, and outputs a score of friendship inference. Higher scores indicate higher potential that two users are acquainted with one another.
- *PGT_AL*: We also extend PGT to have an active learning-based PGT, denoted by PGT_AL. Similar to EBM_AL, the user pair with the highest PGT score will be selected for labeling. After the selected node was labeled, the personal and global scores in PGT will be updated to reflect the spatiosocial interactions between users. Let $L_{i,j}$ be the co-occurrence location list between query user i and user j selected by active learning. We adjust the probability for users i and j to visit location $l_k \in L_{i,j}$. The adjustment for personal mobility is as follows:

$$\rho(i, l_k) = \begin{cases} \rho(i, l_k)^{\frac{1}{s}}, & \text{if user } i \text{ is acquainted with } j \\ \rho(i, l_k)^{s}, & \text{if user } i \text{ is not acquainted with } j \end{cases} \tag{6}$$

  In addition, we also adjust the global mobility. The global mobility of PGT is location entropy. Hence, we adjust the location entropy of location $l_k \in L_{i,j}$ according to the labeling feedback as follows:

$$g(l_k) = \begin{cases} g(l_k)^{\frac{1}{s}}, & \text{if user } i \text{ is acquainted with } j \\ g(l_k)^{s}, & \text{if user } i \text{ is not acquainted with } j \end{cases} \tag{7}$$

  The concept of adjustment is like EBM_AL. The personal mobility and global mobility will be raised when the labeling feedback is positive. On the contrary, the personal mobility and global mobility will be lower down when the labeling feedback is negative. The updated range of location entropy can be changed by adjusting parameter s. We set $s = 2$ by default in all experiments.
- *SRW* [1]: Supervised random walk (SRW) is one of the state-of-the-art random walk-based methods for link prediction. SRW combines the information from the network structure with node-level and edge-level attributes. These attributes are used to guide the random walker starting from a source node to surfer in the graph. Nodes with higher probabilities to be reached are considered as the potential friends for the source node. To implement SRW in our experiment, for each pair of users i and j, we consider five essential features for the supervised learning part in SRW: (1) number of their meeting events, (2) LD score, (3) WF score, (4) EBM score and (5) PGT score. Since SRW is a supervised method, it needs labeled data for training. Therefore, we consider all of the labeling outcomes (by active learning) as the training set. It can be noticed that the training size is extremely small.

The evaluation metric used in our experiments is **Precision@N**, which estimates the percentage of relevant instances among the top-N ones. Here we consider N is a small number since investigators might concern more about those early reported. We empirically set $N = 10$ at most in this work. Let $F_i(N)$ be the set of the top-N inferred acquaintances for user i by a certain method, and $\hat{F}_i$ be the set of user i's ground-truth friends. The score of Precision@N is defined as $\frac{|F_i(N) \cap \hat{F}_i|}{N}$.

