Spatial Influence - Measuring Followship in the Real World

Huy Pham
Computer Science Department
University of Southern California
huyvpham@usc.edu

Cyrus Shahabi
Computer Science Department
University of Southern California
shahabi@usc.edu

Abstract—**Finding *influential* people in a society has been the focus of social studies for decades due to its numerous applications, such as viral marketing or spreading ideas and practices. A critical first step is to quantify the amount of influence an individual exerts on another, termed *pairwise* influence. Early social studies had to confine themselves to surveys and manual data collections for this purpose; more recent studies have exploited web data (e.g., blogs). In this paper, for the first time, we utilize people's movement in the real world (aka *spatiotemporal* data) to derive pairwise influence. We first define *followship* to capture the phenomenon of an individual visiting a real-world location (e.g., restaurant) due the influence of another individual who has visited that same location in the past. Subsequently, we coin the term *spatial influence* as the concept of inferring pairwise influence from spatiotemporal data by quantifying the amount of followship influence that an individual has on others. We then propose the Temporal and Locational Followship Model (TLFM) to estimate spatial influence, in which we study three factors that impact followship: the time delay between the visits, the popularity of the location, and the inherent coincidences in individuals' visiting behaviors. We conducted extensive experiments using various real-world datasets, which demonstrate the effectiveness of our TLFM model in quantifying spatial influence.**

I. Introduction

Social influence is a well-established term in sociology; it indicates the change in attitude, opinion or behavior that one person causes in another as a result of different forms of actions, such as interactions, recommendations or observations [1][2][3]. In the past decade, the study of social influence received a major boost due to the availability of web data (e.g., social networks, blogs, review web sites) [4][5][6][7][8][9]. However, because of the dataset they used, these studies were confined to behaviors that were observed mostly in the virtual world. Instead, in this paper, utilizing the spectacular growth of location data representing people's movements in the real world, we aim to study social influence by observing behaviors in the real world. In particular, we focus on deriving pairwise influence (the amount of influence an individual exerts on another) by analyzing location logs over time.

The pervasiveness of GPS-enabled mobile devices has introduced massive ***spatiotemporal*** data that portrays, at high resolution, the movements of people, specifically by indicating *who* has been *where* and *when* (examples of data sources: Twitter, Flickr, Foursquare, WhatsApp). Such collections of spatiotemporal data constitute a rich source of information for studying various social behaviors. One particular behavior that draws our attention is when an individual visits a location (e.g., a restaurant) due to the influence of another individual who visited that same location in the past. We define this behavior as ***followship***. Hence, followship is an indication of pairwise influence between people in the real world. Subsequently, for the first time we introduce ***spatial influence*** – a concept of inferring pairwise influence from spatiotemporal data by quantifying the followship influence that an individual exerts on another in the real world.

The most notable application of spatial influence is for identifying highly influential people: for target advertising (by giving influential people coupons/promotions so they can further spread the information to many other people under their influence), for political campaigns (by making influential people the seeds/initiators of the campaigns), for cultural studies (choosing influential people to spread ideas and good practice), etc. Spatial influence also has its own *unique* utility when the above-mentioned applications become specific and bounded to a certain geographical area. In such cases, we need to choose the seeds/individuals who are closely related to the area and are influential to people in that area. For example, the president of a university is clearly more influential to the students at that university than some general rappers.

To the best of our knowledge, no study has focused on the systematic and automatic quantification of pairwise social influence from the rich spatiotemporal data (see Section II-C). The reason for this void in the literature is due to the lack of spatiotemporal data in the past. However, this is now no longer the case. For example, Twitter and Foursquare reportedly receive millions of spatiotemporal records per day [10][11]. Consequently, the availability of this new data motivated us to study spatial influence.

Quantifying spatial influence brings up many challenges. First, we do not assume any prior knowledge about friendship information between users since this information is usually absent or *sparsely* available in spatiotemporal datasets [12]. Second, we need to distinguish actual followship from other successive visits that are *not* due to influence, which we call *coincidences*. For example, people who dined at a restaurant earlier do not necessarily influence people who dined there later; therefore, their successive visits are simply coincidences, even though they may share the same pattern as followship. This issue becomes more challenging because we do not have information about the real-world friendships in order to rule out coincidences; furthermore, even using friendships may limit the findings of influence between people who are not friends, but can still influence others, such as celebrities and fans (see Section II-C). Third, even if we can identify successive visits as followship, how should we quantify followship? Should it be a function of location (the popularity of a location), or the time delay (the time interval between visits)? Fourth, how should we measure the individual contribution of each factor (location and time delay) and then how do we combine them in a meaningful manner? Among the above-mentioned issues, the cases related to coincidences and the impact of locations are critical to spatial influence. Indeed, previous studies in geo-social networks [13] [14] showed that up to 70% of human movement is due to location-related reasons. However, previous studies in social-media influence did not consider the location factor (see Section II-C); therefore, we cannot simply adapt their solutions for spatial influence.

To address the above challenges, we propose the **T**emporal and **L**ocational **F**ollowship **M**odel (TLFM) to estimate spatial influence by taking into account the contributions of all the impacting factors. Specifically, the *temporal* followship estimates influence as the *urge* or *how soon* a person wants to visit the location following the initial visit of her influencer (aka the *time delay*), while the *locational* followship discounts the popularity of the location from that measurement. In addition, we utilize the Shannon entropy to eliminate the contribution of coincidences. Finally, we report our extensive experiments on real-world datasets collected from different location-based social networks, which proves the effectiveness of our model in quantifying spatial influence. We used the influence computed from the corresponding social networks as ground-truth to evaluate the spatial influence computed by TLFM and we observed about 70% of our inferred influence closely matches the ground truth. Finally, we developed several baseline approaches by adapting various techniques proposed to compute influence in social media, and our experimental comparisons confirm that none can effectively capture spatial influence.

Note that our solution is not limited to spatial influence; it can also be applied to social media to improve the inference of online influence due to the consideration of the popularity of online actions, which has not been considered in previous work. Also, we do not intend to separate spatial influence from social influence in all scenarios; rather, we measure followship quantitatively by particularly utilizing spatiotemporal data.

Our key contributions are summarized below:

- We introduce *followship* to capture the intuition of pairwise influence. Subsequently, for the first time, we define *spatial influence* as the concept of inferring influence from spatiotemporal data by quantifying followship.
- We propose the TLFM model to derive spatial influence from spatiotemporal data. In addition to considering the time delay between visits, we believe TLFM is the first model to consider the impacts of locations and coincidences.
- Our extensive experiments on real-world datasets collected from different location-based social networks prove the effectiveness of our model and the ineffectiveness of techniques in previous studies (developed for inferring online social influence) in capturing spatial influence.

The remainder of the paper is organized as follows: we introduce spatial influence, define the problem, discuss challenges and related work in Section II. We propose our solution in Section III, and discuss implementation and complexity in Section IV. We show the experimental results in Section V and conclude in Section VI.

II. Background

A. Social Influence Versus Spatial Influence

Social influence: When user u is said to influence user v, there is an implication of a pairwise probability p_{uv} associated with this influential relationship from u to v. Putting this in the context of actions, it means if u performs an action (such as clicking the "like" button of a Facebook fan page), then v will also perform the same action at a later time with probability p_{uv}. This is called *influence probability*, which indicates the extent of influence that u exerts on v. Generally speaking, $p_{uv} \neq p_{vu}$ and therefore influential relationships are directed. When u influences v, we say u is the ***influencer*** and v is the ***influencee***. Social influence is important because of its many potential applications and it is also a critical part of the *influence maximization* problem [7].

Spatial influence: We aim to study influence in the geo-spatial context. For this, we first define followship as the indication of influence in the real world, then we define spatial influence as the study of quantifying followship.

