Measuring Spatial Influence of Twitter Users by Interactions

Hong Wei
Department of Computer Science
University of Maryland
College Park, Maryland 20742
hyw@cs.umd.edu

Jagan Sankaranarayanan
UMIACS
University of Maryland
College Park, Maryland 20742
jagan@umiacs.umd.edu

Hanan Samet
Department of Computer Science
University of Maryland
College Park, Maryland 20742
hjs@cs.umd.edu

ABSTRACT

The three ways of interactions in Twitter–*retweet*, *reply*, and *mention*–comprise of a latent dynamic information flow network between users, which can be utilized to determine influential users. This paper focuses on determining which Twitter users have great influence on a query location Q in the sense that they are assumed to provide information that is of sufficient interest to prompt people at Q to interact with them. Note that an influential Twitter user who is of great influence on Q may not be necessarily from Q. Therefore, we first define ***generalized influential Twitter users*** regardless of whether their location was known or not, meaning that such generalized influencers on Q can be either from inside Q, or outside Q, or even unknown. A more interesting subset of generalized influencers is the ones whose location is in Q, and termed as ***local influential Twitter users***. One potential application of finding local influencers (e.g., local news media) is to detect local events by tracking their tweets.

Using a large amount of data collected from Twitter, we first build a large-scale directed interaction graph of Twitter users and present an analysis of the geographical characteristics of the edges in this interaction graph and make several interesting observations. Based on these findings, we propose two versions of PageRank that measure spatial influence on the interaction graph: Edge-Local PageRank (ELPR), and Source-Vertex-Locality PageRank (SVLPR), which takes into account the spatial locality of edges and the spatial locality of source vertices in edges, respectively. In addition, a Geographical PageRank (GPR) is also proposed trying to incorporate both of these two factors together. In the experimental evaluation, we examine the effectiveness of the proposed methods with regards to 3 different US cities "Boston, MA", "Bristol, CT" and "Seattle, WA", and the results show that our algorithms outperform their baseline approaches including the topological network metrics and the original PageRank. In addition, we also explored the possibility of using local influential Twitter users as potential news seeders and showed that some types of influential users have high credibility in outputting local place-relevant tweets.

CCS CONCEPTS

•**Applied computing** →*Law, social and behavioral sciences;*

KEYWORDS

Spatial Influence, Interaction Graph, Spatial Locality, PageRank, Twitter, Social Network, News Seeders, Local News

LENS'17, Los Angeles Area, CA, USA

DOI: 10.1145/3148044.3148046

ACM Reference format:
Hong Wei, Jagan Sankaranarayanan, and Hanan Samet. 2017. Measuring Spatial Influence of Twitter Users by Interactions. In *Proceedings of 1st ACM SIGSPATIAL Workshop on Analytics for Local Events and News, Los Angeles Area, CA, USA, November 7–10, 2017 (LENS'17),* 10 pages.
DOI: 10.1145/3148044.3148046

1 INTRODUCTION

Twitter, one of the most popular micro-blogging services, allows users to publish short messages, called *tweets*, on various subjects. In Twitter, each user, can subscribe to another user to receive the contents the latter publishes, through "following" the latter. In so doing, the former user becomes one of the latter's "followers", and the latter becomes one of the former's "friends". The definition of making a "friend" on Twitter is different from establishing a reciprocal "friend" relationship in other social network services like Facebook, because such a "following" operation in Twitter completes without requiring the user being followed to grant permission nor follow back the user who initiates the "following", which generates a directed follower-following relationship. With users as vertices, and their directed relations of following and being followed as edges, a social network in Twitter builds up.

Such a direct follower-following social graph, however, is not always available due to the difficulties imposed by the Twitter API rate limits in obtaining an access to a complete social network link structures in current Twitter. On the other hand, Twitter offers a few more dynamic ways to interact with other people such as *retweet, reply, mention*, which have been utilized in some works to determine influential Twitter users such as [1, 2]. Different from the absence of follower-following relationship, interactions are embedded in the meta-data of tweets and need no further request to Twitter API once a tweet dataset is ready. Therefore, one of the popular strategies is to first rebuild a social network from interactions and then determine influential users in the interaction graph [3–12].

But few of the existing approaches to determining the influence of Twitter users on the social network gives credits to where they are from and furthermore how close they are. Therefore, in this paper, we are focusing on answering the following question: ***Given a query location Q, which Twitter users have great influence on it?*** We refer Q to a circular region defined by a geographical center point l_q and a radius ϵ.

We consider a Twitter user to be spatially influential at Q if his authority has been endorsed by the local people from Q. We deem the interactions (*retweet, reply, mention*) one user initiates to another as his endorsement to the latter's authority. In essence, the more people from a location endorse a Twitter user, more spatially influential he becomes on that location. In this definition, we don't require a Twitter user to have to be from location Q (e.g., his home location falls within Q) to be considered influential there. In such sense, the infuential Twitter users are termed as ***generalized influential Twitter users*** on Q. The more interesting subset of the

influential Twitter users on Q is the ones who are also from location Q, termed as ***local influential Twitter users*** on Q.

Solving this problem is beneficial to many applications like targeted advertising, political campaign, trend analysis [13], and location-based recommendation [14]. In particular, finding local influential Twitter users also has the potential to discovering local news and events. For example, the Twitter accounts representing local news media usually cover and deliver information in their posted tweets regarding what is happening at a location and can be utilized as news seeders [15, 16] to help news detection [17, 18]. To test the viability of local influential Twitter users in such applications, we examined the tweets published by a set of top influential Twitter users in Boston for a week. The results show that more than half of the tweets are considered local by virtue of discussing content relevant to the local place, and the ratio of local tweets goes higher if only considering specific group of users such as news person (e.g., News Media and Reporter) and sports person (e.g., Sport Player and Sport Team).

In this paper, we first build a large-scale directed interaction graph. An intuitive solution to finding influential Twitter users then is to append a post-processing location filter step after finding influential people in general. For example, one can rely on the indegrees of vertices in the interaction graph or apply the PageRank schema to yield a ranking order for Twitter users regarding their influence, and then select the ones who fall within Q and identify them as influential Twitter users on Q.

Our proposed methods improve over this strategy by additionally considering spatial locality in the edges and its source vertices respectively. Specifically, by emphasizing on spatially local edges (applying a exponential distance-decay on the edges), i.e., whose two vertices have smaller geographical distances, our method Edge-Locality PageRank (ELPR) more effectively find ***local influential Twitter users*** than the network metrics like indegree and the original PageRank. By focusing on the edges whose source vertices are spatially local to the query location center l_q, our method Source-Vertex-Locality PageRank (SVLPR) doesn't rely on a post-processing location filter step and thereby also capture the Twitter users who are of great influence on but not necessarily from the location Q. The experiments also show that SVLPR outperforms its indegree-based baseline approach. Moreover, our hybrid method Geographical PageRank (GPR) attempts to bring geographical distances among Twitter users into the process of finding spatially influential Twitter users, which improves over PageRank by taking into consideration both link structure and geographical distance during propagating influence among users.

The rest of this paper is organized as follows. In Section 2, we review related work. In Section 3, we describe the dataset we are using, along with an interaction graph built from it. In Section 4, we first present our two methods to determine local and generalized influential Twitter users, respectively and additionally propose a hybrid method trying to combine them. Section 5 describes the experimental evaluation of our methods. Concluding remarks are drawn in Section 6.

2 RELATED WORK

There has been a plural of works on identifying the influential users in the social networks, Twitter in particular. A few of recent surveys Gayo-Avello [19], Kardara et al. [20] and Riquelme and González-Cantergiani [21] provide comprehensive summarizations on the different techniques regarding identifying influential Twitter users.