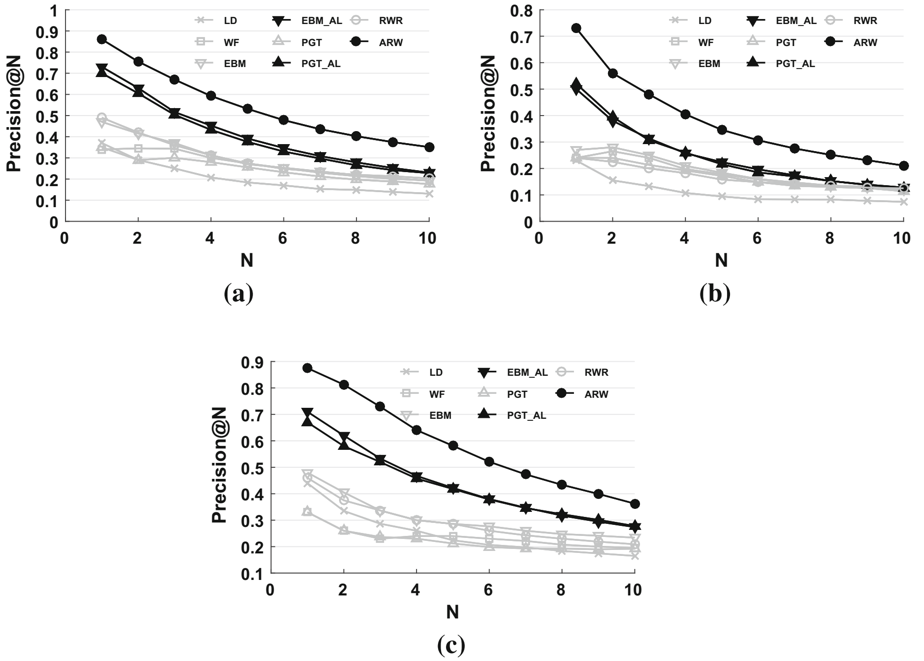


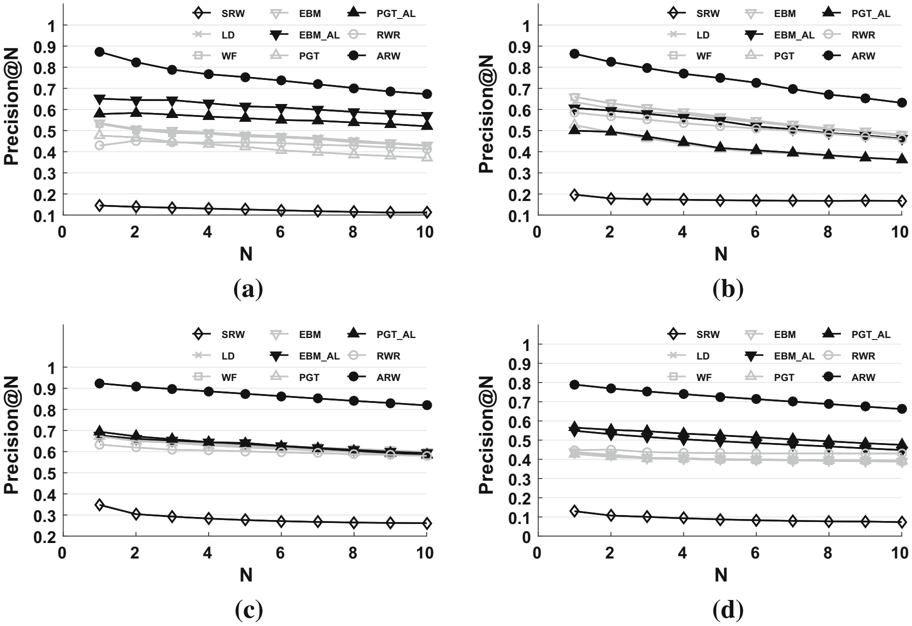
Fig. 5 Performance comparison of different methods in Instagram data. **a** San Francisco, **b** London, **c** New York

5.3 Experimental results

5.3.1 Comparison of inference methods

We first present the performance of our *ARW* framework, comparing to the eight competitors. The experiments are conducted in three cities on Instagram and twelve cities on Meetup. The number of reported acquaintances N is varied from 1 to 10. Note that the candidate selection strategy of our ARW in this experiment is Positive Informativeness Reward (IR+) by default. We will discuss the effectiveness of different candidate selection strategies later.

Figure 5 presents the results in Instagram data. Figures 6 and 7 exhibit the results in small-scale cities and large-scale cities, respectively, in Meetup data. Such results deliver the following findings. First, it can be apparently observed that *ARW* significantly outperforms all competitors. Though the precision score slowly decreases as N increases, ARW always maintains a significant advantage over the competitors. Second, the conventional inference approaches, i.e., PGT, EBM, LD, WF and RWR, result in worse performance, especially in cities with large amounts of data (as shown in Fig. 7 and the sizes of cities are listed in Table 2). It is reasonable since big cities possess much denser area in terms of people and locations, which causes conventional social measures cannot well separate the acquaintances from non-acquaintances based on their meeting events and geographical check-in activities. When facing such challenge, by imposing the active learning strategy with only few labeling actions, ARW can surprisingly lead to outstanding performance. This finding informs us that the intervention by manual investigators can effectively bring social clues to guide the ARW acquaintance search in the constructed heterogeneous graph.



(a)

(b)

(c)

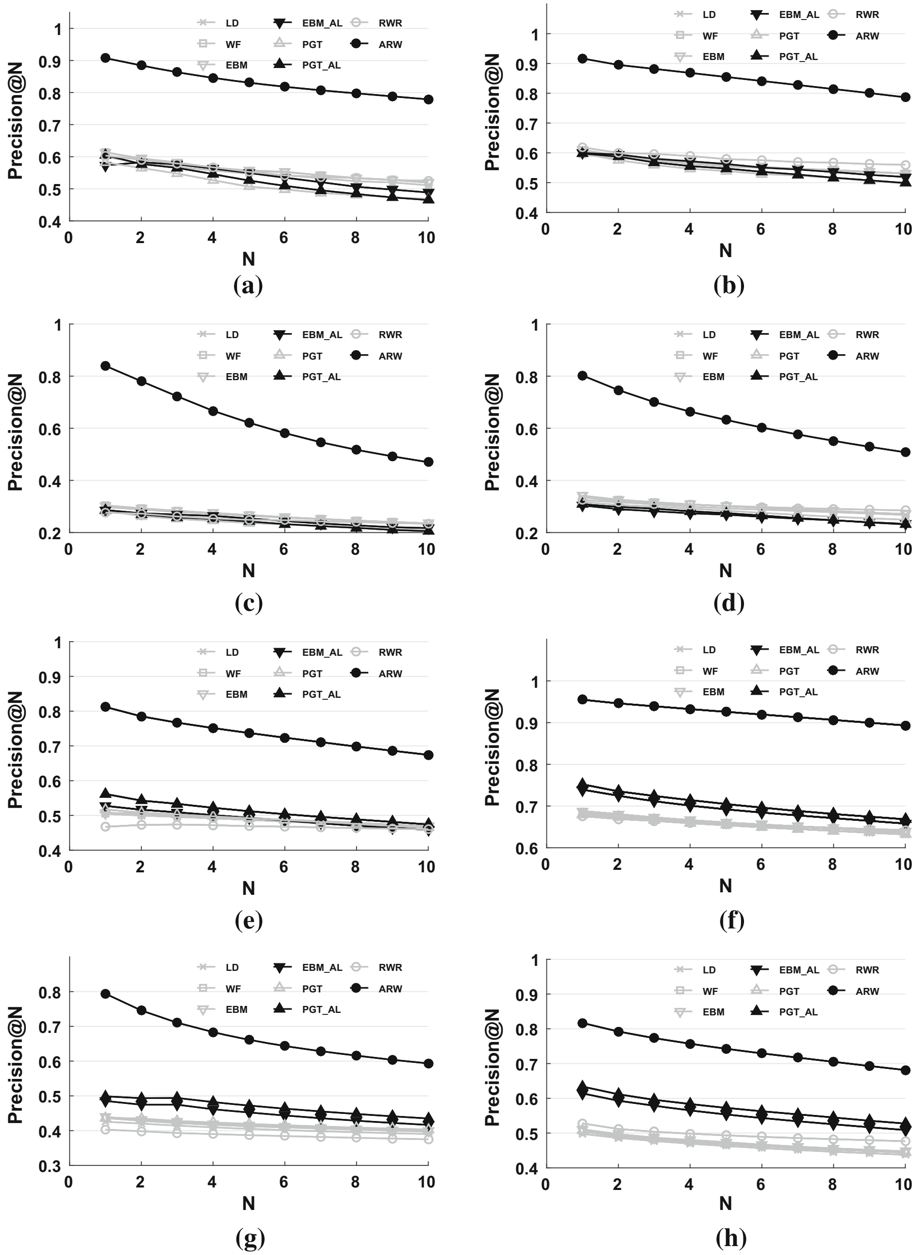
(d)

