Socio-spatial influence maximization in location-based social networks

Mohammad Hosseinpour [a,*], Mohammad Reza Malek [a], Christophe Claramunt [b]

[a] Department of GIS, Faculty of Geodesy and Geomatics Engineering, K.N.Toosi University of Technology, Tehran, Iran
[b] Naval Academy Research Institute, Lanveoc-Poulmic, BP 600, 29240 Brest Naval, France

Keywords:
Location-based social networks
Information diffusion
Line graph
Influence maximization problem
Feature vectors
k-nearest neighbors

A B S T R A C T

Identifying influential nodes in social networks is a key issue in many domains such as sociology, economy, biology, and marketing. A common objective when studying such networks is to find the minimum number of nodes with the highest influence. One might for example, maximize information diffusion in social networks by selecting some appropriate nodes. This is known as the Influence Maximization Problem (IMP). Considering the social aspect, most of the current works are based on the number, intensity, and frequency of node relations. On the spatial side, the maximization problem is denoted as the Location-Aware Influence Maximization Problem (LAIMP). When advertising for a new product, having access to people who have the highest social status and their neighbors are distributed evenly across a given region is often a key issue to deal with. Another valuable issue is to inform the maximum number of users located around an event, denoted as a query point, as quickly as possible. The research presented in this paper, along with a new measure of centrality that both considers network and spatial properties, extends the influence maximization problem to the location-based social networks and denotes it hereafter as the Socio-Spatial Influence Maximization Problem (SSIMP). The focus of this approach is on the neighbor nodes and the concept of line graph as a possible framework to reach and analyze these neighbor nodes. Furthermore, we introduce a series of local and global indexes that take into account both the graph and spatial properties of the nodes in a given network. Moreover, additional semantics are considered in order to represent the distance to a query point as well as the measure of weighted farness. Overall, these indexes act as the components of the feature vectors and using *k*-nearest neighbors, the closest nodes to the 'ideal' node are determined as top-*k* nodes. The node with maximum values for feature vectors is considered as the 'ideal' node. The experimental evaluation shows the performance of the proposed method in determining influential nodes to maximize the socio-spatial influence in location-based social networks.

1. Introduction

The range of research topics in the area of social networks varies from macroscopic to microscopic issues. In the former, the whole structure of the network and its general properties such as small world [1] and scale-free [2] are often evaluated. In the latter, the focus is on the nodes themselves and their respective roles in the network [3]. Identifying influential nodes in a given network is classified into this latter category. Identifying important nodes in a network is likely to give a better understanding and even control over a network. When the objective is to broadcast some information over a given network from some given nodes, it is possible to increase the impact on the other nodes of the network by selecting the ones the most structurally important. In fact, the search for a general index that will reflect the role and importance of some social network nodes is highly dependent on the specific case [4]. However,

* Corresponding author.
E-mail address: mhosseinpour@mail.kntu.ac.ir (M. Hosseinpour).

https://doi.org/10.1016/j.future.2019.06.024

one might consider that a node which has the capability to spread information faster and to a large part of the network is an effective one [5,6]. For instance, when promoting some novel products, the most valuable impacts are often derived from direct recommendations, as people usually trust their relatives. With the advent and development of social networks, this issue has become a crucial one. Existing methods to determine such effective nodes vary from structural measures, either local such as the measure of degree, or global from betweenness centrality or closeness centrality. Another approach, suggested recently, is to apply the concept of the line graph. In graph theory, the line graph of an undirected graph represents the adjacencies between the edges of the initial graph. Line graph has some prominent properties that make it easier to analyze social networks [7–9]. A line graph can retrieve the original network by preserving all its information. In addition, indirect neighbors are considered besides direct neighbors. Furthermore, a line graph of a graph has a higher degree of overlapping. The importance of this issue becomes apparent when searching for overlapped vertices. Finally yet importantly, using a line graph, the problems of resolution limit and the necessity of having prior knowledge about the number of communities, are not relevant anymore.

The objective of the research presented in this paper is to provide a modeling framework together with a series of indexes for identifying nodes that are influential in a location-based social network. Hereafter, influential nodes are those who have a higher social status, and whose "friends" are evenly distributed around an event hereafter denoted as a query point or inside a given region. In fact, we are looking for nodes that have a high social centrality, but along with this, location centrality is also considered. Location centrality means having "friends" that are distributed properly and uniformly across the region or around a query point. The motivation behind the integration of the spatial dimension is, for instance, the need to diffuse some information uniformly and quickly in a specific region or to get feedback from citizens about the quality of the services provided by different organizations within a specified region.

More precisely, we extend the problem of maximizing influence in social networks to location-based social networks, and we call this approach as the Socio-Spatial Influence Maximization Problem (SSIMP). Accordingly, we model a social network as a line graph, and where for each subgraph $k_{1,n}$ $(n \geq 3)$ of the initial graph, there is a complete subgraph k_{n-1} in the line graph. Several indexes are defined to measure the influence level of these subgraphs both locally and globally. They are referred to as the socio-spatial influence index, the query point index and the weighted farness index. Finally, by applying a supervised clustering algorithm through k-nearest neighbors, influential nodes are identified. The outputs of the analysis are evaluated and is shown that how the proposed technique can be applied to identify social and spatial influential nodes. In summary, the main contributions of this paper are as follows:

- Modeling influential nodes in a given network considering both the social structure and the spatial distributions of the neighbor nodes;
- Using dual space and the concept of line graph as the background modeling framework;
- Identifying influential users around a query point or across a given region using feature vectors;
- Evaluating the nodes' influence in a location-based social network using Thiessen polygons.