In the Twitter social graph, an intuitive way to measure a user's influence is by his number of followers, i.e., the indegree of the vertex representing this user. Although it is suggested Twitter itself is also using the same strategy [22], the metric of indegree isn't always able to reflect the real happening information flowing patterns in Twitter [1, 2, 22, 23] and therefore limited in discovering influence patterns. Dynamic interactions are further exploited to determine influential Twitter users. For example, Kwak et al. [1], one of the earliest effort to quantitatively study the topological characteristics of Twitter's social network, have studied ranking users by the number of retweets and find that it is quite dissimilar with ranking users by the number of their followers. Furthermore, some derivatives of interactions have also been investigated like the normalized or averaged retweets and mentions by total tweets or total followers [2], which might yield a slightly different influence ranking result. Nevertheless, such statistical properties of interaction don't attribute credibility to the phenomenon that a user's influence might be propagated to distant users that are not directly connected to (or interacting with him) on social networks.

On the other hand, there have been some works borrowing PageRank from ordering webpages in the connected World Wide Web Page et al. [24] to ranking users in Twitter directed social network graph [1, 22, 23, 25, 26]. PageRank improves over previous measures that are based directly on simple metrics in the sense that it assumes that by following a user, the followers are implicitly conferring some influence to him and then iteratively propagates a user's influence through the whole social graph. To avoid the limitations of Twitter API in obtaining the follower-following network structure and meantime capture the dynamic information flow between Twitter users, the interactions like *retweet*, *reply*, and *mention* have been utilized trying to build similar social graphs [3–12]. With these graphs, the iterative influence propagation schema such as PageRank can be applied. For example, MultiRank [4] builds different graphs for different interactions such as *retweet* and *reply* respectively, in a given topic.

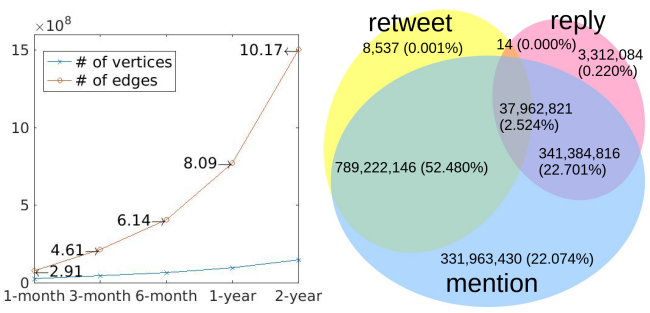
Another strategy of finding influential users in social networks is by *influence maximization*, which is to select k users to maximize the expected number of users being influenced [27]. Location-aware influence maximization methods are also proposed such as Bouros et al. [28] and Li et al. [29], to find top k influential users in a geographical region. Although they have a similar problem context to us regarding identifying local influential Twitter users, our work differs from them by addressing people's geographical proximity (distance) during propagating each other's influence through the social network. Moreover, our algorithms inherently bring ranking orders to all the users by running only once, which is beneficial over their work for queries with varying value of k.

With regard to incorporating geographical proximity between graph vertices into PageRank, the work of Chin and Wen [30] is the most related to ours. They solve a different problem to capture spatial concentration of population movement by running on a geospatial network, where each vertex represents a unique geographical place and edges form between places within reachability in a given travel time.

3 BUILD AN INTERACTION GRAPH $\mathcal{G}$

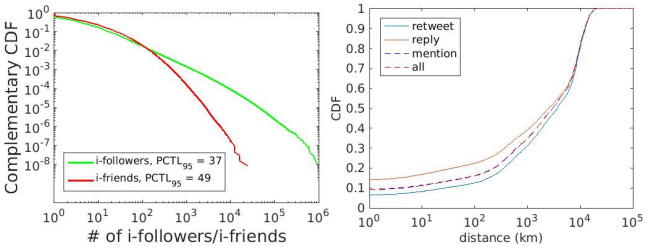
3.1 Dataset

Like the directed follower-following relationships, each Twitter interaction between users has an underlying direction too, pointing from the user who actively initials the action of *rewtweet*, *reply*



(a) (b)

Figure 1: (a) – The number of vertices and edges of interaction graphs built by using 1, 3, 6, 12 and 24-month of tweets. (b) – Venn Diagram of edges in $\mathcal{G}$ by *retweet*, *reply* and *mention*, respectively.



(a) (b)

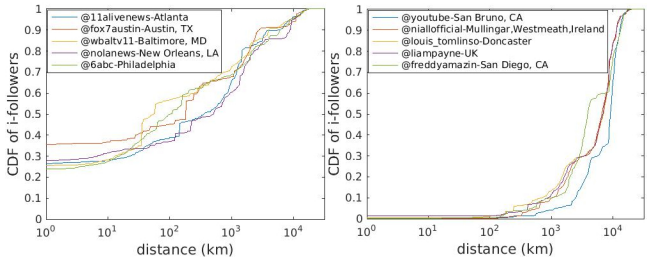
Figure 2: (a) – Distributions of the i-follower/i-friends a Twitter user has. (b) – Distribution of the distances of edges

or *mention* to the other one. Therefore, during building up the interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a directed edge e_{ij} from user v_i to user v_j is constructed and added to $\mathcal{E}$ if there exists at least one interaction pointing from v_i to v_j, both of whom are also added to $\mathcal{V}$ as vertices. For convenience, in a directed *edge* e_{ij} of $\mathcal{G}$, we call v_i one of v_j's *i-followers*, i.e., *interaction followers*, and vice versa, v_j one of v_i's i-friends, i.e., *interaction friends*.

Our dataset consists of 5, 515, 214, 722 tweets collected between January 2015 and December 2016. In these tweets, there are 1, 097, 055, 845 *retweet* interactions, 587, 550, 806 *reply* interactions and 2, 147, 483, 647 *mention* interactions. We therefore build an interaction graph $\mathcal{G}$ of 1, 503, 853, 848 directed edges and 147, 842, 352 users as vertices. $\mathcal{G}$ is relatively sparse in comparison to the one reported in [1]–a complete Twitter social network by July 2009, though $\mathcal{G}$ is on the similar scale in terms of the number of edges. For example, the ratio of number of edges over the number of vertices in $\mathcal{G}$ is 10.17 while the one in [1] has a ratio of 35.25 with a total number of 1, 468, 365, 182 edges. Although collecting more tweets for longer time will increase the ratio, such increase is at a very slow speed as showed in Figure 1a.

Regarding the contribution to building edges in $\mathcal{G}$, as showed in Venn diagram Figure 1b, *mention* is the most significant by covering 99.779% of edges, while *retweet* 55.005% and *reply* only 25.445%. The Venn diagram also shows that *mention* is covering most of the edges constructed from *retweet* and *reply*, indicating that most of the users who retweet or reply each other also mention each other. This, however, is not the case between *retweet* and *reply*, who only share 2.524% edges in common.

Regarding the distributions of indegree/outdegrees in the interaction graph $\mathcal{G}$, we plot the complementary cumulative distributions of the number of i-followers/i-friends each Twitter user has in Figure 2a, which shows a power-law pattern.



(a) (b)

Figure 3: (a) – CDFs of i-followers for 5 local news agencies over their geographical distance. (b) – CDFs of i-followers for the top 5 users (who are selected by maximum indegrees in $\mathcal{G}$).

3.2 Twitter User Locations

In Twitter, there are two sources to know a user's location: the geographical coordinates in his GPS-tagged tweets and the home location in his profile – also termed as *profile-location*. The profile-location is often in the form of place names like "College Park, MD" and can be aligned with databases like GeoNames[1] to decode its geographical latitude/longitude coordinates [31, 32]. In order to assign a unique pair of latitude/longitude coordinates, for users who has multiple pairs of geographical coordinates, we compute the *L1-multivariate median* which essentially finds a point having the minimum sum of distances to a given set of points Z [33]:

$$\underset{z'}{\operatorname{argmin}} \sum_{z \in Z} distance(z', z) \tag{1}$$

After discarding coordinates of (0.0, 0.0) that are likely caused by GPS malfunction, we then have 54, 428, 031 (36.8%) Twitter users having geographical coordinates, and 4, 933, 524 of them from GPS-tagged tweets. Correspondingly, 625, 186, 580 (41.6%) edges of $\mathcal{G}$ have both their vertices with geographical locations.