Fig. 6 Performance comparison of different methods in Meetup data. **a** Oxford, **b** Roma, **c** Melbourne, **d** Paris

Third, to further understand whether active learning can work well on state-of-the-art unsupervised acquaintance estimation methods (i.e., EBM and PGT), our evaluation also reports the results of active learning-based EBM and PGT (i.e., EBM_AL and PGT_AL). It can be found that the performance of EBM_AL and PGT_AL is truly better than EBM and PGT, but the gaps are minor, comparing with the gaps between ARW and RWR. Such result proves the effectiveness of the proposed techniques of candidate selection and graph refinement in Sect. 4.3.

Fourth, we can see the performance of SRW is the worst one, even it produces the inferred acquaintances in a supervised manner. The reason is supposed to that SRW needs sufficient training data so that the random walks can be well guided in the graph. As active learning essentially allows the only very limited number of labeling feedback, the worse training destroys the performance. Moreover, we present the time efficiency (in s) of LD, WF, EBM, PGT, EBM_AL, PGT_AL, SRW, RWR and our ARW. The results are shown in Fig. 8. Note that the method of active learning version executes 20 rounds of active learning. We can find that the running time of all methods except for SRW is less than 5 s in Meetup data. In Instagram data, the running time of ARW and SRW is larger than other methods. Due to the active learning process, ARW needs more time to execute the random walk mechanism. SRW needs much more time to train and execute random walks. As the city scale increases, however, the running time of SRW grows exponentially. With worst performance and inefficient running time, SRW can be considered as the most ineffective method. Therefore, we do not show the results of SRW in Figs. 5 and 7, and the following experiments.

In addition, we not only hope ARW can lead to the best performance, but also the execution time per query must be acceptable. Therefore, we further report the execution time per query (in s), along with Precision@10. The results are as shown in Fig. 9. Though the execution time



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 7 Performance comparison of different methods in Meetup data. **a** New Orleans, **b** Sydney, **c** Miami, **d** Berkeley, **e** Cambridge, **f** London, **g** San Francisco, **h** New York

of RWR is about 13 times faster than that of ARW, the time cost of ARW is still acceptable. As depicted in Fig. 9b, all execution time of ARW in Meetup data is less than 10 s. However, the execution time depends on the graph size. Larger graphs take much time. We can see that the time of ARW in New York is around 17 min in Instagram data, as shown in Fig. 9a.

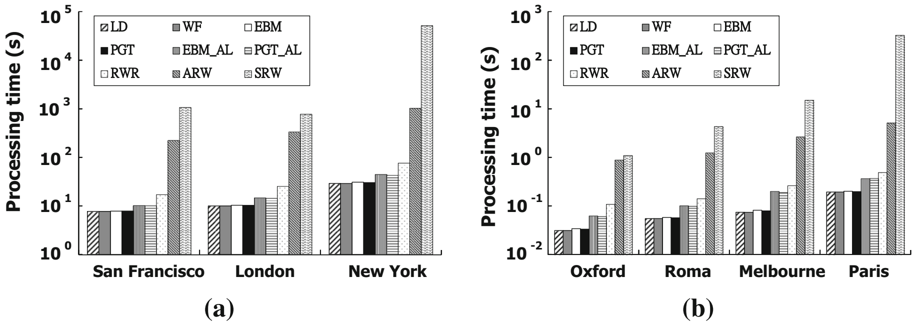


Fig. 8 The running time per query for all of the competitors. **a** Instagram, **b** Meetup