The rest of the paper is structured as follows. Section 2 briefly reviews related works in the field of social networks and location-based social networks. Section 3 develops the problem definition, while Section 4 introduces the proposed methodology. Section 5 applies the proposed method on two real social networks and Section 6 discusses the results. Finally, Section 7 draws the conclusions and outlines future works.

2. Related works

A social network can be constructed from relational data and can be defined as a set of social entities, such as people, groups, and organizations, with some relations or interactions between them [10]. Methods for analyzing social networks are designed to explore patterns of interaction between social network entities [11]. The focus is on the relationships between nodes, rather than entities themselves. Determining influential nodes in order to maximize the spread of influence in social networks is one of the key issues. The problem of selecting effective nodes for marketing purposes was first introduced by Domingos and Richardson [12], while the influence maximizing problem was defined as an optimization problem by Kempe et al. [13]. The authors provided a simple greedy algorithm that evaluates all the nodes in the network and combination of this method with a Monte Carlo simulation is, however computationally expensive. While a series of follow-up works have suggested additional developments such as a combination with evolutionary algorithms to improve computational times, these approaches are still limited in terms of performance, and strictly applied to conventional social networks [14,15]. Another category of approach has applied heuristic algorithms with the advantage of better running time, but still limited to some approximations. For instance, Jiang et al. [16] developed a simulated annealing algorithm and introduced two new heuristic methods in order to accelerate the convergence while searching for influential nodes. Reyes and Silva [17] search for influential nodes in the form of a maximum coverage problem and presented some new heuristics that take into account the network topology. Li et al. [18] suggested a new formulation of the problem, so-called the Keyword-Based Targeted Influence Maximization (KB-TIM). The objective is a search for a set of candidate nodes that maximize the influence on users who are relevant to some advertisements. The authors used a sampling technique based on a weighted reverse influence set and achieved an approximation ratio. However, and as for the previous optimization-based algorithms, most of these works are oriented to conventional social networks, without further consideration of additional geographical dimensions. Indeed, these algorithms can be applied to many application areas, but they are not appropriate for geographical contexts where social networks can be inferred.

On the other hand, the widespread use of smart devices and location-based services has created a new concept of social media so-called location-based social networking. A location-based social network can be roughly defined as a social network that is closely related to a geographical context and where nodes and links are located in space. Location-based social networks generally regroup a set of entities that share some relationships and experiences, and can offer services and opportunities to the users involved. Indeed the issue of influence maximization in such networks is also a key issue to study, in order to analyze the most influential nodes, but indeed the difference being here to not only analyze the network underlying structure, but also the influence of space, topology, and distances between the entities involved. As for conventional social networks, greedy algorithms have been developed to find influential users in location-based social networks. For example, Li et al. [19] developed two greedy algorithms that increase the speed of diffusion in location-based social networks. Bhosale and Kulkarni [20] have also used a community-based greedy algorithm for mining top-k influential nodes. This algorithm has two components: dividing the network into clusters by taking into account the information diffusion,

and then finding the influential nodes in each cluster by dynamic programming.

Beside greedy algorithms, the tree structure is also have been used to identify target users. Li et al. [21] introduced a PR-tree index structure and developed a community-based seed selection algorithm, which frequently selects users with the most adverse influences in their communities using offline indexes. Recently, Su et al. [22] devised a TR-tree index structure where each tree node stores users' topic and geographical preferences. By traversing the TR-tree in depth-first order, targeted users are determined.

However, one of the works most related to ours is by Wang et al. [23], who modeled influential nodes near a particular land use, such as new restaurants and introduced a distance — aware influence maximization model, which integrates two influence factors such as spread and users' distance to some given locations. Another related approach is developed by Bouros et al. [24] whose main goal was to find influential people in a particular geographic area. In order to do this, the initial regional influence of each user is obtained by assigning weights to the edges and calculating the network distance between the users.

Overall, and while the subject of the papers discussed here is mostly about the Location Promotion Problem, which is to select a small set of seed users who can lure other users to the target location well, the location of neighbor nodes is hardly considered. On the other hand, evenly distribution of the neighbor nodes in a given region or around a query point is an issue required for applications such as an incident or for advertising purposes. In fact, the spatial distribution of such neighbor nodes in the graph plays an important role in many information diffusion contexts. This is the challenge we address in this paper.

3. Problem definition

Let us consider a location-based social network as an undirected graph $G = (V, E)$ comprising of n nodes and e edges in a geographic region of R. Each pair of nodes $u, v \in V$ is connected by an edge $(u, v) \in E$ if they have a direct relationship. Every node $v \in V$ has a specific location l_v and a set of immediate neighbors, N_v.