Definition 1: ***Followship*** is a spatiotemporal behavior where an individual v visits a location due to the influence of another individual u who has visited that same location. We say v follows u.

Definition 2: ***Spatial influence*** is the concept of pairwise influence inferred from spatiotemporal data by quantifying followship, which indicates the amount of influence that one individual u exerts on another individual v in the geo-spatial context. We use the notation $[u \to v]$ to denote the *influential relationship* in which u influences v, or v follows u.

Clarifications: First, the time delay between the visits in followship is greater or equal to zero. If it is zero, followship becomes a *co-occurrence* – two users are at the same place at the same time [12]. Second, *not* all successive visits are followship; only those caused by influence are followship, according to Definition 1. An example of followship is when people go to a restaurant because of recommendations from friends who visited the place before. An example of successive visits that are *not* followship is when non-related people happen to shop at the same mall by accident. We call such cases *coincidences*.

Definition 3: ***Coincidences*** are successive visits to the same place by different individuals, which are due to ***non-influence*** reasons – any reason *except* influence.

Fig. 1 shows an example of followship, where person v follows person u in visiting various places in Los Angeles; each may have multiple visits to the same place (not shown in the figure). In Fig. 1, there are three common places (marked in circles) visited by both u and v: a cafe, an opera house and a statue. Unfolding their visits at these common places onto the time axis, we get a more informative picture of followship in Fig.

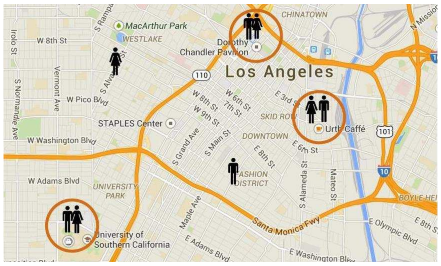


Fig. 1. Example of the visits by u and v.

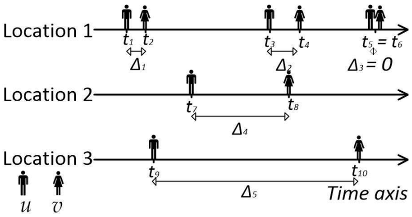


Fig. 2. The visits of u and v are unfolded along the time axis.

2, which show the time delay between visits. In Fig. 2, each horizontal time axis corresponds to one place, and the time delays are shown as Δ_1, Δ_2, etc.

B. Problem Definition & Challenges

First, we would like to clarify a few terms.

Location: Technically, a location is a *point*, which is different from a *place* (e.g., a park), which is an *area*. However, location-based social networks (LBSNs) automatically convert the check-ins at the same place to the same common coordinates at the time users perform check-ins in order to eliminate the *uncertainty* of GPS data. Therefore, we can use the two terms *place* and *location* interchangeably.

Check-in: A check-in is a record of spatiotemporal data in the form of a triplet $\langle u, l, t\rangle$, which indicates the user's ID u, location l and the time t that the user shared her location with LBSN.

Problem 1: **PROBLEM DEFINITION:**
Given a set of locations $L = \{l_1, l_2, ..., l_m\}$, a set of users $U = \{u_i, i = 1, ..., n\}$, a set of check-ins in the form of *user-location-time* triplet $\langle u, l, t\rangle$, the problem is to infer the pairwise spatial influence p between users in U.

Notes: first, the value of p is not necessarily between 0 and 1, but can be normalized to become influence probability. Second, since influential relationships are directed, $p_{u_i \rightarrow u_j}$ can be different from $p_{u_j \rightarrow u_i}$ for user pair u_i and u_j. Third, the only time-related input to the problem is the check-in's time. Another possible input is the length of stay, i.e., how long a user stays at a place. However, most LBSNs, such as Foursquare, Twitter and Gowalla, do not record the length of stay. In fact, users check in at places, but never check out. Therefore, we design our model to capture such reality by limiting the input to only user IDs, locations and check-in times. Length of stay can be a direction for future work once such information becomes readily available.

Issues related to inferring spatial influence are:

- Differentiate followship from coincidences without having to rely on explicit friendships.
- Measure the impact of the time delay between the visits in followship.
- Measure the impact of the location in followship.
- Combine the contributions of the two impacts.

C. Related Work

A large number of studies in social influence focused on an optimization problem called *influence maximization*, in which the focus is to find a small set of influential individuals (seeds), who have the most combined influential impact on an entire social network [4][15][6][7][8]. In the influence maximization problem, the pairwise influence is the most critical input and is generally assumed to be known or generated randomly for experiment purposes. More precisely, the influence maximization problem does not focus on inferring pairwise influence, but it rather utilizes the pairwise influence for optimizing the viral propagation of information in a social network. Hence, technically the influence maximization problem is orthogonal to our work, and it shares similarities with our work only in applications.

Studies focusing on pairwise influence can be categorized into two subgroups: *direct-influence* and *indirect-influence*.

With the *direct-influence* subgroup, the subjects are online blogs/articles that directly reference each other, which offers clear evidence of influence – an article influences another in a referencing event. Saito et al. [9] inferred the influence probability in a network of online blogs. Specifically, the authors modeled the problem as finding the influence probability from a series of past episodes, with each episode offering the evidence of influence events: who succeeded or failed to influence whom at successive points of time. The influence probabilities were modeled so that they would maximize the chance of observing those given episodes. A series of studies by Gomez-Rodriguez et al. [16][17][18] proposed another way of learning social influence by assuming that online events cascade along a hidden tree where nodes are blogs/articles. They then inferred the hidden tree by capturing which node influences which, based on the time when an article started referencing another. Similarly, in several studies by Zhou et al [19], and Du et al. [20], the authors also utilized the cascades of information diffusion to infer the underlying network of influence.

With the *indirect-influence* subgroup, a subject *indirectly* influences another through a common action. For example, user u liked and started following a Facebook fan page, and her friend, user v, later also started following that same fan page. Studies in this subgroup rely on explicit friendships between users (i.e., users list others as their friends in their profiles) in order to infer influence based on the actions performed by friends only, and thus do not need to worry about coincidences. Goyal et al. [5] later proposed another model to learn influence probability from logging data of users' online activities, specifically by looking at users that join online groups after their friends joined the same groups.

They proposed several models, including static, continuous and discrete time models to capture the impact of the time delay between the actions on influence.

The studies in the two above-mentioned subgroups are the most relevant to our work as they focused on inferring pairwise *social* influence. The main reason they cannot be applied to spatial influence is that they focused on using only the time delay between the events, which is not necessarily a research limitation, but rather a scope restriction of the papers. However, as we mentioned in Section I, up to 70% of human movement was found to be attributable to location-related issues: randomness, regularities between home and work places, etc. [13] [14]. Consequently, the presence of users' geo-locations in the input data introduces new challenges, and the impact of locations and coincidences in user location behaviors need to be taken into account in inferring spatial influence. Among the three challenges we mentioned – the impact of time delay, locations and coincidences, only the time delay challenge has been addressed in the previous studies for *social-media* influence [5][9] [16][17][18][19][20], whose solutions could only be a *partial* solution to spatial influence; the two other challenges remain unresolved. In Section V, we will compare our approach with these related studies experimentally.

To the best of our knowledge, there are no studies in social influence that capture the impact of the popularity of online actions, which would be equivalent to considering the impact of the popularity of locations in this paper. Note that we cannot rely on using explicit friendships to rule out coincidences. In spatial influence, friendships are *implicit*, and we have to rely on raw spatiotemporal data only in order to infer influence. Therefore, as a baseline solution, we first infer an implicit social graph from spatiotemporal data (e.g., by using the EBM model [12]), and then use this with the solution in [5] to infer pairwise influence. This is the best effort one can make by simply making use of previous social studies, but it still does not capture the impact of locations. *Therefore, capturing the impact of locations on influence by simply using the early studies in social-media influence is not possible, even when combining them with other geo-social studies* (e.g., [12]). We evaluate this straightforward approach (as a "baseline" solution) in our experiments, and the result shows our own approach was about 2x better in accuracy (see Section V-B3).