Next, we use Twitter users/lookup API to download the profile information for Twitter users whose profile-location have not been exposed in our dataset. These users are usually the ones who had appeared in our dataset but were only being replied to or mentioned by others and thereby lacking the profile location information. After downloading profiles for these users, we get the locations of additional 12, 018, 353 users, making a total 66, 446, 384 (44.9%) users have geographical locations. This makes 756, 737, 542 (50.3%) edges in $\mathcal{G}$ have both of their vertices with geographical locations.

Furthermore, there have been methods proposed trying to estimate locations for Twitter users whose locations are unknown such as [33], we therefore investigate the effect of utilizing such a geotagging procedure in Section 5.5.

Geographical Distribution of Edge Distances In this paper, ***the distance of an edge in*** $\mathcal{G}$ refers to the geographical distance between its two Twitter users. For the edges both of whose two vertices have geographical coordinates to calculate a distance, we plot their distance distribution in Figure 2b, which shows that interactions happen over various distances and not always over shorter distances and should receive different geographical considerations.

4 MEASURING SPATIAL INFLUENCE IN $\mathcal{G}$

4.1 Observation and Motivation

The objective of measuring the spatial influence of Twitter users is to find, given a query location Q, which Twitter users have great influence on it. An intuitive solution to this problem is to append a post-processing location filter step after finding influential people

[1] https://www.geonames.org

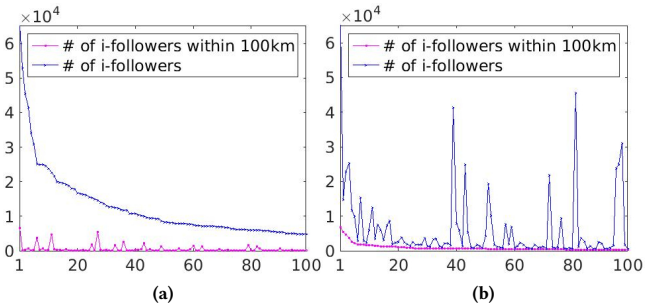


Figure 4: (a) – Top 100 Twitter users in Boston sorted by the number of i-followers. (b) – Top 100 Twitter users in Boston sorted by the number of i-followers within 100 km.

in general. For example, one can use the indegrees of vertices in $\mathcal{G}$ or apply the original PageRank schema (i.e., without giving geographical considerations) to yield a ranking order for Twitter users regarding their influence, and then select the ones who fall within Q and identify them as influential Twitter users on Q.

Such strategy, however, neglects a few important observations.

Regarding Local Influential Twitter Users: First, by utilizing a location filter, the above strategy assumes that, for a Twitter user in Q, his general influence equals to his local influence. This, however, is not necessarily true. For example, as showed in Figure 4a, a Twitter user from location Q who have many i-followers is not guaranteed to have many local people interacting with them, and *vice versa* in Figure 4b. A similar observation is also reported in [34] but on a scale of country-level.

Second, as shown in Figure 3a, for local influential Twitter users, their interaction followers are more aggregated around the local place, and the number of their followers decreases over longer distance, which implies that their influence is more revealed at local places and decays over longer distance. This motivates us to gives different geographical considerations to edges with different distance using a distance-decay function.

Regarding Generalized Influential Twitter Users: A user from a place (or even without specifying his home location) other than the query location Q might have a considerable amount of i-followers from Q and thereby have a chance to exhibit non-negligible and even noticeable influence on Q. For example, even though the Twitter user "@Youtube" is from thousands of miles away from Boston, the number of his i-followers from Boston is larger than any other Twitter users that are in Boston. Another example is "@Patriots" who is a Boston-based Twitter user account, but we found in our dataset, thousands of people from Bristol, CT retweeting, replying and mentioning this user even though those two cities are almost 200km away. Applying a location filter to only keep the Twitter users who are within a limited geographical range to the query location center l_q is likely going to miss such users, and therefore is not suitable for determining the generalized influential Twitter users. This inspires us to alternatively measure a Twitter user's spatial influence on Q based on the spatial locality of his i-followers with respect to the query location center l_q and thereby capture Twitter users of great influence on Q without requiring them to have to be from Q.

In the following subsections, we first give a brief description of the PageRank algorithm. Next, we present our 2 instances of PageRank that address the above two observations, respectively. At last, a hybrid method is proposed trying to combine the 2 instances of PageRank together using distance-decay functions, along with a location query specific teleportation vector, which determines the initial influence values assigned to each vertex.

4.2 PageRank Overview

The mechanism behind PageRank can be briefly explained by an intuitive random surfer model on a given graph where this surfer visits a vertex with a certain probability and follows an outbound edge at random to visit next vertex. The influence of each vertex is then coded in the probability for the random surfer reaching that vertex, calculated as the sum of probabilities of the surfer following all possible edges towards to that vertex.

Additionally, PageRank defines a damping factor h which controls the probability that the random surfer, before starting visiting next vertex, chooses to follow an edge in the given graph to reach next vertex or simply teleport to one which is not connected by edge with the previous vertex. This damping factor is used to avoid the random surfer being trapped in some disconnected components (if exist) in the directed graph and guarantees the convergence of PageRank. In summary, suppose we have a directed graph $G = (V, E)$ in which V is the vertex set and E is the directed edge set, the PageRank procedure can be iteratively defined as follows:

$$\mathbb{R}^{t+1} = (1-h) * \mathbf{\Pi} + h * \mathbb{R}^t \times \mathbb{M} \tag{2}$$

where $\mathbb{R}^t = \begin{bmatrix} r_1^t & r_2^t & \cdots & r_N^t \end{bmatrix}$ is the ranking result after iterating t times, N is the number of vertices $N = |V|$, and each element r_i^t represents the PageRank score of the vertex v_i; h is the damping factor ranging from 0 to 1; $\mathbf{\Pi} = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_N \end{bmatrix}$ is a teleportation vector in which each element π_i denotes the probability that the surfer teleports to the vertex v_i from any other vertices; and $\mathbb{M}$ is the transition probability matrix which is a $N \times N$ matrix with each element m_{ij} specifying the probability that the surfer transits to vertex v_j from vertex v_i by following an existing directed edge in the graph. In the typical PageRank algorithm, the teleportation probability to each vertex is identical by setting $\mathbf{\Pi} = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix}$, and the transition probability m_{ij} is 0 if vertex v_i doesn't have a outbound edge to vertex v_j, and $m_{ij} = \frac{1}{|OUT_i|}$ if such an edge exists, where OUT_i denotes the set of vertices to which v_i has an outbound edge. For simplicity, we use the lowercase script out_j (or in_j) to denote the cardinality of the set OUT_j (or IN_j which denotes the set of vertices from whom vertex v_j has an inbound edge).