Definition 1 (*Influence Area*)**.** The influence area of a node $v \in V$ in a location-based social network denoted as IA_v is the geographic area affected by that node and is defined using Thiessen polygons. In other words, the area of Thiessen polygon covering the node v is considered as the influence area of that node. Let X be a metric space with a distance d. The influence area associated with the node v is the set of all points in X whose distance to node v is lower than their distance to the other nodes in V.

$$IA_v = \{x \in X \mid d(x, v) \leq d(x, u)\}, v, u \in V \tag{1}$$

Definition 2 (*Geographic Coverage*)**.** The Geographic Coverage GC_v of a node $v \in V$ is given by the geographical area to which its immediate neighbors are distributed. It is given as follows where N_v denotes the neighbor set of v, and IA_u is the area covered by a neighbor node u:

$$GC_v = \sum_{i=1}^{n} IA_u, u \in N_v \tag{2}$$

Definition 3 (*Influence Area Index*)**.** the influence area index of a node $v \in V$, denoted as $I_{ia}(v)$, is defined as the aggregate influence area of its neighbor set, N_v, divided by the total area of the extent of the whole location-based social network considered A_R.

$$I_{ia}(v) = GC_v / A_R, v \in V \tag{3}$$

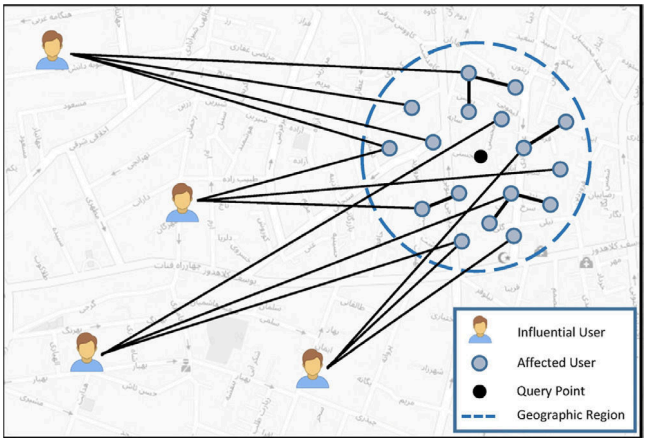

Fig. 1. (Color online) Graphical representation of the problem definition.

Definition 4 (*Influential User*)**.** A node is considered as an influential user in a location-based social network if it meets the following two conditions:

1. The greater number of its direct or indirect neighbors is located inside the query region or around a query point compared to other nodes. Indirect neighbor of a given node is considered as the one not directly connected to that node, that is, located at a graph distance equal or higher than two from that node. Conversely, direct neighbor of a node is the one directly connected to that node.

2. These neighbors are evenly distributed within the query region or around the query point, that is, they have a greater coverage area.

Influential node is determined in such a way that it has the highest number of followers inside a query region or around a query point.

Problem Statement. The influence maximization problem as applied to a conventional social network is to find a minimum subset of nodes so that the information diffusion provided by this subset has the most expected influence on the network. The objective of the present paper is to extend the influence maximization problem for location-based social networks and considers it as a socio-spatial influence maximization problem. Given a location-based social network $G = (V, E)$ and a query point q, the problem is to find a set $k \subseteq V$ so that the extent of the geographic coverage of their neighbors around that query point is maximal. In other words, the total Influence Area Index of subset k is the highest over any other arbitrary subset with the same number of nodes:

$$\forall f \subseteq V | \, n(f) = n(k), \sum I_{ia}(f) \leq \sum I_{ia}(k) \tag{4}$$

Fig. 1 depicts a graphical representation of the discussed problem. The context we assume is to diffuse information within a specified geospatial region or around a query point as soon as possible. We are looking for users having an influence over more possible number of users, which are located inside the query region or around the query point.

We consider the following assumptions in the process of information diffusion:

1. Networks are considered either directed or undirected. Each node can affect the other connected nodes so a node may be affected by two or more neighboring nodes.

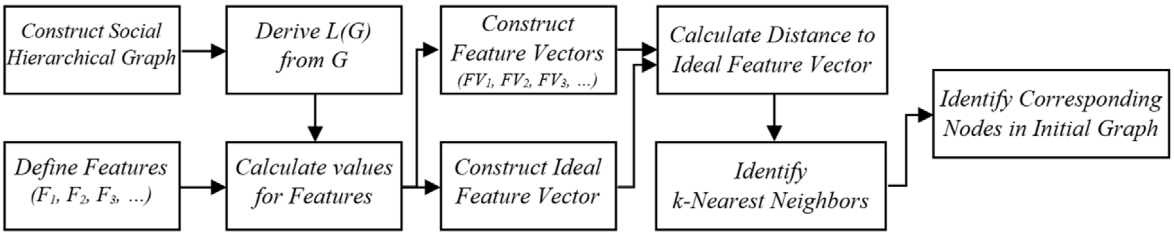

Fig. 2. Overview of the proposed method.

2. The cost of activating primary nodes for all nodes is considered the same. If the costs of activating or convincing the primary nodes are different, the cost minimization issue should be considered from the outset. That is, the initial nodes are chosen in a way to have the highest influence at the lowest cost.

3. The probability of affecting each neighbor node follows the Independent Cascade Model (ICM) [25]. Under the ICM model, time unfolds in discrete steps. At any time-step t, each newly activated node $u \in V$ gets one independent attempt to activate each of its outgoing neighbors with a probability function.

4. Affected nodes are considered as active nodes and other nodes as inactive, and each affected node will remain active until the end of the diffusion process.