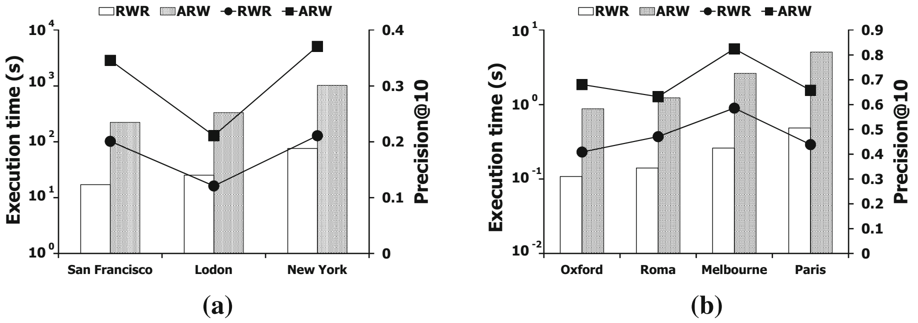


Fig. 9 The execution time per query versus Precision@10. **a** Instagram, **b** Meetup

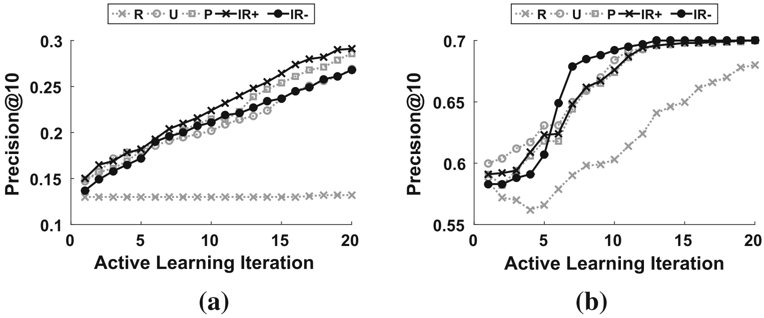


Fig. 10 Performance of candidate selection strategies. **a** Instagram, **b** Meetup

5.3.2 Comparison of candidate selection strategies

In this section, we compare the performance of different candidate selection strategies using our ARW method, i.e., Random (R), Uncertainty (U), Proximity (P), Positive Informativeness Reward (IR+) and Negative Informativeness Reward (IR−). We vary the number of iterations in active learning (i.e., the labeling budget) from 1 to 20 and report the score of the Precision@10, and the results are shown in Fig. 10. Note that the resulting precision scores are obtained by averaging the scores over different cities in each dataset.

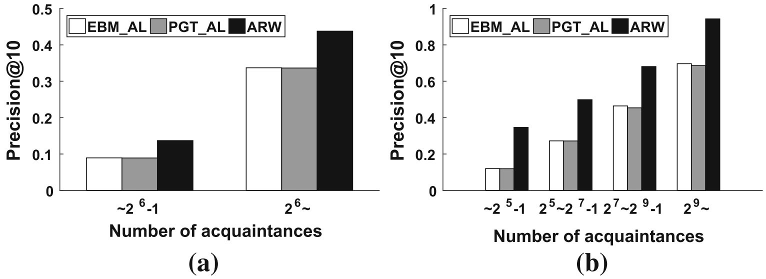


Fig. 11 Effect on different levels of acquaintance numbers by fixing Precision@N and $N = 10$. **a** Instagram, **b** Meetup

It can be apparently found that the proposed IR+ and IR− strategies lead to the better performance in both datasets as the active learning iteration increases. Detailed speaking, IR+ works better on Instagram data, while IR− outperforms others in Meetup data. We think such performance difference between IR+ and IR− results from the density (i.e., $\frac{\#Edges}{\#Nodes}$) of the constructed heterogeneous graph. According to Table 2, the graph density values of Instagram cities are much higher than those of Meetup cities. For graphs with lower density (i.e., higher sparsity), ARW needs to be advised with more positive responses so that the random surfer can have more knowledge about the positions of acquaintances in graphs among the huge amount of non-acquaintances. If graphs are relatively dense, on the contrary, negative responses are more useful since they can avoid ARW walking toward non-acquaintances, while acquaintances have higher proximity from the query node in essence.

5.3.3 *Effects on the number of acquaintances*

We evaluate how does the number of query user's acquaintances affects the performance of active learning-based methods, i.e., *ARW*, PGT_AL and EBM_AL. By quantizing the number of acquaintances of the query user into several levels, i.e., ($\leq 2^6 - 1$) and ($\geq 2^6$) in Instagram data, while ($\leq 2^5 - 1$), ($2^5 \sim 2^7 - 1$), ($2^7 \sim 2^9 - 1$), and ($\geq 2^9$), we report the experimental results in Fig. 11. The Precision@N ($N = 10$) for different levels of acquaintance numbers is shown. For example, "$2^9 \sim$" in x-axis means all users whose numbers of acquaintances are greater than $2^9 - 1$. It is different from Figs. 5, 6 and 7. Their Precision@N scores are calculated by all users in that city with different N values. Note that the resulting precision scores are obtained by averaging the scores over different cities in each dataset.