III. The TLFM Model

We organize our discussion as follows: First we discuss the contribution of a doublet's span in Section III-A, which we call *temporal dependency* of followship. We then discuss the contribution of a doublet's location in Section III-B, which we call *locational dependency* of followship. Finally, in Section III-C we discuss the issue related to *influence causality* and coincidences.

Definition 4: ***Doublet*****:** A followship-doublet (or doublet for short) is the finest element of followship; it consists of a pair of visits or check-ins, denoted as c_1 an c_2, at the *same* location l by two *different* users.

Clarifications: First, being the finest element of followship, doublet does not contain other doublets (or check-ins) by these two users. For example, in Fig. 2, at Location 1 two check-ins at t_1 and t_2 form a doublet, but two check-ins at t_1 and t_4 do not form a doublet because it is not the finest element – there are other check-ins by the users between t_1 and t_4 in that *same* location. Second, a doublet is defined with regard to an influential relationship – *who follows whom*. Therefore, the order of the check-ins in a doublet depends on the direction of followship: the check-in by the influencer precedes the check-in by the influencee. For example, from location 1 in Fig. 2, if v follows u, then the doublets are (t_1, t_2), (t_3, t_4) and (t_5, t_6). However, if u follows v, then the doublets are (t_2, t_3), (t_4, t_5) and (t_6, t_5).

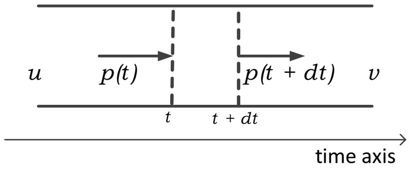


Fig. 3. Influence that u exerts on v at a location over time.

Definition 5: ***Doublet's span*****:** The ***span*** of a doublet is the time *interval* Δ between the two check-ins in the doublet.

Definition 6: ***Co-occurrence*****:** A co-occurrence between two users is a doublet whose span is 0.

The span of a co-occurrence is 0, theoretically, but we can accept a small value [12], such as less than 15 minutes due to the fact that users may perform check-ins separately in time even though they are together at a place.

Definition 7: ***Multiplet*****:** A followship-multiplet (or a multiplet for short) at a location is the set of doublets by the two users at that location.

A. Temporal Dependency of Followship

In this section, we study the relation of the doublet's span with influence (and thus the name temporal dependency).

Consider a doublet when two users u and v visited location l at time t_1 and t_2, respectively. Without loss of generality assume that u influences v in this doublet and thus $t_2 \geq t_1$. The span of the doublet $\Delta = t_2 - t_1 \geq 0$.

When u influences v, and u visits location l, it is natural to expect that v would like to visit location l under the influence of u [5][1]. The desire of v to visit l is nothing but the influence that u exerts on v. The stronger the desire, the sooner v will visit l, and thus the more influence that u exerts on v. At best, v will try to visit l together with u, which creates a co-occurrence in both space and time. On the other hand, the later v visits l, the less influence u has on v. Consequently, the influence that u exerts on v weakens over time starting from the moment t_1, or in other words, the span of the doublet inversely indicates the influence of u on v. This observation is in accordance with early studies [5][16][17], which empirically proposed that influence decays over time by a law; subsequently, our goal is to derive this law theoretically and study the parameters that govern it. Consider the geo-spatial channel that u influences v shown in Fig. 3. This illustrates the influence over a time period after u visited location l, meaning after t_1. Let $t \geq t_1$ be any moment *after* u visited location l, and $p_{u \to v}(t, l)$ be the influence of u on v at time t ***as if*** v visits location l

at t; note that t should not be confused with the span $\Delta = t - t_1$. To simplify the presentation, we omit the symbols $u \rightarrow v$ and l in $p_{u \rightarrow v}(t, l)$, which therefore become just $p(t)$. We will restore these symbols later. Furthermore, let $dt > 0$ be a temporal differential that represents an infinitely small change of variable t. At time $t+dt$, the influence of u on v is $p(t+dt)$. Correspondingly, the change of the influence over the interval dt is $dp = p(t+dt) - p(t)$. As mentioned earlier, the influence decreases over time; therefore, $dp = p(t+dt) - p(t) < 0$ shows the loss of influence over the infinitely small time interval dt.

The amount of influence expected to be lost during the infinitely short time interval dt (between t to $t + dt$) depends on the *current* value of influence $p(t)$ at t and is proportional to $p(t)$. We assume this behavior because it is also observed in other phenomena, such as in radioactive decay [21] (the more particles present, the higher the number of decays), and also in the Maxwell-Boltzmann distribution statistics [22]. Note that we will verify this experimentally. Formally:

$$\frac{p(t + dt) - p(t)}{dt} \propto p(t) \tag{1}$$

or we can write this in the form of an equation as follows:

$$\frac{dp}{dt} = \frac{p(t + dt) - p(t)}{dt} = -\frac{p(t)}{\tau}$$

where the minus sign indicates that $p(t)$ decreases over time, and τ has time unit and indicates the proportional constant, whose meaning we will clarify shortly. Rewriting the above equation in a more compact form, we have:

$$\frac{dp}{dt} = -\frac{p}{\tau} \tag{2}$$

Rewriting the equation and integrating both sides, we have:

$$\frac{dp}{p} = -\frac{dt}{\tau} \Longleftrightarrow$$

$$\int \frac{dp}{p} = -\int \frac{dt}{\tau} \Longleftrightarrow ln(p) = -\frac{t}{\tau} + C \Longleftrightarrow$$

$$e^{ln(p)} = e^{-\frac{t}{\tau}+C} = e^{C} e^{-\frac{t}{\tau}} = p_0 \, e^{-\frac{t}{\tau}}$$

where C is an integral constant. After simplifying the above equation and de-compacting it (writing $p(t)$ instead of just p), we have:

$$p(t) = p_0 e^{-\frac{t}{\tau}} \tag{3}$$

Equation (3) shows that the influence of u on v with respect to location l decays exponentially over time ($p(t) = 0$ when $t < 0$). Thus, our theoretical derivation agrees with the empirical proposals for the exponential decay of influence over time in early studies [5][16][17]. p_0 comes out as an integral constant. Recall that the doublet span is $\Delta = t - t_1$; if we set $t_1 = 0$ (i.e., the initial visit of u to location l is at the origin of the time axis), then $\Delta = t$, thus t becomes the span of the doublet if v visits the location at time t. Therefore, we omit symbol Δ and simply use t to imply the span because we can always assume $t_1 = 0$ without loss of generality.

Furthermore, by setting $t = 0$, we have $p(0) = p_0$. Therefore, p_0 has the meaning of the ***initial*** influence that u would exert on v before it starts to decay over time. In other words, p_0 is the influence of u on v at location l if they *co-occur* at l. To clarify the meaning of the constant τ, we set $p(t) = p_0/2$ – the time that the influence has decreased by half. Solving equation $p_0/2 = p_0 \; e^{-t/\tau}$, we get $t = \tau \ln 2$. Consequently, $\tau \ln 2$ has the meaning of *half-life* – the time interval, through which the influence decreases by half. We will denote this quantity as $h = \tau \ln 2$.

It is important to elaborate on the constant p_0 – the *initial* influence in Equation (3). Specifically, what does p_0 depend on? Attributes related to a doublet are the span t and the location l. As we see, the temporal factor t has already been considered and included in the exponential part $e^{-t/\tau}$ of the Equation (3). Consequently, location l is the only remaining factor that affects p_0. It may be tempting to think that everyone has a different influential capability and thus p_0 should also depend on the user. However, recall that we do not assume that each user has her own initial level of influence at the beginning, but rather we infer this entire information from followship. Therefore, p_0 should only depend on the location l. Consequently, we separate Equation (3) into two parts: the *temporal dependency* $e^{-t/\tau}$ and the *locational dependency* p_0. To better structure the locational dependency, we rewrite p_0 as $p(l)$ – a function that depends only on the location l.