4. Proposed methodology

A general solution to the socio-spatial influence maximization problem is to determine effective features and evaluate nodes' performance with respect to these features. The type and the number of these features vary according to the application and the type of the used network. Using these features as feature vector components, the status of each node could be drawn in an n-dimensional space where n is the number of specified features. Next, by introducing the best possible values for these features, that is, the so-called 'ideal' feature vector, the Euclidean distance between each vector and the 'ideal' vector in this n-dimensional space can be derived. The lower the Euclidean distance, the greater the probability of that node to fall into the set of top-k nodes. As a peculiarity of this work, all these steps and computations are carried out under the line graph instead of the initial graph. This reflects the interest of the line graph, that is, the fact that in such applications, relationships between nodes are more important than nodes themselves. Accordingly, the nodes are replaced with the complete subgraphs. In addition, converting the initial graph to line graph removes leaf nodes and nodes with degree two, those being hardly classified as influential users in the large networks. This has the advantage of reducing the computation load. Finally, by performing reverse operations, the nodes associated with these top-k subgraphs are identified in the initial graph.

The first step of the proposed methodology is to model the propagation of influence between the users by the Social Hierarchical Graph (SHG). Three major features named socio-spatial influence index, query point index and weighted farness index are also defined. By applying the line graph to the SHG, features and computations are transferred to the new mathematical space. At the next step, the ideal feature vector is derived and comparing the feature vectors with the ideal vector leads to the top-k subgraph in the line graph and correspondingly top-k nodes in the initial graph. Fig. 2 gives an overview of the proposed method.

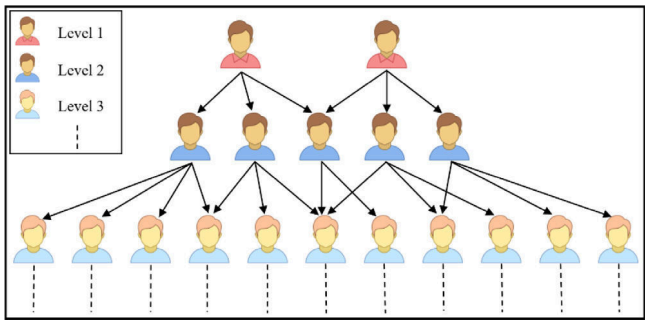

Fig. 3. (Color online) Constructing the social hierarchical graph based on the direction of links between the users.

4.1. Social hierarchical graph

Directed user relations in a social network graph are crucial to model their respective influence. In many applications, followers are often affected by opinion leaders. This indeed stresses the role of direct relations in assessing the influence propagation in a given social network. A social hierarchical graph is a structure that can nicely take into account directed links and the underlying hierarchical structure among the users of a network. This graph also has the advantage of easily identifying and eliminating the users who do not have any follower or connection within the query region or around the query point and hence reduce the computational load. Fig. 3 schematically shows the structure of a sample social hierarchical graph with different levels of users.

4.2. Line graph

An appropriate method to have access directly to neighbor nodes, and to assess their interactions with each other, is given by the line graph. By converting the initial graph to the line graph, each node with degree 3 or higher will be converted to a complete subgraph (Fig. 4), this being not the case for leaf nodes or nodes with degree two so these nodes will be eliminated from the calculations and hence reduce the computational load. By doing this conversion, neighbor nodes, which now are organized as complete subgraphs are directly processed. First, by defining a social and spatial index, the status of these subgraphs is examined locally. Next, the status of the complete subgraphs must be evaluated relative to each other and to a query point. Therefore, by defining a weighted farness index and a query point index, the distance between subgraphs and their distance to the query point are calculated, respectively. Considering these indexes as elements of feature vectors and then clustering subgraphs using the k-nearest neighbors method, the final seed set is identified.

Definition 5 (*Line Graph*)**.** The line graph of a graph G is the graph $L(G)$ that represents the adjacencies between the edges of G.

The line graph of $G = (V, E)$, has the following properties [26]:

1. For each edge in graph G, there is a node in $L(G)$ such that $|n_L| = |e_G|$

2. For each path of length 2 in G, there exists an edge in $L(G)$

The number of edges in the line graph of G, is $e' = \frac{1}{2}\sum_{v\epsilon V(G)} d(v)^2 - e$ and for each subgraph $k_{1,n}(n \geq 3)$ of the G there is a complete subgraph k_{n-1} in $L(G)$.

In a line graph, each node represents an edge in the initial graph, and each edge corresponds to a pair of connected edges. The common node between the connected pair of edges is the intermediate node. Therefore, by switching to the line graph, these intermediate nodes are removed and direct links are established between the neighbors.

In order to analyze the neighbor nodes, a series of local and global parameters have been defined. The local parameter represents the status of the neighbor nodes using node degree and the standard distance between the neighbors, both socially and spatially. These two subparameters are referred to as Social Influence Index and Spatial Influence Index respectively. Combining these two indexes provides a single index called Socio-Spatial Influence Index, which is a local measure for assessing the status of the neighbors.

4.3. Defining features

In order to identify the top-k nodes so that their neighbors have the best spatial distribution around a query point or across a given region, these must meet the simultaneous following conditions: (1) these nodes must have high social and spatial influences; (2) their neighboring nodes must have lower spatial distribution towards a query point; and (3) the spatial correlation between the neighbor nodes of top-k nodes must be minimized.