Three findings can be learned from the results. First, the performance gets better as the number of acquaintances increases. It is natural that users possessing more friends tend to result in higher accuracy since more friends bring more evidences of acquaintance behaviors for the query user. Second, no matter how the number of acquaintances is varied, our *ARW* can outperform PGT_AL and EBM_AL. The advantage gap of ARW over the other methods gets significant for the least acquaintance level. Such results demonstrate not only the stability of *ARW*, but also its strength in discovering acquaintances even though the query user has few friends. Being able to find acquaintances with less social clues would be very useful in identifying the companions of a known terrorist in Homeland Security. Last, by comparing the precision scores in Fig. 11a, b, we can find the performance on Instagram is worse than that on Meetup. By taking ARW as an example, the former ranges from 0.1 to 0.45, while the latter ranges from 0.35 to 0.95. We think such differences result from the fact that Meetup has a stronger setting of acquaintance, while the acquaintances in Instagram are relatively

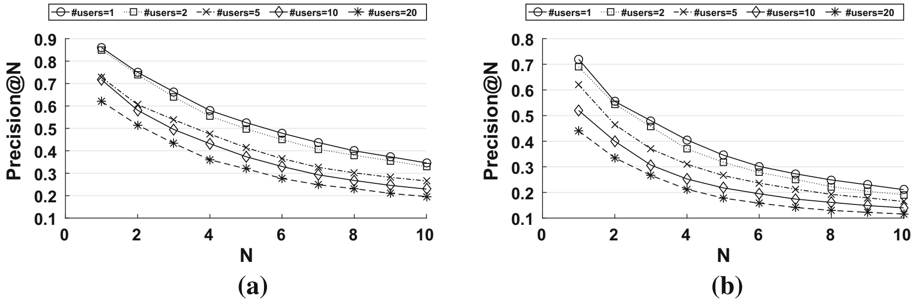


Fig. 12 Effect of #user queried by active learning in Instagram data. **a** San Francisco, **b** London

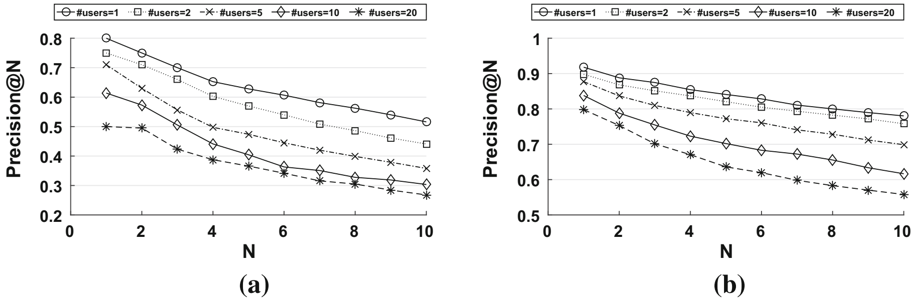


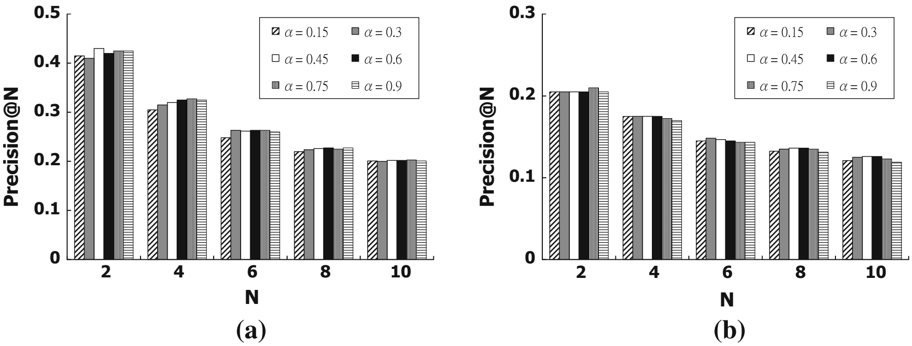
Fig. 13 Effect of #user queried by active learning in Meetup data. **a** Berkeley, **b** New Orleans

weak. In Instagram, two users with mutual followship are regarded as an acquaintance. But in Meetup, two users are acquainted with each other only if they co-participate in a sufficient number of online groups. Co-participation provides more evidences on acquaintances than mutual followship since co-participation implies two users share common traits or attributes in online groups, while mutual followship deliver less potential in ensuring common interests.