Equation (3) obtains a new and more informative form:

$$p(t, l) = p(l) \times e^{-\frac{t}{\tau}} \tag{4}$$

Function $p(l)$ only depends on location l and it is considered a constant during the course of the derivation of Equation (3). Therefore the derivation remains valid.

Note: The derivation also shows how to *combine* the impacts of the time delay and the location – the multiplication in Equation (4).

To verify the validity of the exponential temporal decay of spatial influence, we conducted an experiment using a real dataset from Foursquare. This dataset has two parts: one contains the social information between users – who is friend of whom; and another part contains the spatiotemporal data. The spatiotemporal data is a set of check-ins, where each check-in is a tuple: $\langle$ *user-id, location-id, latitude, longitude, time* $\rangle$ which indicates, from left to right, the ID of the user, the ID of the location (auto-generated by Foursquare), the latitude and longitude of the location, and the time of the check-in, respectively. For the purpose of this experiment, we randomly chose 200,000 pairs of friendships out of roughly $1.4M$ friendships, for whom we computed the spans of all their doublets. In Fig. 4, we show the behavior of temporal dependency of this real dataset as to how the number of doublets of *friends* decreases as the span grows. The $x-$axis shows the doublet's span in hours, and the $y-$axis shows the number of doublets that corresponds to the span in the $x-$axis. The leftmost cross ×, for example, indicates that there are about 35,000 doublets with spans between 0 and 10 hours, and the second leftmost cross indicates there are about 32,000 doublets with spans between 10 (exclusive) and 20 (inclusive) hours.

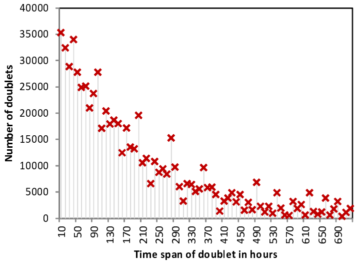


Fig. 4. Exponential Decay – Foursquare Data

From Fig. 4 we observe that the number of doublets between *friends* drops exponentially as their span grows. This in turn tells us that the tendency of people repeating their friends' actions, or their followship, decreases exponentially starting from the moment when their friends first initiated the actions. This implies that spatial influence decays exponentially over time, and hence, verifies the validity of our theoretical derivation of temporal dependency in Equations 3 and 4. From Fig. 4 we also estimated the *half-life* $h = \tau \ln 2 = 130$ hours or roughly five and a half days, which is the time when the influence decreases by half (the value on the y-axis decreases by 50%), and therefore parameter $\tau = 130/\ln 2 = 188$ hours. From this experiment, we also observed that the influence drops to a negligible level when $t > 900$ hours or roughly 38 days (not shown in Fig. 4 as it is out of range of the $x-$ axis). We call this value $T = 900$ hours the ***span threshold***, which therefore can be used as the threshold or limit for searching for doublets in spatiotemporal data.

Note that other possible considerations for the decay of $p(t)$ include linear function, power law or Rayleigh distribution. However, we will show later in Section V-B1 that experiments for spatial influence across different datasets of different LBSNs show a consistent exponential behavior; thus, we eliminate these options from further consideration, particularly in our study of spatial influence.

Next, we consider the temporal dependency in a multiplet. Without loss of generality, assume the influential direction is also from u to v, meaning u influences v. Let the multiplet m consist of multiple doublets $\{d_1, d_2, ..., d_n\}$ of two users u and v at the same location l, and since we assume that u influences v, in each of these doublets the check-in by u preceded the check-in by v.

The overall influence of u on v at l consists of the influence scored from the component doublets d_i of the multiplet m.

$$p(m) = p(l) \sum_{d_i} e^{t_{d_i}/\tau} \qquad (5)$$

where t_{d_i} is the span of doublet d_i, and we formally added m as an argument because d_i and l are defined in m. Note that because all the doublets are related to the same location l, they share the same factor $p(l)$. Equation (5) shows the ***depth*** of the influence of u on v at location l.

Note that Equation (5) implicitly assumes that the latest check-in by u, for example at t_3 in Fig. 2, immediately preceding the check-in by v at t_4, is the most and the only significant action by u in influencing v to visit the location at t_4. An example may be that Alice has been influenced by frequent visits by Bob to a cafe, and she decides to visit that same cafe. Similarly, the most recent visit by v (say $c_{v,1}$) following the visit by u at location l accounts for the most influence that v receives from u, and the subsequent visits by v to location l, are not significant. Therefore, we need to relax these assumptions by including the prior check-ins by u and the subsequent check-ins by v, if they are not already included in other doublets, in Equation (5). As a result, we obtain a new equation.

$$p^*(m) = p(l) \sum_{d_i^*} e^{t_{d_i^*}/\tau} \qquad (6)$$

where d_i^* denotes a *relaxed* doublet that allows one visit to be included in more than one doublet. The only disadvantage of this new equation is that it will add complexity to the algorithm. In the experiments, we will show that even by using Equation (5) we can achieve a high level of accuracy in quantifying spatial influence.

Last, to determine the direction of influence in a multiplet (who influences whom), we propose the following basic rule. We identify who is the first to initiate the action at location l. By looking at their check-in at the location of the multiplet, the person who performed the first check-in is the influencer and the other is the influencee. If both of them initiated this first action together (a co-occurrence), then they are considered friends [12], and thus this is a bi-directional influential relationship where friends influence each other [12][7][15]. Consequently, we have two multiplets, one for each influential direction, from u to v and from v to u. Note that we only mention this most basic rule. A more thorough consideration is also possible, for example, by considering who is the first to visit the location after neither of the two users has visited the location for a period greater than the span threshold T.

B. Locational Dependency of Followship

As we mentioned earlier, in Equation (4) $p(l)$ is the initial influence at location l that u would exert on v if they co-occur or before the influence starts to decay over time exponentially. We also argued that $p(l)$ only depends on location l and thus, in general, it shows the effect of each individual location in a doublet on social influence in the real world. In this section we elaborate on the function $p(l)$ – the impact of a location.

It is natural to expect people to visit more famous and popular places because those are frequently mentioned on online media, advertisements or word of mouth. Therefore, it is easier to convince or to influence a person to visit a popular place, such as Times Square, because the place is already well-known and thus that person's motivation may have come from many different sources of information. On the other hand, it is harder to convince someone to visit a less popular place, such as a mediocre restaurant that has a low ranking and is little or poorly known among people. Consequently, it requires a *higher* level of trust or influence from a person in order to convince other people to visit unpopular places, and if the person succeeds in doing so, then she should deserve a higher score in the quantification of her influence. This score is the function $p(l)$. The less popular the location l, the higher the rewarding score. Subsequently, we will build function $p(l)$

based on the popularity of a location in order to capture this intuition.

Note that it is ***not** impossible* to find some counter-examples (as *single-time* events) for the above-mentioned relationship between the popularity of a location and the influence. However, we are discussing followships that happen on *regular basis* that result in considerable measurements of both temporal dependency $p(t)$ and locational dependency $p(l)$. See Section III-C for more details.

To study the popularity of locations, we use Location Entropy [12][23]. At location l, let $V_{l,u} = \{< u, l, t >: \forall t\}$ be the set of check-ins by u and $V_l = \{< u, l, t >: \forall t, \forall u\}$ be the set of check-ins by *all* users. The probability that a check-in, which is selected randomly from V_l, was performed by user u is $P_{u,l} = |V_{l,u}|/|V_l|$. The Shannon entropy of location l based on this probability is given as follows:

$$H_l = -\sum_{u, P_{u,l} \neq 0} P_{u,l} \ln P_{u,l} \tag{7}$$

This entropy measures how popular a location is in terms of the people who visited it. In information theory, it is the *amount of information* about the users who visited location l. The first obvious observation is that the more visitors at l, the higher the entropy. However, the popularity of a location cannot always be described by just using the number of the visitors at the location, and this is where the entropy comes into play. The advantage of using entropy is that it measures the location popularity based on the distribution of check-ins over the users who performed them.