4.3.1. Social influence index

The influence of individuals on one another in social networks depends on a variety of parameters, including the type, intensity, and frequency of the relationships between them. In order to measure the influence of some individuals on a given social network, several models have been so far suggested, e.g., ICM. Under the ICM model, time unfolds in discrete steps. At any time-step t, each newly activated node $u \in V$ gets one independent attempt to activate each of its outgoing neighbors with a probability $p(u, v) = W(u, v)$. In other words, $W(u, v)$ denotes the probability of u influencing v [27]. For implementation purposes, the probability of affecting each neighbor node is considered as 1, $(P(u \rightarrow v) = 1)$, which assumes that each affected node can affect certainly its neighbor nodes. Based on the social hierarchical graph, users can activate their neighbors in a cascading mode.

Assume that $N_v = \{u_1, u_2, \ldots, u_n\}$ denotes the set of activated neighbors of the node v. The social influence index for node v is denoted as $I_{so}(v)$ and is defined as the normalized version of its number of affected neighbor nodes:

$$I_{so}(v) = (n_v - \min(n))/(\max(n) - \min(n)),$$
$$v \in V, n \in \mathbb{N}, 0 \leq I_{so}(v) \leq 1 \quad (5)$$

Where n_v denotes the number of affected neighbor nodes of node v while $\min(n)$ and $\max(n)$ denote minimum and maximum number of affected neighbors of nodes in V, respectively.

4.3.2. Spatial influence index

Section 3 first introduces the influence area index that gives the geographical area covered by the neighbor nodes using the Thiessen polygons. Next, the Spatial Influence Index reflects the amount of spatial dispersion between the neighbor nodes. The larger the extent of the spatial distribution of the neighbors of a given node, the higher the spatial influence of that node. Several measures can evaluate the spatial distribution of a set of nodes. The standard distance is one of these measures, which provides a direct measure for the spatial distribution of these nodes. The standard distance measures the degree of concentration or dispersion of a set of nodes relative to the geometric center of those nodes. Scattered and clustered nodes have a larger and smaller standard distance respectively. The standard distance for a node v is given by Eq. (6):

$$SD_v = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} + \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}} \quad (6)$$

In which, n is the number of affected neighbors, (X_i, Y_i) is the location of ith active neighbor and $(\overline{X}, \overline{Y})$ is the geometric center of the nodes included in the set of affected neighbors and is calculated as:

$$(\overline{X}, \overline{Y}) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \quad (7)$$

Normalizing the standard distance gives the spatial influence index for node v denoted as $I_{sp}(v)$:

$$I_{sp}(v) = (SD_v - \min(SD))/(\max(SD) - \min(SD)) \quad (8)$$

4.3.3. Socio-spatial influence index

One might search for the relation between the number of neighbors and their spatial distribution, and the relationship with the influenceability of a node. In other words, this allows us to compare the respective influence level of two nodes, one with a higher number of neighbors, but lower spatial distribution and the other with a lower number of neighbors but a larger spatial distribution of its neighbors. A linear combination of these two indexes gives a local comprehensive index so-called the socio-spatial influence index and denoted as I_{ss}. I_{ss} establishes an equilibrium between two separate indexes using the α parameter and provides a single quantity to assess the local performance of each influential node, both socially and spatially. Another benefit of the I_{ss} is that it reduces the dimension of the vector features and thus decreases the computational cost. For a node v, I_{ss} is computed through a linear combination as below:

$$I_{ss}(v) = \alpha . I_{so}(v) + (1 - \alpha) . I_{sp}(v) \quad (9)$$

The greater the value of this index for a given node, the more effective that node. α is a parameter that controls the weight of each index and it can vary from 0 to 1 as follows:

- $\alpha = 0$: the final index is merely equal to the spatial index, this leading to the superiority of nodes that only have the spatial effects regardless of social popularity.
- $0 < \alpha < 0.5$: the spatial index is more effective than the social index and, for example, a node with two neighbors that have a greater spatial distribution has priority over a node with three neighbors which have lower spatial distribution.
- $\alpha = 0.5$: both indexes contribute to the final index equally, and spatial influence will be as important as social popularity and vice versa.
- $0.5 < \alpha < 1$: the social index is more effective than the spatial index and, for example, a node with three neighbors which have a smaller spatial influence has priority over a node with two outspread neighbors.
- $\alpha = 1$: the final index is merely equal to the social index, this leading to the superiority of nodes that are only socially popular, regardless of spatial influence.

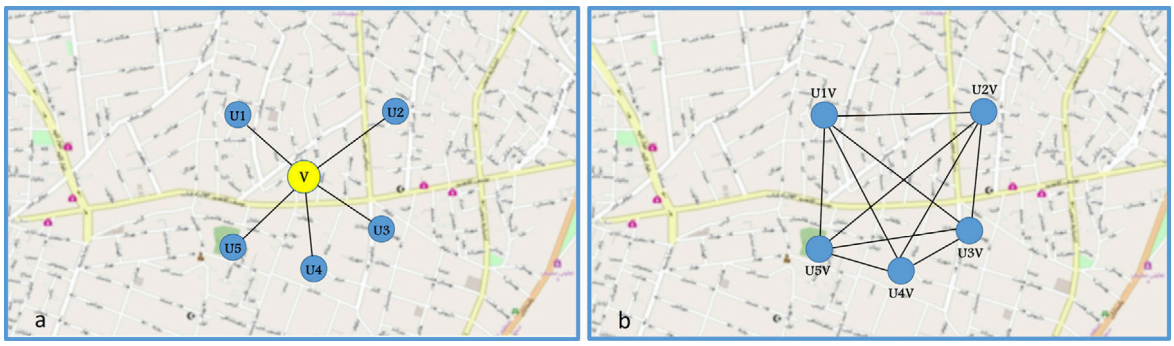

Fig. 4. (a) A sample graph with its direct neighbors and (b) the line graph of the sample graph.