Transition Probability in Weighted Graph Generally, given an weighted graph, for example, $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ where $w_{ij} \in \mathcal{W}$ denotes the weight of an edge $e_{ij} \in \mathcal{E}$, the transition probability from vertex v_i to vertex v_j can be calculated defined as [35]:

$$m_{ij} = \frac{w_{ij}}{\sum\limits_{v_k \in OUT_i} w_{ik}} \tag{3}$$

4.3 Edge-Locality PageRank

Figure 3a shows that, as the representatives of local influential Twitter users, the local news agencies have more followers aggregated around the local place but less and less as geographical distance increases, indicating that their influence might decay over distance. This, therefore, inspires us that in determining local influential Twitter users, one Twitter user transfers more influence to another if they have a shorter geographical distance. We therefore propose to use a distance-decay function [30] to assign edges weights as follows and hence give more geographical considerations to edges

who have shorter distances, e.g,, those who are spatially local.

$$f^{EL} := \frac{1}{(d_{ij}+1)^{\kappa}} * \delta(e_{ij}) \tag{4}$$

where $\delta(e_{ij})$ is a binary checking function that outputs 1 if both of the two vertices v_i and v_j have geographical locations in $\mathcal{G}$, otherwise outputs 0. d_{ij} is the distance between vertex v_i and v_j when $\delta(e_{ij}) = 1$, otherwise, set to 0. Adding 1 to distance is to avoid zero-divisions. The parameter κ is the scale factor of distance decay [30] and determines the degree at which the power-law curve declines. In general, a larger κ yields a steeper curve and more significant effect on distance decay.

With the above weight function f^{EL}, we calculate a Edge-Locality Transition Matrix $\mathbf{M^{EL}}$ using Equation 3. Along with the identical transportation vector Π in Equation 2, we now define Edge-Locality PageRank (**ELPR**) as follows:

$$\mathbb{R}^{t+1}_{ELPR} = (1-h) * \mathbf{\Pi} + h * \mathbb{R}^{t}_{ELPR} \times \mathbf{M^{EL}} \tag{5}$$

The ranking result yielded in *ELPR*, however, is not specific to the query location Q and thereby needs a location filtering post-process to find the Twitter users who are from Q.

Location Filtering Given a ranking list of Twitter users and a query location $Q : (l_q, \epsilon)$, an location filtering step is to output a Q-specific ranking list in which only the Twitter users at a distance of $\leq \epsilon$ to l_q are kept and their relative orders in the original ranking list are also reserved.

4.4 Source-Vertex-Locality PageRank

The preferences for spatially-local edges in the Edge-Locality PageRank (**ELPR**) proposed in previous section might miss some Twitter users who have been *retweeted*, *replied* or *mentioned* by a considerable amount of people from the query location Q, even though they are not from Q. To remedy this, in this section, we propose Source-Vertex-Locality PageRank (**SVLPR**) which addresses if a vertex is spatially local to the query location l_q, defined follows:

Definition 4.1. A vertex $v_i \in \mathcal{V}$ is spatially local to a query location l_q if their $distance(l_i, l_q)$ is within a threshold ϵ. Vertices that don't have a location l_i are not spatially local to l_q.

We then propose the following source-specific weight function $f^{SVL} : e_{ij} \rightarrow w_{ij}$ in $\mathcal{G}$:

$$f^{SVL} := \delta'(v_i, l_q, \epsilon) \tag{6}$$

in which, $\delta'(v_i, l_q, \epsilon)$ is a binary spatial locality checking function that outputs 1 if the v_i is spatially local to l_q, otherwise outputs 0. In other words, Equation 6 essentially removes all edges $e_{i,j}$ where v_i is not in range ϵ of the query location l_q.

A Source-Vertex-Locality Transition Matrix $\mathbf{M^{SVL}}$ is then calculated using this weight function and Equation 3. Since $\mathbf{M^{SVL}}$ is already l_q-specific, we adopt identical transportation vector Π and define Source-Vertex-Locality PageRank (**SVLPR**) as follows:

$$\mathbb{R}^{t+1}_{SVLPR} = (1-h) * \mathbf{\Pi} + h * \mathbb{R}^{t}_{SVLPR} \times \mathbf{M^{SVL}} \tag{7}$$

4.5 Geographical PageRank

Edge-Locality PageRank (*ELPR*) and Source-Vertex-Locality PageRank (*SVLPR*) find influential Twitter users on location Q by addressing two different geographical considerations of $\mathcal{G}$. The former emphasizes on the edges that are formed within a shorter spatial distance, while the latter focuses on the edges whose source vertices fall spatially-locally to the query location center l_q. In this section, we combine these two factors in a Geographical PageRank (**GPR**) algorithm. Specifically, concatenating these two factors in **GPR** is completed by the operation of multiplication via the weight function f^{GEO} defined as:

$$f^{GEO} := \begin{cases} \frac{1}{(d_{ij}+1)^{\kappa}} * \delta(v_i, l_q, \epsilon), & \text{both } v_i, v_j \text{ have locations,} \\ \frac{|IN_j^{l_i,\epsilon}|}{|IN_j|} * \delta(v_i, l_q, \epsilon), & \text{only } v_i \text{ has a location} \\ \frac{1}{(d_{max}+1)^{2\kappa}}, & \text{otherwise.} \end{cases} \tag{8}$$

where d_{ij} is the distance between vertex v_i and v_j, d_{max} is the maximum value of all d_{ij} and used to punish edges both of whose two vertices don't have available locations. And $\delta(v_i, l_q, \epsilon)$ has the same definition with the one in Equation 6.

Recall that IN_j is the set of vertices from whom vertex v_j has an inbound edge, i.e., the set of interaction followers of user v_j. We additionally define $IN_j^{l_i,\epsilon}$, a subset of IN_j in which each vertex is within ϵ distance to the location l_i. When a vertex v_j doesn't have a location label, our intuition of using $\frac{|IN_i^{l_i,\epsilon}|}{|IN_i|}$ as its spatial influence to l_i is out of consideration its likeliness of falling nearby l_i by treating its interaction followers' locations as its potential location distribution.

From Equation 8 and 3, we then calculate a Geographical Transition Matrix $\mathbf{M^{GEO}}$. To give more preferences to Twitter users who have shorter distance to the query location center l_q, we propose the following Q-Specific Teleportation Vector Π^Q to complement the Geographical Transition Matrix $\mathbf{M^{GEO}}$.

Query Location Q-Specific Teleportation Vector For a query location Q, we first define a vertex v_i's spatial relevance to Q as follows:

$$rel(l_q, v_i) = \begin{cases} \frac{1}{\left(distance(l_i, l_q)+1\right)^{\kappa}}, & v_i \text{ has geo-coordinates } l_i \\ \frac{|IN_i^{l_q,\epsilon}|}{|IN_i|}, & \text{otherwise} \end{cases} \tag{9}$$

where $distance(l_i, l_q)$ calculates the geographical distance between the query location l_q and the location l_i of vertex v_i. And the definition of $\frac{|IN_i^{l_q,\epsilon}|}{|IN_i|}$ is similar to the one defined in Equation 8 but now measures the likeness of an unknown-location vertex v_i falling in Q by treating v_i's interaction followers' locations as its potential location distribution. Now, we normalize the spatial relevance to get a vertex's teleportation probability $\pi_i = \frac{rel(l_q, v_i)}{\sum_{v_k \in \mathcal{V}} rel(l_q, v_k)}$ and use $\mathbf{\Pi^Q}$ to denote such a teleportation vector.

Combining the Geographical Transition Matrix $\mathbf{M^{GEO}}$ and the query location Q-Specific Teleportatoin Vector $\mathbf{\Pi}^Q$, we define the Geographical PageRank (**GPR**) as follows,

$$\mathbb{R}^{t+1}_{GPR} = (1-h) * \mathbf{\Pi}^Q + h * \mathbb{R}^{t}_{GPR} \times \mathbf{M^{GEO}} \tag{10}$$

Like *SLVPR*, we don't apply a location filtering post-process on *GPR*, which will miss the generalized influential Twitter users on the query location Q. This distinguishes from *ELPR*, which is not query location specific and thereby runs only once for different Qs, although a location filter is needed.