An important property of Location Entropy is $\exp(H_l)$, which is called the *effective number* of visitors at location l [12]. This is *not* the *actual* number of users who visited l, but it is rather an *equivalent* number of users who visited l *as if* they all visited l an equal number of times. For example, the effective number of visitors is $\exp(0.35) = 1.4$ for location l_1, and $\exp(0.69) = 2$ for location l_2. Note that the effective number does not need to be integer; it is an index to show how popular a location is.

With the support of Location Entropy, let us go back and discuss function $p(l)$ – the locational dependency part of Equation (4). As we mentioned earlier in this section, $p(l)$ is to capture the *initial* influence that u would exert on v before this value starts to decay over time. This initial influence should be *inversely* proportional to the popularity of location l, for which we accept the inverse quantity of the effective number of visitors at l, that is $p(l) = \exp(-H_l)$. We rewrite Equation (4) as follows:

$$p(l, t) = \exp(-H_l) \times \exp(-t/\tau)$$

$$= \exp\left(\sum_{u, P_{u,l} \neq 0} P_{u,l} \ln P_{u,l}\right) \times \exp\left(-\frac{t}{\tau}\right) \tag{8}$$

The first exponential expression shows the locational dependency, and the second shows the temporal dependency of followship. Equation (5) also obtains a similar form due to the same factor $p(l)$ that is common for all doublets at location l.

$$p(m) = \exp\left(\sum_{u, P_{u,l} \neq 0} P_{u,l} \ln P_{u,l}\right) \times \sum_{d_i} \exp\left(-\frac{t_{d_i}}{\tau}\right) \tag{9}$$

When u influences v in multiple locations $\{l_1, l_2, ..., l_s\}$, and produces in each location l_k a corresponding multiplet m_{l_k}, the overall influence value is:

$$p_{u \to v} = \sum_{l_k} p(m_{l_k}) \tag{10}$$

1) Credit distribution via locational dependency: An important issue concerning the influencer-influencee relationship is the distribution of credit among multiple influencers when they all share the credit of influencing the same person to visit a location. Particularly, when user v visited a location l as a result of being influenced, we may expect that v was influenced by many people, and thus the credit for convincing v to visit l must be distributed appropriately among the influencers. Hence, the question: how do we distribute the influence credit among the influencers?

The answer is that we have already accomplished the credit distribution *implicitly* via the locational dependency $p(l)$. $p(l)$ means the initial credit for an influencer for convincing someone to visit location l. This function is built based on all the visits to location l in order to take into account the popularity of the location; the more popular, the less credit. Thus it implies that we are implicitly distributing the influence credit via $p(l)$; the more visitors to l means there are more influencers to v, but at the same time, also means that the initial credit $p(l)$ becomes smaller (according to Equations 7 and 9); consequently, each influencer will get less credit $p(l)$ as a result of sharing the total credit with the others.

C. Eliminating Coincidences

So far, we have elaborated on estimating spatial influence by measuring followship between two people via Equations (8) (9) and (10). However, we left out one important question: are all successive visits by two users to the same location (multiplets) *definite* indication of influence? The answer is obviously "no". For example, two people may alternate their work schedule on different days at the same cafe, or two professors may teach the same class at different times. In such cases, there is usually little or no influence between individuals. Recall that we refer to such cases as coincidences (see Section II). Hence, we need to distinguish followship from coincidences. We refer to this issue as the ***causality*** of influence. We set aside this issue temporarily in the previous two sections is because we wanted to focus purely on measuring followship without worrying about causes. However, we now discuss this important matter.

We rely on social statistics to reduce/discount the measurement of influence in the cases that are more probable of coincidences. For this purpose, one can use different statistics to take into account various factors to help eliminate coincidences, such as the similarity between users in socio-demographics (age, ethnicity, religion, education, profession), or the similarity in users' location behaviors (such as cosine similarity in [14]), etc. However, to keep our approach within the scope of focusing on spatiotemporal data, we propose to use the findings in previous social studies [12][24], which showed that strong relationships in the real world should be associated not only with the frequency of their common actions, but also with the diversity of their common actions.

Specifically, putting this in the context of our problem, the strong influence of u and v not only results in a large number of times that v followed u, but also results in a high number of unique locations, in which v followed u. For example, if we observe that v followed u often, but in only a single location, then that may indicate coincidences, but not influence, because u and v may just be two workers who alternate their schedules or two professors who teach the same class at different time. However, if we observe their followship not only frequently, but also in various locations (e.g., friends hang out together in various places), then that is a better indication of influence and less probable of coincidences. We refer to the later as the ***diversity*** of followship in terms of locations. Consequently, successive visits with higher diversity are less prone to coincidences, while those with low diversity are more prone to coincidences.

Diversity captures another aspect of spatial influence, in addition to the ***depth*** of influence in *each* location described by Equation (5).

1) Measuring Diversity of Successive Visits: To measure the diversity, we can use Shannon entropy or Renyi entropy (a more advanced method as shown in [12]). For demonstration, we show how to use Shannon entropy to compute diversity. To ease the presentation, we still keep using the terms doublets and multiplets, even though they may refer to followship or coincidences. However, we will use the term *successive visits* instead of *followship*.

Let $M_{u,v} = \{m_{l_1}, m_{l_2}, ..., m_{l_s}\}$ be the set of multiplets between users u and v where each multiplet m_{l_k} corresponds to location l_k. Without losing generality, assume that the direction of influence is u influences v. We aim to use Shannon entropy to measure the *amount of location information* contained in this set of multiplets. Recall from Section III-B where we argued that using the number of unique visitors to a location may not correctly describe how popular a location is. For the same reason, using the number of unique locations in $M_{u,v}$ also may not correctly characterize its diversity of locations. That is why we need entropy.

Let $p_{t,k} = \sum_{d_i} \exp(t_{d_i}/\tau)$ be the *depth* of $M_{u,v}$ at location l_k (the temporal dependency of m_{l_k}), and let $p_t = \sum_k p_{t,k}$. The Shannon entropy is defined as follows:

$$H(M_{u,v}) = -\sum_k \left(\frac{p_{t,k}}{p_t}\right) \ln \left(\frac{p_{t,k}}{p_t}\right) \qquad (11)$$

Note that because we are interested in the diversity of $M_{u,v}$ in terms of locations, the inherent property of each location should not matter. Thus, in Equation (11) we do not include $p(l)$ (locational dependency) which only characterizes the inherent property (popularity) of each location.

A high value of $H(M_{u,v})$ indicates that the successive visits by u and v are followship (due to the influence of u on v), while a low value indicates the high possibility of coincidences. Subsequently, by applying $H(M_{u,v})$ in form of a filter (named *causality filter*), we can discount the effect of coincidences. Let $\sum_{l_k} p(m_{l_k})$ (Equation 10) be the measure of the successive visits of u and v, the real influence of u on v after discounting coincidences is defined as follows.

$$p_{u \to v} = (H(M_{u,v}) + \alpha) \times \sum_{l_k} p(m_{l_k}) \qquad (12)$$

We include parameter α in Equation (12) because $H(M_{u,v}) = 0$ in case where the successive visits occurred in only one location. This parameter is application-dependent and can be set to a small value if we do not want to completely eliminate successive visits with one location from consideration. Generally, $f(u,v) = H(M_{u,v}) + \alpha$ needs to be normalized to the range $[0, 1]$.