Table 1
Calculating the socio-spatial influence index of two sample nodes for different values of α.

Node	N	N_{min}	N_{max}	SD(m)	SD_{min}	SD_{max}	α	I_{ss}	Rank
U	5	1	10	100	10	500	0.25	0.249	1
							0.75	0.379	2
V	8			30			0.25	0.225	2
							0.75	0.593	1

The α parameter can be used to determine the preference and superiority of each of the indexes. Table 1 shows, how does the value of α affect the prioritization of the nodes shown in Fig. 5. As shown is this figure, in the case of $\alpha = 0.25$, the node U is prioritized to the node V, but by increasing the value of α to 0.75, this priority has been reversed.

After examining the status of nodes using I_{ss}, one can sort them and choose the appropriate k nodes as the effective ones. However, the neighbors of the selected nodes may have high geographic correlation or may be located at a distance far away from the query point. Therefore, it is necessary to assess the status of the neighbor nodes globally. This assessment includes calculating the distance between the nodes and the distance to the query point.

4.3.4. Query point index

Distance to the query point, so-called query point index, is one of the main features in maximizing the spatial influence in a network. In other words, people whose neighboring nodes are closer to the location of an event denoted as a query point will have a higher priority to the information diffusion process. In cases other than the occurrence of an event, and without loss of generality, the purpose of which is to disseminate the influence within a given geographical area, the query point is considered as the geometric center of that area. In order to measure the distance of subgraphs to the query point, the concept of standard distance is applicable, except that the instead of the geometric center of the neighbors, the location of the query point is used. The query point index for the node v denoted as $I_{qp}(v)$ is given by Eq. (10):

$$I_{qp}(v) = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - X_q\right)^2}{n} + \frac{\sum_{i=1}^{n}\left(Y_i - Y_q\right)^2}{n}} \tag{10}$$

Where, n is the number of affected neighbor nodes and (X_q, Y_q) is the coordinates of the query point. In order to control the values obtained for this quantity, the normalized value of this index is defined according to Eq. (11):

$$\hat{I}_{qp}(v) = (I_{qp} - \min\left(I_{qp}\right))/(\max\left(I_{qp}\right) - \min(I_{qp})) \tag{11}$$

4.3.5. Weighted farness index

The spatial correlation between the affected neighbors of nodes should also be taken into account in order to avoid choosing the nodes with overlapping neighbors. In other words, using a global index, the spatial distribution of the neighbors of a node is determined relative to the other ones. Given that, the final nodes are selected in a way that they have the best spatial distribution around a query point. So, one should use a parameter of farness or closeness for nodes. By definition, the closeness is reciprocal of the farness and for large social networks is given as follows:

$$C(x) = N/\sum_{y} d(y, x) \tag{12}$$

Where N is the number of nodes in the graph and $d(y, x)$ is the distance between vertices x and y.

Relying solely on the notion of farness might lead to a selection of nodes that are located at the margin of the considered region. In order to avoid this and to enhance the distance between important nodes, we introduce a notion of Weighted Farness Index denoted as I_{wf} as given by Eq. (13):

$$I_{wf}(v) = \sum_{i=1}^{f-1}(I_{ss}(v) * d((k_n)_v, (k_m)_i)),\ v \in V;\ n, m \in \mathbb{N} \tag{13}$$

Where, k_n represents the subgraph (a node along with its affected neighbors) which, I_{wf} is calculated for and k_m denotes the neighbor subgraphs while f is the number of subgraphs in the line graph.

The distance between two subgraphs is also equal to the sum of distances between their pairwise nodes, which is given by Eq. (14):

$$d((k_n)_v, (k_m)_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij},\ v \in V;\ n, m \in \mathbb{N} \tag{14}$$

In Eq. (13), the I_{ss} is used as the weight parameter. Applying the weights enhances the farness of the influential nodes from each other. The more distant an important node from others, in particular, important ones, the more preference it gets. In order to control this quantity the normalized version of this index is defined as follows:

$$\hat{I}_{wf}(v) = (I_{wf}(v) - \min(I_{wf}))/(\max\left(I_{wf}\right) - \min(I_{wf})) \tag{15}$$

4.4. Selecting top-k subgraphs in $L(G)$

After deriving the local and global indexes required examining the status of top-k complete subgraphs, those with the best values for these three indexes are selected as the final influential subgraphs. In order to do this, the feature vector for each subgraph is drawn based on the results obtained for its features. Feature

Fig. 5. Two sample nodes along with their immediate neighbors: (a) a node with more neighbors and lower spatial distribution; (b) a node with fewer neighbors and higher spatial distribution.

vectors are used to represent the qualitative or quantitative properties of an object mathematically and analytically. A feature vector contains various elements of an object and is represented as a point in the feature space.