5 EMPIRICAL EVALUATION

In this section, we first describe the experimental settings including the related baselines approaches, the evaluation methods and default parameters settings. Next, we report the results of measuring spatial influence of Twitter users by different methods regarding 3 cities in USA. Afterwards, we choose the city of Boston, MA to report the comparing results. Furthermore, we study the effects of the interaction's types, along with the effects of applying a geotagging procedure to estimate locations for unknown-location Twitter users, followed by a study on the sensitivity of the distance-decay factor κ in *ELPR* and *GPR*. At last, we discuss the potential applications of using local influential Twitter users as news seeders regarding to local news (event) detection.

5.1 Experimental Settings

5.1.1 ***Baseline Approaches.*** Because the difference in types of influential Twitter users *ELPR* and *SVLPR* are trying to find – the former finds *local influential Twitter users* who are not only having great influence on a location but also from there while the latter find *generalized influential Twitter users* and doesn't have a requirement regarding where they are from, we put them in two different control groups and list their related baseline approaches separately as follows. In addition, we also present the results of the hybrid method *GPR* to investigate its effects of combing the two types of spatial locality defined in Section 4.3 and Section 4.4, respectively. Baseline approaches to *ELPR* (Edge Locality Group):

- *InD*: measures the influence by a user's In-Degree in $\mathcal{G}$, i.e., the number of i-followers a Twitter user has.
- *LocInD*: measures the influence of a user by the number of its i-followers who are within ϵ distance to this user.
- *PR*: i.e., PageRank, measures the influence by a user's score by running PageRank on $\mathcal{G}$.

Baseline approache to *SVLPR* (Source Vertex Locality Group):

- $iFol - l_q$: measures the influence of a user by the number of its i-followers who are within ϵ distance to l_q.

Since *InD*, *LocInD*, *PR* and *ELPR* are not Q-specific, a location filter is applied to only keep the Twitter users within ϵ distance to l_q.

5.1.2 ***Evaluation Methods.*** Choosing the city of Boston, MA, we study two aspects of the ranking algorithms: correlation and effectiveness.

1) Correlation. The correlation is measured by a modification of Kendall's τ [36] used in Kwak et al. [1]. This modification overcomes the the limit in the original Kendall's τ that rankings in comparison must have the same element and allows for comparing only top k elements in each rankings. The correlation ranges from 0 to 1, and a larger value indicates a stronger agreement. In this paper, we only compare the top 100 in each algorithm's ranking result.

2) Effectiveness. It is very difficult to evaluate the effectiveness of rankings in lack of ground-truth. To approach the effectiveness evaluation, for the methods in the group of "Edge Locality", we utilize a set of manually-collected local influential Twitter users in Boston, MA and compute the **average ranking order** in each of the methods; for the methods in the group of "Source Vertex Locality", we calculate **the number of verified Twitter accounts** in the top 100 influentials identified in each of the methods.

Average Ranking Order: We first manually collect 20 locally influential Twitter users accounts from 4 different categories in Boston metropolitan area and list them as follows:

News Agencies – "@wcvb", "@bostondotcom", "@cbsboston", "@7news", "@bostonhearld";

Sports Team – "@redsox", "@celtics", "@nhlbruins", "@thebostonpride", "@bostoncannons";

Government – "@marty_walsh", "@cityofboston", "@bostonpolice", "@bostonfire", "@masddot";

University – "@bu˙tweets", "@harvard", "@mit", "@berkleecollege", "@northeastern";

As describe in the following, the selection of the representative local influential Twitter users is very much completed by using an external authority, i.e., Google Search Engine, and such knowledge is not known a prior. More importantly, the experimental evaluation is not only to identify these local influential users but instead to compare the average ranking order of them.

Collecting Twitter users in first 2 categories are completed by first typing the keywords in Google "Boston local news", and "Boston Sports team" to find top related websites and then locating their officially Twitter accounts on the webpages. We didn't choose the news agency of "Fox 25 Boston" because it changes its Twitter account from "@fox25news" to "@boston25" in April 2017. The Twitter accounts in the category of *Government* are the official accounts of Boston Mayor, Boston Government, Boston Police Department, Boston Fire Department and Massachusetts Department of Transportation, respectively. And the Twitter accounts in the category of *University* are the official accounts of Boston University, Harvard University, Massachusetts Institute of Technology, Berklee College of Music and Northeastern University, respectively.

The order of Twitter users in a ranking starts from 0. The smaller order a Twitter user has, the more influential he is in that ranking. The average ranking order of a set of influential Twitter users in a ranking is the average of the orders of each influential Twitter in that ranking. ***Therefore, a smaller average ranking order indicates a better ranking algorithm.***

Number of Verified Accounts: In Twitter, verified accounts are the ones that have been examined to be authentic by Twitter itself and considered as high-quality Twitter users. The status of verification can be found in the Twitter user's profile information. We therefore propose to check the quality of a ranking algorithm by counting how many verified Twitter accounts in its top 100 elements. ***The more verified accounts a ranking algorithm has in its top*** 100***, the higher quality this ranking algorithm is of.***

Note that in the evaluation, we also report the performance of the "Source Vertex Locality" methods regarding the metric of Average Ranking Order; and *vice versa.*, the performance of the "Edge Locality" methods regarding the metric of Number of Verified Accounts is also given.

5.1.3 ***Default Parameter Setting.*** The default parameters used in our methods and related baseline approaches are set as follows.

- l_q: the query location centers of "Boston, MA", "Bristol, CT" and "Seattle, WA" are set to 42.3584/-71.0598, 41.6812/-72.9407 and 47.6062/-122.332, respectively, using GeoNames database.
- ϵ: the radius ϵ in the query location Q (also the spatial locality threshold in Definition 4.1), is set to 100km, which we think is large enough for majority of the cities.
- h: the damping factor in PageRank is set to 0.85 for the algorithm *PR*, *ELPR*, *SVLPR* and *GPR*.
- κ: the distance-decay factor in *ELPR* and *GPR* are set to 4 in default. The sensitiveness of κ will be reported in Section 5.6.
- PageRank Iterations: 100 for *PR*, *ELPR*, *SVLPR* and *GPR*.
- Distance Unit: the distance is in the unit of ϵ, i.e., 100km. .

Table 1: The top 5 influential Twitter users identified for 3 different cities.

City	Edge Locality				Source Vertex Locality		Hybrid
	InD	*LocInD*	*PR*	*ELPR*	$iFol - l_q$	*SVLPR*	*GPR*
Boston, MA	Patriots	Patriots	Patriots	Patriots	YouTube	YouTube	Patriots
	CrazyFightz	OnlyInBOS	OITNB	BostonGlobe	realDonaldTrump	realDonaldTrump	Youtube
	DrJillStein	BostonGlobe	JohnCena	OnlyInBOS	Patriots	Patriots	BostonGlobe
	Diaryforcrush	RedSox	BostonGlobe	RedSox	GIRLposts	BostonGlobe	OnlyInBOS
	TWICHISTE	stoolpresidente	RedSox	NHLBruins	HillaryClinton	OnlyInBOS	RedSox
Bristol, CT	SportsCenter	SportsCenter	SportsCenter	SportsCenter	YouTube	YouTube	Youtube
	espn	espn	espn	espn	realDonaldTrump	realDonaldTrump	SportsCenter
	ESPNNFL	SmackHighCT	ESPNNFL	SmackHighCT	GIRLposts	GIRLposts	WSHHFANS
	ESPNStatsInfo	ESPNNFL	ivoryella	MikeAndMike	SportsCenter	SportsCenter	realDonaldTrump
	darrenrovell	ESPNStatsInfo	darrenrovell	ESPNStatsInfo	SincerelyTumblr	CauseWereGuys	Patriots
Seattle, WA	amazon	Seahawks	amazon	Seahawks	YouTube	YouTube	Seahawks
	OriginalFunko	Mariners	Starbucks	Mariners	Seahawks	Seahawks	YouTube
	Starbucks	KING5Seattle	Seahawks	SoundersFC	realDonaldTrump	realDonaldTrump	Mariners
	Seahawks	seattletimes	BillGates	seattletimes	HillaryClinton	Mariners	seattletimes
	XSTROLOGY	SoundersFC	Microsoft	KING5Seattle	GIRLposts	DangeRussWilson	SoundersFC

5.2 Top 5 Influential Twitter users identified regarding 3 different cities in US

In this section, we analyze and compare the top 5 Twitter users identified by our methods and the ones by related baseline approaches with regards to 3 cities "Boston, MA", "Bristol, CT" and "Seattle, WA". The results are listed in Table 1, in which the symbol "@" ahead of a Twitter username is omitted for compactness.