In Equation (12), even though coincidences may create a large number of successive visits and result in high values of the measure $\sum_{l_k} p(m_{l_k})$, the filter $f(u,v)$ can effectively reduce their effect on $p_{u \to v}$ because $f(u,v)$ is negligibly small due to the lack of location diversity in coincidences.

IV. Implementation

The computational complexity of TLFM comes from the search for doublets in spatiotemporal data. Therefore, we provide the implementation of searching for doublets in Algorithm 1. Note that the algorithm is straightforward; however, our goal of presenting it is to analyze the complexity. Inputs include (i) a list $c[i]$ of check-ins at a given location l sorted by increasing time; (ii) an empty hashmap H where the key stores the user pair's IDs (e.g., "12345:678" – influencer precedes influencee), and the value stores a list of their doublets; (iii) the span threshold T. The algorithm returns H. Primitive functions used are $u(c_i)$ and $t(c_i)$, which return the user and the time of a check-in c_i; clear(I) – erases the content of a set I.

The algorithm is self-explanatory. It scans through the list of check-ins at each location only once (index i in Algorithm 1). For each check-in by index i, it only scans for subsequent check-ins within the time interval T (the span threshold) because check-ins by two users that span beyond this interval are considered to have negligible relations with each other in terms of influence. The time complexity for each location is $O(n\xi)$ where ξ is the maximum number of check-ins within a time interval T, and n is the total number of check-ins at that location. The space complexity is also $O(n\xi)$; for each check-in indexed by i, there are at most ξ doublets that can be constructed with $c[i]$ (which is the preceding check-in in a doublet). The overall time and space complexities for a set of spatiotemporal data with N check-ins are $O(N\xi)$ and $O(N\xi)$.

Since the search for doublets in each location can be performed separately and independently, one can use MapReduce to parallelize the doublet search for a spatiotemporal dataset to further improve the efficiency. In a related study in social connections [12], the authors provide a straightforward MapReduce implementation for the search of co-occurrences, which can also be applied for searching doublets. We refer the readers to [12] for more details.

Algorithm 1: Doublet Search

```
Input: $c[i]$ – a list of check-ins, $H$ – empty hashmap;
       $T$ – span threshold
Output: $H$
Create an empty set $I$ for users' IDs; $n =$ size of $c$; $i = 0$;
while ($i < n$) do
    $j = i + 1$;
    while (true) do
        if (($j == n$) or ($t(c[j]) - t(c[i]) > T$) or
                ($u(c[i]) == u(c[j])$) )
            clear($I$);
            break; end if;
        if ($I$ does not contain $u(c_j)$)
            Add ($c[i]$,$c[j]$) to hashmap $H$
                        under key "$c[i]$:$c[j]$";
            Add $u(c[j])$ to $I$;
        end if;
        $j$++;
    end while;
    $i$++;
end while;
return $H$;
```

V. Performance Evaluations

A. Data and Experiment Set-up

For experiments we used three different datasets from three LBSNs: Gowalla, BrightKite and Foursquare, which were collected during the periods Feb 2009 to Oct 2010, Feb 2008 to Oct 2010 and Feb 2012 to Jul 2012, respectively. Each of these datasets contains spatiotemporal data which we use to derive spatial influence, and an explicit social graph which we use as ground truth. Table 2 summarizes these datasets. *M* denotes million, *k* denotes thousand, and "Check-ins' # " means the number of check-ins.

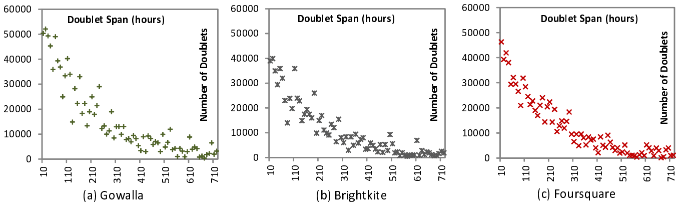
TABLE I. Datasets

	Check-ins' #	Users' #	Friendships' #
Gowalla	6.4*M*	197*k*	951*k*
Brightkite	4.5*M*	58*k*	214*k*
Foursquare	1.3*M*	669*k*	1.4*M*

We divide each dataset into two parts: ***training*** and ***evaluation***. We randomly pick three out of every five *locations* and choose all the check-ins in those locations to add to the training set, and use the check-ins of the remaining locations for the evaluation set. In the end, the numbers of check-ins for the training set and evaluation set for each of the datasets Gowalla, Brightkite and Foursquare are (4.3*M*, 2.1*M*), (2.5*M*, 2.0*M*) and (0.8*M*, 0.5*M*), respectively. Our experiments are written in Java and run in a 64-bit OS X with 16GB memory and 2.20 GHz CPU quad-core.

B. Experiments

One of the major challenges of this work is to obtain a dataset, for which the ground truth of pairwise influence is available for every pair of users. However, this lack of ground-truth is not unique to our work, but is also a challenge for all the other studies in social influence [5][7][14]. Because of this challenge, Kempe et al. [7] proposed an indirect technique for measuring the ground truth for pairwise influence based on another study [25] by Liben-Nowell et al., which was verified by numerous experiments and since then has been widely accepted and used in social network studies. Therefore, in our study, we also adapted this widely accepted technique. While this is not as good as direct ground-truth, it is the best that we (or any other study in social networks for that matter) could do for experimental evaluations considering that a dataset with direct ground truth for pairwise influence is nearly nonexistent today. We also applied this similar approach in our previous study for evaluating social strength, EBM [12].



(a) Gowalla (b) Brightkite (c) Foursquare

Fig. 5. Temporal dependency is shown as how the number of doublets depends on their time span in hours.

1) Temporal Dependency: In our first set of experiments with the training datasets, our goal is to verify the exponential behavior in the temporal dependency of spatial influence (see Section III-A), and learn the parameter *half-life* h and *span threshold* T. Specifically, we examine how the number of doublets of friendships depends on the doublet's span. Recall that we already showed this experiment with a smaller and different subset of data from Foursquare in Section III-A as an illustration of our theoretical derivation. However, we repeated this experiment with larger data from different LBSNs to show the consistency of the exponential decay for spatial influence across different networks. Note that we only perform the experiments for user pairs who are explicit friends, because we can assume the evidence of existing influence [3][5].

The results for data from three LBSNs are shown in Fig. 5. The description of Fig. 5 is similar to that of Fig. 4. We observe that all three datasets exhibit consistent behavior: the number of doublets of friendships drops exponentially as their span grows. This indicates that the tendency of people repeating their friends' actions (aka followship) drops exponentially over time starting from the moment when their friends first initiate the actions. Consequently, the spatial influence that people exert on their friends decays exponentially over time, which confirms the validity of our theory about temporal dependency of spatial influence in Section III-A. In addition, based on this experiment we estimated the *half-life* and the *span threshold* parameters (in hours) for each dataset to be approximately (h,T) = (130,1000), (135,850) and (130,900) for Gowalla, Brightkite and Foursquare, respectively; the span threshold is beyond the limit of the figures. Overall, the fluctuations in these parameters across the datasets are not significant considering experimental uncertainty and possibly different motivations for sharing locations in each LBSN. Parameter $\tau = h/\ln 2$, learned from this experiment with the training datasets, will be used in subsequent experiments with the evaluation datasets to compute temporal dependency.

Note that as an alternative to our temporal dependency we could use one of the solutions in the previous studies in social influence mentioned in Section II-C. In the next section, we implement the solution proposed in [5] as a baseline solution.

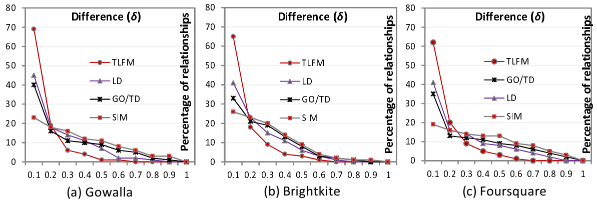
2) Comparison of TLFM with Related Work: In this set of experiments, we use the explicit social graph of each network as ground truth to evaluate our method and the methods proposed in the related studies discussed in Section II-C.