5.3.4 Effects on different parameter settings

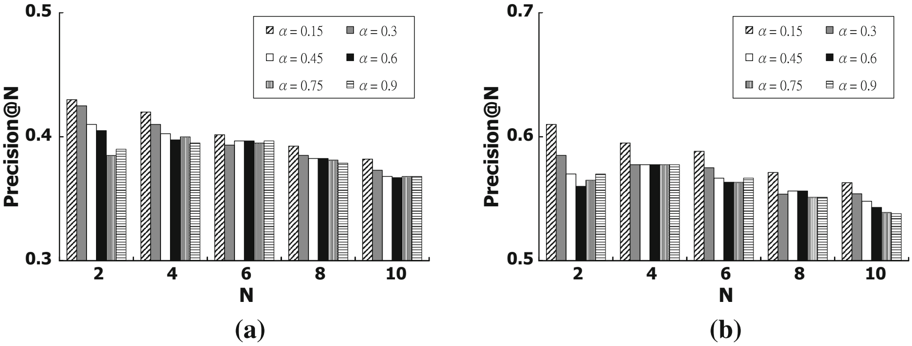
Effects on the Number of Users per Active Learning Round We also evaluate how does the number of peoples asked by active learning affect the performance of ARW. The number of total asked peoples is 20. Then, we set 1, 2, 5, 10 and 20 peoples to ask per each active learning round. That is to say, active learning will execute (#total asked peoples / #peoples asked per active learning round) times. The experimental results are shown in Figs. 12 and 13. The experimental results exhibit that more times of active learning lead to the better performance. ARW that inquires the label of one user per active learning round can generate the best precision scores in different cities of both datasets. Such results reflect more fine-grained active learning process can better learn the labels of acquaintances.

Effects of Restart Probability α We compare the performance of different values of restart probability α under our ARW method. We set $\alpha = 0.15, 0.3, 0.45, 0.6, 0.75$ and 0.9. The experimental results are shown as in Figs. 14 and 15. We can find that there is no significant difference between different α values. Though the proximity will be different when the restart probability is different, the proximity-based ranking results of nodes will not be affected.



(a) (b)

Fig. 14 Effect of α in Instagram data. **a** San Francisco, **b** London



(a) (b)

Fig. 15 Effect of α in Meetup data. **a** Berkeley, **b** New Orleans

5.3.5 Analysis of labeling feedback

In general, one may think obtaining positive feedback can better benefit the active learning so that the accuracy of a prediction task would get higher. To understand whether it is true in acquaintance inference, we conduct the analysis of labeling feedback. We aim to report the percentages of positive and negative labeling feedback acquired from our *ARW*. We choose to present the results of Oxford, Paris and Miami in Meetup Data, and New York, San Francisco and London in Instagram, while other cities exhibit similar trends.

Figure 16 shows the percentages of positive and negative labeling feedback. We can find that the percentages of negative feedback are significantly less than positive ones, while the results in Meetup tend to draw an opposite consequence (except for in Miami). Such results correspond to the empirical studies of the effects of the acquaintance number and the candidate selection strategies in previous two subsections. Specifically, first, the Instagram data exhibit lower graph density. A large amount of non-acquaintances result in a higher probability of labeling negative users get higher, comparing to Meetup data that possess a relatively high graph density. Second, the formation of social acquaintance is weaker for Instagram due to mutual followship than for Meetup resulting from co-participation. Weak acquaintance would introduce more noises for the random walk mechanism and thus guide ARW to arrive at nodes with negative feedback.

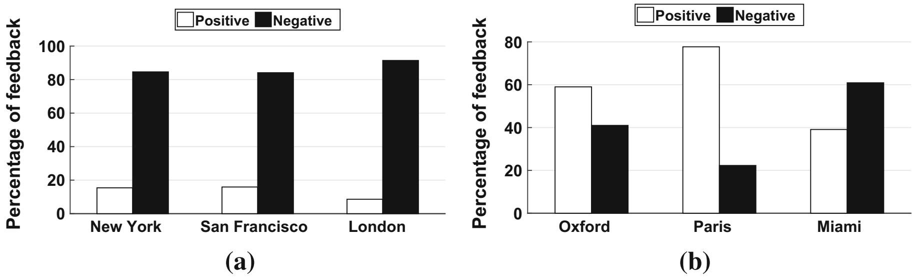


Fig. 16 Percentage of positive and negative feedback. **a** Instagram, **b** Meetup

6 Conclusions and discussion

This paper proposed and solved the location-aware acquaintance inference (LAI) problem. We develop a semi-supervised inference framework ARW to deliver three contributions. First, it seamlessly integrates random walk with restart with active learning. Second, while ARW is based on the heterogeneous information network, it can be applied to unsupervised node ranking tasks by constructing the corresponding graphs. Finally, to implement the concept of active learning, we devise five strategies to select the candidates to be labeled and two refinement mechanisms that incorporate positive and negative labeling feedback. Empirical studies conducted on Instagram and Meetup datasets show that ARW can significantly outperform state-of-the-art methods. The results also reveal that only a small number of budgets for labeling can lead to a satisfying boost of performance.

Future extensions of ARW could involve incremental updating of RWR proximity scores when some nodes and edges are removed from and added into a large-scale graph. In addition, ARW can also be extended to the discovery of top-*k* nodes, instead of computing the scores of all nodes, for the purpose of boosting the time efficiency. Some efficient algorithms can be exploited to reduce the complexity to $O(|\mathcal{V}| + |\mathcal{E}|)$ without sacrificing the accuracy, e.g., K-dash [13]. Also, it would be interesting to select users whose labeling can directly maximize the proximity scores of top users. Moreover, the performance of ARW can be boosted if we can effectively learn the edge weights in the activity graph.

Acknowledgements This work was supported in part by Ministry of Science and Technology, ROC, under Contract 105-2221-E-006-140-MY2, 107-2636-E-006-002 and 106-3114-E-006-002. In addition, we also thank the support from Academia Sinica under the Grant AS-107-TP-A05. Also, we would like to thank anonymous reviewers and the editor for their very useful comments and suggestions.