The feature vectors are drawn in a 3D space that takes into account the socio-spatial influence index, the query point index and the weighted farness index as their three axes. After projecting subgraphs into this 3D space, the better-performing ones are identified by a clustering technique. The *k*-Nearest Neighbor algorithm is used for this purpose. *kNN* is a supervised clustering technique that is simple but effective in practice. This method assumes that the data is distributed in a metric feature space and then finds the *k*-nearest neighbors to the sample data. Since *kNN* is a supervised method, it needs to have one or more sample data so that it could measure the distance between the input data and the sample data. Here, the sample data is considered as the 'ideal' point that has the best performance for all three features. This performance is related to the type of the query point. If the query point is the location of an incident, then the 'ideal' point will have the highest values for I_{ss} and I_{wf} and the lowest value for I_{qp}. In addition, if the query point is the geometric center of the given region, then the greatest values of all three indexes represent the 'ideal' point. By calculating the Euclidean distance between this 'ideal' point and the other ones, the top-*k* subgraphs are identified.

4.5. Selecting top-k nodes in G

The last step in detecting influential users includes an inverse operation. This is to retrieve the corresponding nodes of top-*k* complete subgraphs by returning to the root graph, which is, converting the line graph to the original graph. Several algorithms can perform this function. Roussopoulos [28] introduced one of such algorithm; its complexity is non-polynomial and is based on an algorithm that searches for the maximal connected common subgraphs in graphs. The Matrix Relabeling Inverse Line Graph Algorithm is another approach proposed by Liu et al. [29], and is more effective than the previous algorithm, but its complexity is also non-linear. Later, the authors of this algorithm introduced the Inverse Line Graph Algorithm (ILIGRA) [30]. The time complexity of this algorithm is linear in the number of nodes in the line graph. This algorithm assumes that the given graph is a line graph and starts to construct the root graph. During the process, this algorithm checks whether the given graph is a line graph or not and stops when it finds the graph is not a line graph. We adopted the ILIGRA algorithm to convert the line graph to its original graph because of its efficiency and decreased computational complexity.

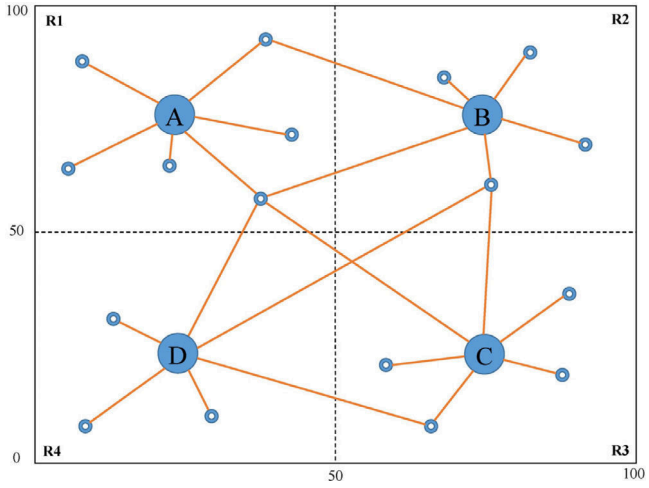

Fig. 6. A sample network with four important nodes *A*, *B*, *C*, *D*, and their neighbors distributed in different sub-regions.

Table 2
Ranking of important nodes shown in Fig. 6 using local and global features.

No	Degree	I_{so}	I_{sp}	I_{ss}	I_{qp}	I_{wf}	$d(v,p)$	Rank
A	8	1	0	0.5	0.51	1	0.49	2
B	6	0.33	0.3	0.32	1	0.18	0.84	3
C	5	0	0.55	0.28	0	0	1.43	4
D	5	0	1	0.5	0.89	0.71	0.31	1

5. Implementation

Prior to larger experimentation, a small network with 21 nodes and 24 edges, illustrated in Fig. 6, is used as a test network. The underlying region is divided into four areas, and in each area, a node with an alphabetical tag is considered as an important node. The geographic distribution for the neighbors of these nodes is different. The neighbors of the node *A* are only distributed in one sub-region, and nodes *B*, *C*, and *D* have neighbors from 2, 3 and 4 sub-regions, respectively. The goal is to identify a node with high social and spatial influence. The geometric center of the area acts as the query point and the 'ideal' node is defined as $p\left(\max\left(I_{ss}\right), \max\left(I_{qp}\right), \max\left(I_{wf}\right)\right)$. The values of indexes for all four important nodes are calculated and presented in Table 2. Finally, these nodes are ranked according to their distance to the 'ideal' node and, as we expected, the node *D* was identified as the most influential node due to the spatial distribution of its neighbors.

Real datasets have been collected from two location-based social networks, Brightkite and Gowalla [31]. Brightkite was once

Table 3
Statistics of the used datasets.

Dataset	$\|V\|$	$\|E\|$	$\#Check-In$	$Ave.Deg$	*Mode*
Brightkite	58,228	214,078	4,491,143	7.4	Undirected
Gowalla	196,591	950,327	6,442,890	9.7	Undirected

a location-based social networking service provider where users shared their locations by checking-in. The friendship network consists of 58,228 nodes and 214,078 edges. Gowalla is another location-based social networking website where users share their location by checking-in. The friendship network is undirected and consists of 196,591 nodes and 950,327 edges. Table 3 shows the general properties of these datasets. They contain information about the locations visited by each user, with a total of about 6.5 million check-ins for the former network and about 4.5 million check-ins for the latter one. Since there is a need to have a specific location for each user, the average locations recorded by the users can be considered as their specific locations.