We created the ground truth's influence from an explicit social graph by using a standard well-known technique [7] as follows. First, we used the Jaccard index (proposed in [25]) $J(u, v) = |F(u) \cap F(v)|/|F(u) \cup F(v)|$ to compute social strength for each friendship u and v ($F(u)$ denotes the set of friends of u). Then we applied a technique in [7] to divide this strength $J(u, v)$ by the number of friends of v to get the influence of u on v, which is denoted as $s(u \rightarrow v) = J(u, v)/|F(v)|$. Similarly, $s(v \rightarrow u) = J(u, v)/|F(u)|$.

On the other hand, we computed predicted spatial influence from spatiotemporal data by using the TLFM model and the models proposed in the related studies discussed in Section II-C. Specifically, we evaluate the model using Goyal et al. [5] (denoted as **GO**); GO is an approach to estimate influence from a social network. To create a baseline, we used our prior work, EBM [12], to first build a social network from users' locations, and then used GO on top of this network to compute influence. We built another baseline solution using SIM [14] by considering the cosine similarity of two visit patterns of two users as influence. Neither GO nor SIM consider the location factor, including the popularity of locations and coincidences; SIM does not consider time delay. Thus, these two models may mistakenly conclude that two random visitors, say at Times Square, may influence each other. Therefore, we expect them to perform poorly with location data.

In addition, we also evaluate each component of TLFM separately, which includes the *temporal dependency* alone (denoted as **TD**) by setting $p(l) = 1$ for every location in Equation 9, and the *locational dependency* alone (denoted as **LD**) by using only the first exponential part of Equation 8. Note that in this set of experiments, GO is the same as TD, and we do not need to use a filter because we intend to compute influence between friends only in order to perform evaluations against the ground truth.

We normalized the inferred influence p and the ground truth influence s to $[0, 1]$. For each influential relationship, say $u \rightarrow v$, we computed a quantity $\delta = |p(u \rightarrow v) - s(u \rightarrow v)|$, which shows how much the inferred influence differs from the ground truth. Finally, we computed the number of influential relationships, for whom our inferred influence differed from the ground truth by a given δ. In other words, we want to answer the question: *For how many relationships does the result differ from the ground truth, and by how much?* Note that we used percentage for "how many".



(a) Gowalla (b) Brightkite (c) Foursquare

Fig. 6. Comparison of TLFM with Related Work.

Figure 6 shows the results. The x-axis shows the *difference* δ. The y-axis shows the percentage of influential relationships. How to read the graph: the left most dot for the TLFM curve in Fig. 6(a) states that for 70% of relationships, our inferred influence differs from the ground truth by a value between 0 and 0.1. The next dot states that for 19% of relationships, our result differs from the ground truth by a value between 0.1 and 0.2, etc. **Note:** A good method should have a *high* percentage for *low* δ (meaning a large portion of the inferred pairwise influence is close to the ground truth), and a *low* percentage for *high* δ (meaning only a small portion of the results differs widely from the ground truth).

Observations: We start the observations with the performance of the SIM and GO methods from the related studies. Fig. 6 shows that only small portions of the results (25% for SIM and 35% for GO) can be considered accurate (differing from the ground truth by less than 0.1), while a large portion of the results differs significantly from the ground truth. SIM has the worst performance due to the fact that if two users visited the same locations for a similar number of times, it generally does not indicate influence for two reasons. First, one user might visit a location a long time (months or years) before the other does; thus the influence, even if present, may have already decayed to null. Second, if their common locations are popular and famous, their visits might just be random as argued by the study [14]. Note that the main purpose of this study is to account for the impact of randomness and locations in location behaviors of users. The GO model (or TD) also has poor performance because a significant portion of followships is not due to influence, but is because users visited popular locations due to their own knowledge (e.g., from word-of-mouth, social media, etc.). Thus, a person may be over-credited for influencing her friends to go to a well-known and popular place. Locational dependency alone (LD) also has low performance because the decay of influence over time is completely ignored in LD.

However, when TD and LD are combined together into the TLFM model, we observe a significant improvement in the performance for the obvious reason: both temporal decay and impact of locations are captured by TLFM. Specifically, Figure 6(a) depicts that the high accuracy of our inferred influence (δ between 0 and 0.1) is observed in 70% of relationships (improved by 30% as compared to TD). Moreover, the large δ or poor accuracy is observed in significantly fewer relationships: δ between 0.2 and 0.3 is observed in only 6% for TLFM, as compared to 11% for TD and 16% for LD. This improvement is observed consistently for all three datasets Gowalla, Brightkite and Foursquare in Figs. 6(a), (b), and (c).

3) The Effect of Coincidences: We divide this section into two parts.

Part 1: In this part we evaluate the impact of coincidences and the effectiveness of the causality filter based on the Shannon entropy in reducing the effect of coincidences (see Section III-C1). We do this in two steps: first, we prove the correctness of the filter by verifying it against only friendships, for whom we can assume the evidence of existing influence

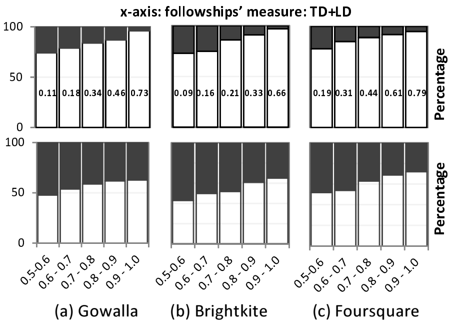


Fig. 7. The effect of the causality filter for friendships only (upper graphs), and for all relationships (lower graphs). The white bars indicate the percentage of pairs that pass the filter; black bars – fail the filter.

[3][5]; second, we apply the filter to all user pairs (including friendships and non-friendships) to show its effectiveness in eliminating coincidences.

Our experimental methodology is as follows. (**a**) First, we make an incorrect assumption that all the successive visits (doublets/multiplets) found in the evaluation dataset are due to influence, aka followship. Subsequently, we measure and normalize the successive visits for *all* user pairs by using only a combination of temporal dependency (TD) and locational dependency (LD), denoted as **TD+LD**, according to Equation (10), without using any causality filter. (**b**) Second, we compute and normalize the diversity of influence (Equation (11)) for those relationships in step (a) whose followship's measure is above 0.5. (**c**) Third, we divide the relationships obtained from step (b) into five subgroups, each with TD+TL measure falling on each of the sub-intervals $[0.5, 0.6]$, $[0.6, 0.7]$, etc. For *each* subgroup, we compute (i) the percentages of friendships who *pass* and *fail* the filter (diversity above and below the threshold 0.1, respectively), and similarly (ii) the percentages of *total* pairs (including friendships and non-friendships) which *pass* and *fail* the filter. For friendships in each subgroup, we also computed their average ground truth influence using the method that we used earlier in Section V-B2. Note that only for the purpose of experiment, we set a threshold for the causality filter; in general, however, one does not need to set a threshold for the filter.

Figure 7 shows the results. The x-axis shows the normalized followship's measure computed by TD+LD, the y-axis shows percentage. For each dataset (for example, Figure 7(a) shows the results for Gowalla's dataset), the upper figure shows that the percentage of friendships in each subgroup passed and failed the filter, and the lower figure shows the percentage of all relationships in each subgroup that passed and failed the filter. How to read the graph: for the two leftmost graphs (for Gowalla's dataset), the leftmost column in the upper graph states that for the subgroup with medium followship's measure by TD+LD ($[0.5, 0.6]$), 74% of friendships passed the filter while the remaining 26% failed; the leftmost column in the lower graph states that 48% of total relationships in that same subgroup passed the filter while 52% failed. The average ground truth influence for friendships in the subgroup is shown inside each column of the upper graphs.