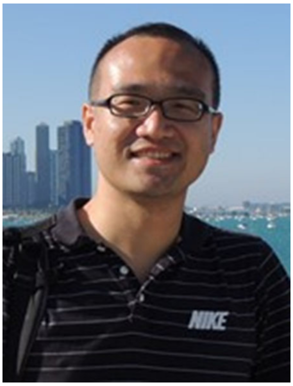
References

1. Backstrom L, Leskovec J (2011) Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the forth international conference on web search and web data mining, pp 635–644
2. Bagci H, Karagoz P (2016) Context-aware friend recommendation for location based social networks using random walk. In: Proceedings of the 25th international conference on world wide web, pp 531–536
3. Bagci H, Karagoz P (2016) Context-aware location recommendation by using a random walk-based approach. Knowl Inf Syst 47(2):241–260
4. Barbieri N, Bonchi F, Manco G (2014) Who to follow and why: link prediction with explanations. In: The 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1266–1275

5. Chen K, Han J, Li Y (2014) HALLP: a hybrid active learning approach to link prediction task. J Comput 9(3):551–556
6. Cheng R, Pang J, Zhang Y (2015) Inferring friendship from check-in data of location-based social networks. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining, pp 1284–1291
7. Cho E, Myers SA, Leskovec J (2011) 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, pp 1082–1090
8. Choudhury MD, Mason WA, Hofman JM, Watts DJ (2010) Inferring relevant social networks from interpersonal communication. In: Proceedings of the 19th international conference on world wide web, pp 301–310
9. Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J (2010) Inferring social ties from geographic coincidences. Proc Natl Acad Sci 107(52):22436–22441
10. Cranshaw J, Toch E, Hone J, Kittur A, Sadeh N (2010) Bridging the gap between physical location and online social networks. In: Proceedings of 12th ACM international conference on ubiquitous computing, pp 119–128
11. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci 106(36):15,274–15,278
12. Feng K, Cong G, Bhowmick SS, Ma S (2014) In search of influential event organizers in online social networks. In: International conference on management of data, pp 63–74
13. Fujiwara Y, Nakatsuji M, Onizuka M, Kitsuregawa M (2012) Fast and exact top-k search for random walk with restart. PVLDB 5(5):442–453
14. Hsieh H, Yan R, Li C (2015) Where you go reveals who you know: analyzing social ties from millions of footprints. In: Proceedings of the 24th ACM international conference on information and knowledge management, pp 1839–1842
15. Hu B, Ester M (2014) Social topic modeling for point-of-interest recommendation in location-based social networks. In: 2014 IEEE international conference on data mining, pp 845–850
16. Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, pp 538–543
17. Li H, Ge Y, Hong R, Zhu H (2016) Point-of-interest recommendations: Learning potential check-ins from friends. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 975–984
18. Li J, Xia F, Wang W, Chen Z, Asabere NY, Jiang H (2014) Acrec: a co-authorship based random walk model for academic collaboration recommendation. In: 23rd international world wide web conference, pp 1209–1214
19. Li K, Lu W, Bhagat S, Lakshmanan LVS, Yu C (2014) On social event organization. In: The 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1206–1215
20. Liben-Nowell D, Kleinberg JM (2003) The link prediction problem for social networks. In: Proceedings of the ACM CIKM international conference on information and knowledge management, pp 556–559
21. Liu S, Wang B, Xu M (2017) Event recommendation based on graph random walking and history preference reranking. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, pp 861–864
22. Liu X, He Q, Tian Y, Lee W, McPherson J, Han J (2012) Event-based social networks: linking the online and offline social worlds. In: The 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1032–1040
23. Lu L, Zhou T (2011) Link prediction in complex networks: a survey. Physica A 390(6):1150–1170
24. Manikonda L, Hu Y, Kambhampati S (2014) Analyzing user activities, demographics, social network structure and user-generated content on Instagram. arXiv:1410.8099
25. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444
26. Medina RM, Hepner GF (2008) Geospatial analysis of dynamic terrorist networks. Springer, Dordrecht, pp 151–167
27. Medina RM, Hepner GF (2013) The geography of international terrorism: an introduction to spaces and places of violent non-state groups. CRC Press, Boca Raton
28. Mengshoel OJ, Desai R, Chen A, Tran B (2013) Will we connect again? machine learning for link prediction in mobile social networks. In: Proceedings of the 11th international workshop on mining and learning with graphs
29. Myers S, Leskovec J (2010) On the convexity of latent social network inference. In: Advances in neural information processing systems, pp 1741–1749