We notice that quite a few of the top 5 influential Twitter users listed in Table 1 are related to commercial accounts. This doesn't come at a surprise in the sense that such users usually have more interactions from other Twitter users due to their population and thus would rank at top positions. More examples of influential Twitter users from various walks of life (news media, reporters, sports team, sports player, politicians, musicians etc.) with respect to Boston can be found in the supplement table[2].

But by listing only the top 5 influential Twitter users, Table 1 is able to show that in general, taking into geographical proximity into consideration, our proposed methods yield better results than the baseline approaches. Such difference becomes more significant when the ranking orders (i.e., as the one listed in the table) are taken into account. We were surprised to observe such differences even for only the top 5 users. In the following, we describe the details of such difference observed in different methods.

InD* vs. *LocInD: In general, *InD* might return noise Twitter users. For example, we do not consider the Twitter users "@Diaryforcrush" and "@TWICHISTE" for the city "Boston, MA" and the Twitter user "@XSTROLOGY" for the city of "Seattle, WA" identified by *InD* are of great influence on their cities because they have very few people from their cities to interact with them. Take "@TWICHISTE" for example, out of the 38, 187 i-followers he has, only 10 are within $100km$ to the center of Boston, MA. In contrast, the 5-th local influential Twitter user "@stoolpresidente" identified by *LocInD* only has 11, 754 i-followers, but 2, 356 of them are within $100km$ to the center of Boston, MA. Although "@Diaryforcrush" and "@XSTROLOGY" get the locations from their geotagged tweets, disuse of such type of geographical information is not going to totally eliminating noisy users because of the existence of users like "@TWICHISTE" who indeed has a profile-location as "Boston, MA", and might also miss some important Twitter users like "@Patriots" and "@Mariners", neither of them giving valid profile-locations.

[2] http://www.cs.umd.edu/~hyw/twiinf-supplement-table.pdf

In contrast, by finding Twitter users who have most interaction followers from the local area, *LocInD* gives high quality results. For example, most of the Twitter accounts identified by *LocInD* are officially accounts of either sports teams, or local news agencies or reporters in each of the three cities, with an exception of "@SmackHighCT" in the city of "Bristol, CT", which is a branch account of a social platform SmackHigh. This account usually posts hilarious contents on high school lifestyle and receives lots of "retweets" from almost one thousand of people in "Bristol, CT".

PR* vs. *ELPR: Both *PR* and *ELPR* improve over their indegree counterparts *InD* and *LocInD* by not just considering how many i-followers (or local i-followers) a user has but also the influence of these i-followers. For example, in comparison with *InD*, the Twitter users in *PR* are all official and verified accounts. Similarly, "@NHLBruines" ranked in the top-5 in *ELPR* but not in *LocInD* because all the top-4 users in *ELPR* (or *LocInD*) are i-followers of "@NHLBruines" while only 2 of them are i-followers of "@stoolpresidente" even though "@NHLBruines' has less i-followers from Boston than "@stoolpresidente".

Taking the spatial locality of edges into consideration, *ELPR* generally outputs a different set of top 5 influentials in comparison with *PR* across the 3 cities because it focuses more on the interactions happened geographically within a city-level. Take the city of Boston for example, the top 5 influentials in *ELPR* has an average of 4204.2 people from Boston actively interacting with them, while the ones in in *PR* has only 2900.2. The numbers for the cities of Bristol and Seattle are 1286.8, 1423.8 and 1396.0 and 2371.6, respectively. This means, *ELPR* more effectively finds Twitter users that are locally influential.

iFol*$-l_q$ vs. *SVLPR: Comparing to previous algorithms, $iFol-l_q$ and *SVLPR* find Twitter users who are influential on a place but not necessarily from there. For example, either the profile-locations of "@YouTube", "@realDonaldTrump" or "@HillaryClinton" is specfified as the 3 cities. Another Twitter user "@GIRLposts" doesn't has a profile-location. But this doesn't mean they don't have influence or negligible influence on the 3 cities. For example, for each of the 3 cities, "@YouTube" has the most number of i-followers from that city than any other accounts, even the ones who are at the city.

In comparison with $iFol - l_q$, the portion of the Twitter users who are from the query city identified by *SVLPR* slightly increases due to its additional consideration of link structures.

Furthermore, in these two methods, several Twitter users are found influential across all the three cities such as "@YouTube" and "@realDonaldTrump", indicating that the influence of these Twitter users are not limited on a local place and goes beyond their profile-locations. This corresponds to the distance distributions of such Twitter users to their i-followers as showed in Figure 3.

GPR: Taking both the spatial locality of edges and source vertices into consideration, *GPR* outputs a combination of the top influentials in *ELPR* and in *SVLPR*. The most interesting finding is that "@Patriots", the official Twitter account of a sport team based in Boston, ranked 5th in *GPR* regarding its influence on the city of Bristol, CT. This is because on one side, "@Patriots" has thousands of people from Bristol, CT to interact with it and on the other side, Boston,MA is at a moderate distance of 170km from Bristol, CT.

5.3 Correlation and Effectiveness

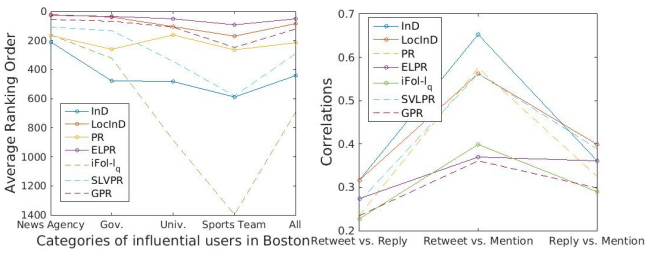
5.3.1 ***Correlation.*** The correlations between the algorithms are listed in Table 2, in which the highest correlation in each row is in bold font. It is clearly that indegree methods are more related to their PageRank counterparts, for example, *InD* vs. *PR*, *LocInD* vs. *ELPR* and $iFol - l_q$ vs. *SVLPR*. In contrast, our proposed methods have lower correlation with the existing metrics *InD* and *PR*, indicating they generate different ranking results to them. This implies identifying spatial influential Twitter users is not a simply procedure of first determining general influence in interaction graph $\mathcal{G}$ by *InD* and *PR* and then applying a location filter post-processing.

In addition, the methods *InD*, *LocInD*, *PR* and *ELPR* have lower correlations to the methods $iFol - l_q$ and *SVLPR* because the former group of methods require that a Twitter user who is influential on a location Q is also from that location, while the latter group of methods don't have such a requirement.

In default, our hybrid method *GPR* is slightly more correlated with *SVLPR* than *ELPR*, indicating that it emphasizes more on the spatial locality of source vertices than the spatial locality of edges and might have more Twitter users who are not from the query location Q in its ranking results as showed in Table 1.

Table 2: Correlation Matrix between different algorithms.