Observation 1: the upper graphs show that the majority of friendships passed the filter, especially those friendships with higher ground truth influence (96% of friendships in Gowalla with an average ground truth influence of 0.73 passing the filter). Since friendships result in social influence [5][7][1], their successive visits are generally not considered coincidences. Consequently, this observation proves that the causality filter behaves correctly.

Observation 2: Relying on this filter, we then analyze the results shown in the lower graphs, which show the pass/fail for all user pairs (friendships and non-friendships). We observe that more than half of the total pairs with extensive successive visits (measured above 0.5) still failed the filter (extensive coincidences), while they would have been considered influential if no filter had been applied. By applying the filter, we therefore can reduce the impact of coincidences effectively. Note that in Fig. 7, we only show the results for the user pairs with extensive successive visits (measured above 0.5) to illustrate the effectiveness of the filter.

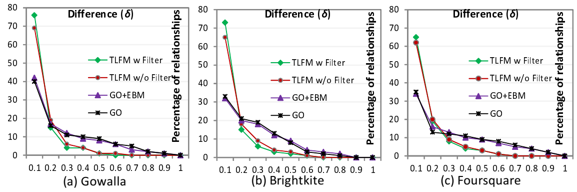


Fig. 8. The impact of filter on spatial influence.

__Part 2:__ In this part we repeat the experiments in Section V-B2 with the following changes. First, even though we still conduct the experiments for only pairs of friends (in order to be able to perform evaluation against the ground truth), we do not assume this knowledge. Therefore, we need to use Equation (12) to compute pairwise influence (by including the filter in the value) (denoted as **TLFM w Filter**). This is the main difference as compared to the experiments in Section V-B2, where we did not include the filter. As for comparison, we also compute influence by using the solution in a related study [5]; this solution needs to rely on friendships, thus, we use the EBM model [12] to infer implicit friendships from spatiotemporal data, following our discussion in the last paragraph of Section II-C (denoted as **GO+EBM**). For comparisons, we also include the result of TLFM without using the filter (denoted as **TLFM w/o Filter**), and of the GO solution from Section V-B2 (denoted as **GO**). The results are shown in Fig. 8.

Observations: We observe that the Shannon entropy-based filter improves the performance by reducing the difference between the inferred pairwise influence and the ground truth, as compared to when no filter is used. This confirms our argument in Section III-C1: the diversity of influence (ability to influence people in various locations) is as important as the depth of influence (strong influence in only one or few locations). On the other hand, the performance of GO+EBM does not improve as compared to GO, and both differ significantly from the ground truth. We attribute this to the fact that the impact of locations on influence is not considered in the GO model, which, consequently, does not discount the impact of location popularity from followship. Recall that Zhang et al. also showed that about 60% of the similarity of the location behaviors of users is due to randomness and location, but not due to influence; this study, together with the result of Location Dependency (LD) in Section V-B2, confirm that the solutions proposed for social media cannot be applied for spatial influence.

4) Efficiency: Table 3 shows the running time of the experiments. Time is measured in minute (m) and second (s). Note that the GO model [5] considers time-delay, therefore, it also needs to utilize doublets. Subsequently, the TLFM and GO models share similar time complexity due to the same doublet searching process.

TABLE II. The running time of searching for doublets.

Dataset	Gowalla	Brightkite	Foursquare
Number of Check-ins	$6.4M$	$4.5M$	$1.3M$
Running Time	$11m\ 42s$	$8m\ 13s$	$3m\ 19s$

As we see in Table 3, it took approximately 12 minutes to search for doublets in the Gowalla dataset in our simple experimental settings (without the use of MapReduce). Generally, the running time depends not only on the number of check-ins, but also on the number of unique locations in the set of check-ins because the search for doublets in each location is processed sequentially and the transitions between the locations in the input can prolong the running time. Finally, our solution is an offline process because it uses data collected over a time period (months or years), and the result needs not be updated every time a person performs a check-in.

VI. Conclusions

In this paper we focused on deriving pairwise influence from spatiotemporal data by analyzing people's movements in the real world – a concept we named spatial influence. We identified a number of challenges with spatial influence, and to address them we presented the TLFM model. TLFM includes formulations to account for the decay of spatial influence over time, to capture the impact of each individual location by considering its popularity, and to account for the role of coincidences. We conducted extensive experiments with real world datasets and confirmed the high accuracy of our model. Specifically, for up to 70% of the influential relationship, our predicted influence closely matches the ground truth. Compared to the related studies in social influence, our work is a valuable complementary method in measuring influence. Finally, our thorough comparative experiments with related studies in computing influence from social media showed that those solutions cannot capture spatial influence.

Acknowledgment

This research has been funded by NSF grants IIS-1320149 and CNS-1461963, and the USC Integrated Media Systems Center (IMSC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the sponsors such as NSF.

References

[1] W. Chen, L. V. Lakshmanan, and C. Castillo, "Information and influence propagation in social networks," *Synthesis Lectures on Data Management*, vol. 5, no. 4, pp. 1–177, 2013.

[2] H. C. Kelman, "Compliance, identification, and internalization: Three processes of attitude change," *Journal of conflict resolution*, pp. 51–60, 1958.

[3] E. M. Rogers, *Diffusion of innovations*, 2010.

[4] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *ACM SIGKDD*, 2010, pp. 1029–1038.

[5] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *ACM WSDM*, New York, NY, USA, 2010, pp. 241–250.

[6] ——, "A data-based approach to social influence maximization," *VLDB*, vol. 5, no. 1, pp. 73–84, 2011.

[7] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *ACM SIGKDD*. ACM, 2003, pp. 137–146.

[8] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, and N. Glance, "Cost-effective outbreak detection in networks," in *ACM SIGKDD*, 2007, pp. 420–429.

[9] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *KBII - ES*. Springer, 2008, pp. 67–75.

[10] C. Weidemann. http://geosocialfootprint.com/.

[11] tech.fortune.cnn.com/2013/03/18/today-in-tech-hulu-tk/.

[12] H. Pham, C. Shahabi, and Y. Liu, "Ebm: an entropy-based model to infer social strength from spatiotemporal data," in *ACM SIGMOD*. ACM, 2013, pp. 265–276.

[13] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *ACM SIGKDD*. ACM, 2011, pp. 1082–1090.

[14] K. Zhang and K. Pelechrinis, "Understanding spatial homophily: the case of peer influence and social selection," in *WWW*. ACM, 2014, pp. 271–282.

[15] P. Domingos and M. Richardson, "Mining the network value of customers," in *ACM SIGKDD*. ACM, 2001, pp. 57–66.

[16] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *In ACM SIGKDD*, 2010, pp. 1019–1028.

[17] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in *In ACM WSDM*, 2013, pp. 23–32.

[18] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," *arXiv preprint arXiv:1105.0697*, 2011.

[19] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes," in *AI and Statistics*, 2013, pp. 641–649.

[20] N. Du, L. Song, M. Yuan, and A. J. Smola, "Learning networks of heterogeneous influence," in *Advances in Neural Information Processing Systems*, 2012, pp. 2780–2788.

[21] K. S. Krane, "Introductory nuclear physics," 1987.

[22] D. Sivukhin, "A course of general physics. vol. ii, thermodynamics and molecular physics," 1990.

[23] J. Cranshaw, E. Toch, J. Hong, and A. Kittur, "Bridging the gap between physical location and online social networks," in *ACM Ubiquitous computing*, 2010, pp. 119–128.

[24] D. J. Crandall, L. Backstrom, D. Cosley, and J. Kleinberg, "Inferring social ties from geographic coincidences," *NAS*, pp. 22 436–22 441, 2010.

[25] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the ASIST*, vol. 58, no. 7, pp. 1019–1031, 2007.