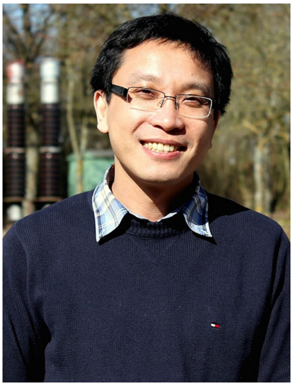
30. Pan J, Yang H, Faloutsos C, Duygulu P (2004) Automatic multimedia cross-modal correlation discovery. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp 653–658
31. Pham H, Shahabi C, Liu Y (2013) EBM: an entropy-based model to infer social strength from spatiotemporal data. In: Proceedings of the ACM SIGMOD international conference on management of data, pp 265–276
32. Redi M, Quercia D, Graham LT, Gosling SD (2015) Like partying? your face says it all. Predicting the ambiance of places with profile pictures. In: Proceedings of the ninth international conference on web and social media, pp 347–356
33. Sadilek A, Kautz HA, Bigham JP (2012) Finding your friends and following them to where you are. In: Proceedings of the fifth international conference on web search and web data mining, pp 723–732
34. Scellato S, Noulas A, Mascolo C (2011) 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, pp 1046–1054
35. Settles B (2010) Active learning literature survey. University of Wisconsin, Madison 52(55–66):11
36. Souza F, de Las Casas D, Flores V, Youn S, Cha M, Quercia D, Almeida V (2015) Dawn of the selfie era: the whos, wheres, and hows of selfies on Instagram. In: Proceedings of the 2015 ACM on conference on online social networks, pp 221–231
37. Sparrow MK (1991) The application of network analysis to criminal intelligence: an assessment of the prospects. Soc Netw 13(3):251–274
38. Sun J, Qu H, Chakrabarti D, Faloutsos C (2005) Neighborhood formation and anomaly detection in bipartite graphs. In: Proceedings of the 5th IEEE international conference on data mining, pp 418–425
39. Sun Y, Han J, Yan X, Yu PS, Wu T (2011) Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In: PVLDB
40. Tong H, Faloutsos C (2006) Center-piece subgraphs: problem definition and fast solutions. In: Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining, pp 404–413
41. Tong H, Faloutsos C, Pan J (2006) Fast random walk with restart and its applications. In: Proceedings of the 6th IEEE international conference on data mining, pp 613–622
42. Wang C, Ye M, Lee W (2012) From face-to-face gathering to social structure. In: 21st ACM international conference on information and knowledge management, pp 465–474
43. Wang D, Pedreschi D, Song C, Giannotti F, Barabasi AL (2011) Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1100–1108
44. Wang H, Li Z, Lee W (2014) PGT: Measuring mobility relationship using personal, global and temporal factors. In: Proceedings of the 14th IEEE international conference on data mining, pp 570–579
45. Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge
46. Xiao X, Zheng Y, Luo Q, Xie X (2014) Inferring social ties between users with human location history. J Ambient Intell Humaniz Comput 5(1):3–19
47. Yin P, He Q, Liu X, Lee WC (2014) It takes two to tango: Exploring social tie development with both online and offline interactions. In: Proceedings of the 2014 SIAM international conference on data mining, pp 334–342
48. Ying JJ, Kuo W, Tseng VS, Lu EH (2014) Mining user check-in behavior with a random walk for urban point-of-interest recommendations. ACM TIST 5(3):40:1–40:26
49. Zhang JD, Chow CY (2015) Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp 443–452
50. Zhuang H, Tang J, Tang W, Lou T, Chin A, Wang X (2012) Actively learning to infer social ties. Data Min Knowl Discov 25(2):270–297

Bo-Heng Chen received the B.S. degree from Computer Science and Information Engineering, National Changhua University of Education, Changhua, Taiwan, in 2011, and the M.S. degree from Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, in 2013. He is currently a Ph.D. student in Department of Computer Science and Information Engineering in National Cheng Kung University. His research interests include data mining and social media analytics.



Cheng-Te Li received the M.S. and Ph.D. degrees from the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan, in 2009 and 2013, respectively. He was an Assistant Research Fellow at the Research Center for Information Technology Innovation in Academia Sinica, Tainan, Taiwan. He is currently an Assistant Professor with the Department of Statistics, National Cheng Kung University, Tainan, Taiwan. His current research interests include social and information networks, data mining and social media analytics. Dr. Li was a recipient of the Facebook Fellowship 2012 Finalist Award, the ACM KDD Cup 2012 First Prize, the IEEE/ACM ASONAM 2011 Best Paper Award and the Microsoft Research Asia Fellowship 2010

Kun-Ta Chuang currently serves as an Associate Professor in Department of Computer Science and Information Engineering in National Cheng Kung University. He was a senior engineer at EDA giant Synopsys during 2006–2011. He received the PhD degree from Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, in 2006. His research interests include data mining, web technology and cloud computing.



Jun Pang received his Ph.D. in Computer Science from Vrije Universiteit, Amsterdam, the Netherlands, in 2004. Currently, he is a senior researcher in the Security and Trust of Software Systems research group at the University of Luxembourg. His research interests include formal methods, security and privacy, social media mining and computational systems biology.

Yang Zhang completed his Ph.D. in Computer Science from University of Luxembourg in 2016. Prior to that, he obtained his bachelor (2009) and master (2012) degree from Shandong University, China. Currently, he is a postdoctoral researcher at CISPA, Saarland Informatics Campus. His research mainly concentrates on privacy in the modern society. Topics include social network privacy, genomic privacy, location privacy, access control, urban informatics, data mining and applied machine learning.