	Edge Locality				Source Vertex Locality		Hybrid
Corr.	*InD*	*LocInD*	*PR*	*ELPR*	$iFol - l_q$	*SVLPR*	*GPR*
InD	1.0	0.35	**0.60**	0.35	0.17	0.17	0.28
LocInD	0.35	1.0	0.36	**0.60**	0.30	0.38	0.30
PR	**0.60**	0.36	1.0	0.40	0.20	0.18	0.29
ELPR	0.35	**0.60**	0.40	1.0	0.30	0.19	0.50
$iFol - l_q$	0.17	0.30	0.20	0.30	1.0	**0.60**	0.53
SVLPR	0.17	0.38	0.18	0.19	**0.60**	1.0	0.56
GPR	0.28	0.30	0.29	0.50	0.53	**0.56**	1.0



(a) (b)

Figure 5: (a) – The average ranking orders of 4 different categories of the local influentials in Boston, MA. (b) – The correlations between different types of interactions.

5.3.2 ***Effectiveness.*** **Average Ranking Order:** Figure 5a shows the average ranking orders of the 4 categories of manually-collected locall influential Twitter users in Boston by different algorithms. Clearly, our method *ELPR* outperforms its baseline approaches *InD*, *LocInD* and *PR*. In addition, *LocInD* outperforms both *InD* and *PR*, justifying the benefits brought by considering the spatial locality of edges in graph $\mathcal{G}$ in determining the spatial influence of Twitter users. Moreover, our another method *SVLPR* that is aware of the spatial locality of source vertices to query location l_q, also achieves better performance than its baseline approach $iFol - l_q$ by additionally taking into account of link structures. At last, our hybrid method *GPR* considering both of the two types of spatial locality in *ELPR* and *SVLPR* has a moderate performance because it introduces popular users like "@YouTube" who are not in Boston, MA.

Number of Verified Accounts: Table 3 list how many verified accounts are there in the top 100 Twitter users identified by different methods. The results show that in the group of "Source Vertex Locality", our proposed method *SVLPR* is slightly better than its baseline approach $iFol - l_q$ because its additional awareness of link structures; and in the group of "Edge Locality", our proposed method *ELPR* clearly outperforms other related methods because in reality, most of the local influential accounts are official accounts of entities like organizations etc and such accounts are usually get verified by Twitter. Our hybrid method, *GPR*, again achieves a moderate performance. This is because, comparing the *ELPR*, it also retrieves Twitter users that are pouplar among people but not necessarily get verified like "@WSHHFANS" because such Twitter users may not represent any organization entities in real world.

Table 3: Number of Verified Twitter users in the top-100

Edge Locality				Source Vertex Locality		Hybrid
InD	*LocInD*	*PR*	*ELPR*	$iFol - l_q$	*SVLPR*	*GPR*
46	76	63	81	55	59	60

5.4 Different Types of Interactions

As showed in Figure 1b, the 3 types of interactions *retweet*, *reply* and *mention* contribute differently in building the edges in the interaction graph $\mathcal{G}$. To investigate how much difference in the influential Twitter users identified by different types of interactions, we run the algorithms on the graphs constructed from only using *retweet*, *reply* and *mention*, respectively and calculate the correlations for *retweet* vs. *reply*, *retweet* vs. *mention* and *reply* vs. *mention*. The results are plotted in Figure 5b, which shows that *retweet* vs. *mention* generally has stronger correlations than *retweet* vs. *mention* and *reply* vs. *mention*. This corresponds to Figure 1b where *retweet* and *mention* have more edges in common than the other two.

5.5 Effects of Geotagging Twitter users

In this section, we study the effects of applying a geotagging procedure to estimate locations for unknown-location Twitter users. We use relative changes (+/-) in Average Ranking Order and relative changes (+/-) in the Number of Verified Accounts to evaluate an algorithm's changes before and after geotgagging. The relative changes of Average Ranking Order is calculated with respect to all the 20 manually-collected Twitter users in Boston in Section 5.1.2.

We estimate location for unknown-location Twitter users by utilizing a Twitter user geotagging procedure [33], which is reported to have the state-of-the-art city-level accuracy. In essence, [33] assigns a location estimation to a Twitter user by using his reciprocal

friends' locations as a set of points in Equation 1 to calculate a median point. After the geotagging, we have 74,846,116 (50.6%) Twitter users assigned with geographical locations, and 1,084,772,048 (72.1%) edges whose two vertices both have geographical locations.

We then again run the different algorithms using the new set of location labels of vertices for $\mathcal{G}$, and list their results in Table 4. For the changes in Average Ranking Order, in the control group of "Edge Locality", methods *InD*, *LocInD* and *PR* get more affected by the geotagging procedure while *ELPR* receives less effects with the Average Ranking Order by only increasing 5.4. The method $iFol-l_q$ is much more susceptible of geotagging than *SVLPR* because $iFol-l_q$ only considers the the location of source vertices and would exhibit larger difference when more Twitter users are geotagged as in Q.

For the changes in the Number of Verified Accounts, all methods yield slight changes before and after geotagging, indicating geotagging unknown-location Twitter users has less effect on the verified official accounts. This is because in most cases such verified official accounts are likely to provide a profile-location, which lets us have their geographical location at hand before geotagging.

Table 4: Effects of Geotagging on Different algorithm

	Edge Locality				SV Locality		Hybrid
	InD	*LocInD*	*PR*	*ELPR*	$iFol-l_q$	*SVLPR*	*GPR*
Avg. Order	477.8	97.6	232.8	56.7	764.9	301.9	107.7
Avg. Order +/-	+38.1	+13.35	+19.15	+5.4	+72.4	+10.75	-13.25
Veri. Accts.	48	75	62	80	52	56	58
Veri. Accts. +/-	+2	+1	+1	-1	-3	-3	-2

5.6 Sensitivity of Distance-Decay Parameter κ

To investigate the sensitivity of κ in *ELPR*, we compare its correlations to *InD* under different values of κ. Similarly, for the sensitivity of κ in *GPR*, we compare its correlation to *ELPR* and *SVLPR* respectively. The results are listed in Table 5.

Table 5 shows that until $\kappa = 2^3$, *ELPR* is becoming more similar to *InD* as the value of κ continues to increase to generate a more significant effect of distance-decay in $\mathbf{M}^{EL}$. This indicates that a larger κ makes *ELPR* more prefer Twitter users having shorter distance to the query location center l_q. The correlation drops at $\kappa = 2^3$ because *LocInD* relies on a location filter and doesn't geographically distinguish the Twitter users within $\epsilon = 100km$ to the query location l_q, while *ELPR* continues preferring to rank higher for those who have even shorter distance to l_q.

Table 5 also shows that *GPR* has more similarities to *SVLPR* across different κ than *ELPR*. In the meantime, as κ increases, *GPR* gives more weights on edges having shorter distances and thereby its correlation with *ELPR* increases. Such trade-off stabilizes after 2^2.

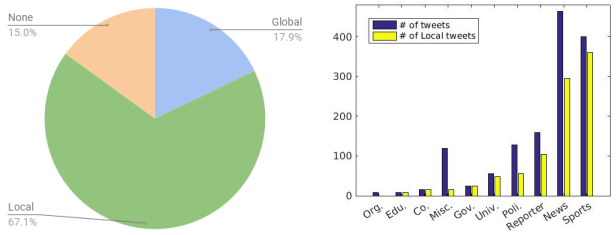
Table 5: Sensitivity of κ in R_{ELPR} and R_{GPR} regarding their correlation with R_{LocInD}, R_{ELPR} and R_{SVLPR}, respectively.

κ	2^{-3}	2^{-2}	2^{-1}	2^0	2^1	2^2	2^3
ELPR vs. *LocInD*	0.01	0.01	0.05	0.33	0.54	0.60	0.57
GPR vs. *ELPR*	0.02	0.02	0.02	0.11	0.39	0.50	0.50
GPR vs. *SVLPR*	0.46	0.46	0.47	0.52	0.57	0.56	0.56

5.7 Application to News Detection

In this section, we explore the potential of local influential Twitter users in acting as news sources (e.g., news seeders [15, 37]) by examining how many of their tweets are about local news (events).

Using the dataset in Section 3, we collected 1,306 tweets posted by the top 70 Boston influential Twitter users identified by the ELPR method (which is given in the supplement table[3]) between Dec 01, 2016 and Dec 07, 2017, and manually categorize these tweets into "Local", "Global" and "None", indicating whether a tweet is about Boston's local news (events), or generally global news or neither of both. The mean and median number of tweets in each user are 18 and 6, respectively. The distribution is presented in Figure 6a, showing that 75% of the tweets are about news, and more importantly, 67.1% are considered local. This supports the viability of using local influential users as potential local news seeders.



(a) (b)

Figure 6: (a) – Ratio of "Local", "Global" and "None" tweets by top 70 Boston influential users in method ELPR. (b) – Number of total tweets vs. number of "Local" tweets for different categories of users.

However, not every local user tweets about the local location. For example, although "@HarvardBiz" is considered influential in Boston, his tweets are mostly reviews on business and technology etc, and are rarely about local news or events. Thus, to investigate which category of Twitter users (the category information of users is provided in supplement table[4] in *tiny* font) are contributing "Local" tweets, for each category of users, we plot the number of their total tweets and the number of their "Local" tweets in Figure 6b, which shows that the users in the categories of *Reporters*, *News* and *Sports* are contributing most of the "Local" tweets and meanwhile maintain a high fraction of "Local" tweets in their own tweets. In addition, most of the tweets posted by users in *University*, *Government* and *Education* are considered "Local", though they have fewer tweets. Therefore, these users might be considered as news seeders, and additional procedures such as classification or topic-sensitive ranking [22] might be exploited in the future to pick out such types of users to improve the quality of news seeders.

6 CONCLUSIONS

This paper focuses on finding spatial influential Twitter users on a query location Q based on the interactions sent out by the local people from Q. The experiments show that by making use of the spatial local edges, our proposed method Edge-Locality PageRank (ELPR) outperforms other related algorithms in finding local influential Twitter users. As local influential Twitter users don't include the ones who are from other places but still have great influence on Q, we furthermore present a method Source-Vertex-Locality PageRank (SVLPR) to find generalized influential Twitter users on the query location Q without requiring them to be from location Q. A hybrid method Geographical PageRank (GPR) taking into account both edge locality and source vertex locality to determine influential Twitter users is also presented. In addition, we also investigate the influence determined by using different types of interactions and also the effects of applying a geotaggging procedure.

There are still many aspects of interactions to explore, such as the frequency and temporal properties [38]. The reciprocity of interactions is also another interesting factor. Also, it is an interesting topic to investigate the typical patterns how user's influence evolves across regions.In addition, *SVLPR* or *GPR* can be modified

[3] http://www.cs.umd.edu/~hyw/twiinf-supplement-table.pdf
[4] http://www.cs.umd.edu/~hyw/twiinf-supplement-table.pdf

to find local influential Twitter users by appending a location filter and therefore to compare with *InD* and *ELPR*. At last, as discussed in Section 5.7, topic-sensitive technologies like LDA might be explored further to identify local influential Twitter users that are in the topic of local news (events).

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? WWW '10.
[2] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM' 10*.
[3] A. Khrabrov and G. Cybenko. Discovering Influence in Communication Networks Using Dynamic Graph Analysis. SocCom '10.
[4] Z. Ding, J. Yan, Z. Bin, and H. Yi. Mining topical influencers based on the multi-relational network in micro-blogging sites. *China Comm.*, 2013.
[5] F. Xiao, T. Noro, and T. Tokuda. Finding News-topic Oriented Influential Twitter Users Based on Topic Related Hashtag Community Detection. *J. Web Eng.*, 2014.
[6] L. A. Overbey, B. Greco, C. Paribello, and T. Jackson. Structure and prominence in Twitter networks centered on contentious politics. *SNAM '13*.
[7] A. E. Cano, S. Mazumdar, and F. Ciravegna. Social influence analysis in microblogging platforms–a topic-sensitive based approach. *Semantic Web*, 2014.
[8] J. Herzig, Y. Mass, and H. Roitman. An Author-reader Influence Model for Detecting Topic-based Influencers in Social Media. HT '14.
[9] G. Katsimpras, D. Vogiatzis, and G. Paliouras. Determining Influential Users with Supervised Random Walks. WWW '15 Companion.
[10] J. Hu, Y. Fang, and A. Godavarthy. Topical Authority Propagation on Microblogs. CIKM '13.
[11] J. Zhang, R. Zhang, J. Sun, Y. Zhang, and C. Zhang. TrueTop: A Sybil-Resilient System for User Influence Measurement on Twitter. *TON*, 2015.
[12] L. B. Jabeur, L. Tamine, and M. Boughanem. Active Microbloggers: Identifying Influencers, Leaders and Discussers in Microblogging Networks. SPIRE '12.
[13] A. Magdy, A. M. Aly, M. F. Mokbel, S. Elnikety, Y. He, S. Nath, and W. G. Aref. GeoTrend: Spatial Trending Queries on Real-time Microblogs. SIGSPATIAL '16.
[14] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and Preference-aware Recommendation Using Sparse Geo-social Networking Data. SIGSPATIAL '12.
[15] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in Tweets. SIGSPATIAL '09.
[16] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of Live News Events Using Twitter. LBSN '11.
[17] F. Chen and D. B. Neill. Non-parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs. KDD '14.
[18] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading News with Maps by Exploiting Spatial Synonyms. *Commun. ACM*, 2014.
[19] D. Gayo-Avello. Nepotistic Relationships in Twitter and their Impact on Rank Prestige Algorithms. *Inf. Process. Manage.*, 2013.
[20] M. Kardara, G. Papadakis, A. Papaoikonomou, K. Tserpes, and T. Varvarigou. Large-scale evaluation framework for local influence theories in Twitter. *Inf. Process. Manage.*, 2015.
[21] F. Riquelme and P. González-Cantergiani. Measuring user influence on Twitter: A survey. *Inf. Process. Manage*, 2016.
[22] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding Topic-sensitive Influential Twitterers. WSDM '10.
[23] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa. TURank: Twitter User Ranking Based on User-tweet Graph Analysis. WISE '10.
[24] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.
[25] T. H. Haveliwala. Topic-sensitive PageRank. WWW '02.
[26] B. Hajian and T. White. Modelling Influence in a Social Network: Metrics and Evaluation. PASSAT '11 / SocialCom '11.
[27] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence Through a Social Network. KDD '03.
[28] P. Bouros, D. Sacharidis, and N. Bikakis. Regionally Influential Users in Location-aware Social Networks. SIGSPATIAL '14.
[29] G. Li, S. Chen, J. Feng, K.-l. Tan, and W.-s. Li. Efficient Location-aware Influence Maximization. SIGMOD '14.
[30] W.-C.-B. Chin and T.-H. Wen. Geographically Modified PageRank Algorithms: Identifying the Spatial Concentration of Human Movement in a Geospatial Network. *PLOS ONE*, 2015.
[31] M. D. Lieberman and H. Samet. Multifaceted Toponym Recognition for Streaming News. SIGIR '11.
[32] M. D. Lieberman and H. Samet. Adaptive Context Features for Toponym Resolution in Streaming News. SIGIR '12.
[33] R. Compton, D. Jurgens, and D. Allen. Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization. *BigData '14*.
[34] I. Anger and C. Kittl. Measuring Influence on Twitter. i-KNOW '11.
[35] W. Xing and A. Ghorbani. Weighted PageRank algorithm. CNSR '04.
[36] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 1938.
[37] N. Gramsky and H. Samet. Seeder Finder: Identifying Additional Needles in the Twitter Haystack. LBSN '13.
[38] M. Franzke, J. Bleicher, and A. Züfle. *Finding Influencers in Temporal Social Networks Using Intervention Analysis*. ADC '16.