First, let us estimate the appropriate value of α for these networks. The value of α should be estimated and selected in such a way that top-k nodes in terms of I_{ss} have the highest spatial distribution and social influence. Three subgraphs with 1000, 2000, and 3000 nodes are randomly selected from each network as training data. By changing the value of α from 0 to 1 with step 0.05, the value of socio-spatial influence index, the distance to the query point and the weighted farness index are calculated for all nodes in each subgraph. Regarding the number of training datasets, the value of k is considered as equal to 50. Furthermore, the query point is considered as the geometric center of the entire geographical region. After determining the top-k nodes, proportional to the different values of α in each of the three subgraphs, the standard distance for each k set is calculated to obtain the spatial distribution of the selected nodes. By drawing the values of standard distance versus α, the appropriate value for α is obtained in all three subgraphs. As shown in Fig. 7, the value of α of the Brightkite dataset is estimated as 0.85 and for Gowalla, its value is 0.75. It is also clear from the diagram that by increasing the number of nodes from 1000 to 3000, the spatial distribution of the selected nodes also naturally increases.

After determining the optimal value of α using the training data, in order to ensure the accuracy of these values, it is also necessary to evaluate them using the test data. Accordingly, from each dataset, 5000 nodes are selected as the test data, and with α being known, the top-k nodes are identified. In order to ensure that the selected nodes are optimal, the final nodes are determined using other values of α and again the spatial distribution of the selected nodes is compared with the value of α using a diagram. As expected, for the same values of α obtained in the previous step, the selected nodes had the highest spatial distribution. Fig. 8 shows the results for the test data from Brightkite and Gowalla.

Then top-k nodes are identified for two real datasets after ensuring the values of alpha. For the Brightkite social network, the set of the final nodes has a total social index of 0.06375. This means that of 58,228 nodes included in this network, 3713 of them are as direct neighbors of the nodes in the final k set. The spatial distribution of the neighbor nodes of this k set gives an area of 9,922.63 Km^2. In the case of Gowalla, the total social index for selected nodes is 0.15636 and the number of direct neighbors is 30740 of 196591. The spatial distribution of these nodes is also equal to 8,747.29 Km^2.

6. Evaluation

The objective of this paper and of the experiment is to identify influential nodes whose neighbors have the maximum geographic coverage within a query region or have the best spatial distribution around a query point. As compared to the previous works, the peculiarity of our approach is the two new heuristic methods that assess the efficiency of the suggested algorithm. The first evaluation method relies first on the selection of subsets of size k from the real networks, secondly by using the α value and measuring the spatial distribution and the total social influence index of the neighbors of the selected nodes, and thirdly by comparing it with the characteristics of the top-k nodes of that network. We randomly selected nodes, one hundred times from both datasets and measured the spatial distribution and the total social influence index of their neighbors. The results of these measurements are shown in Fig. 9 for each dataset.

As shown in Fig. 9, the best set of randomly selected nodes for the Brightkite social network has a spatial distribution of 9,289.91 Km^2 and its total social index is 0.06985. While for this network, the top-k nodes obtained in $\alpha = 0.85$ has a higher spatial distribution of 9,922.50 Km^2 and its total social index is 0.06375, which is close to the previous one. In the case of Gowalla, as shown in Fig. 9, the best two sets of nodes selected by random have the highest social indexes with values of 0.152 and 0.15755 and spatial distributions of 3,105.19 Km^2 and 4,266.73 Km^2 respectively. While the top-k nodes obtained for this network with $\alpha = 0.75$ has a total social influence index of 0.15637 and a spatial distribution of 8,747.29 Km^2, which totally has a better performance than the randomly selected sets from this network.

The second method applied for evaluating the efficiency of the proposed algorithm is to compare the total geographical area covered by the neighbors of the top-k nodes obtained using α value with the total area covered by the neighbors of the nodes selected by using other values of α for each dataset. The suggested method for measuring the total area covered by neighbors is to apply weighted Voronoi diagrams [32]. These diagrams are a special case of space gridding in which, every node has a distinct weight. In other words, the distance between the nodes will be a function of their weights. The value of socio-spatial influence index of each node acts as the weight parameter, so the influence area of each node becomes more or less in proportion to this index. Thus, thanks to the weighted case of the Thiessen polygons, the given region is divided between all nodes and it is assumed that each polygon is affected by the diffusion of the influence to its corresponding node. These polygons denote the influence areas of these nodes (see Fig. 10). The total area of the polygons in which the neighboring nodes are located in is also considered as the influence area of that node. Division of this influence area by the total area of R also results in the influence area index for each node.

The total area covered by the neighbors of the randomly selected subsets of size k versus the sum of the social index of these subsets for both Brightkite and Gowalla networks is shown in Fig. 11.

As for the discussion conducted for Fig. 9, the extremums in both networks are also lower in terms of social influence index and the influence area compared to the ones for top-k nodes for these datasets. In the case of the Brightkite dataset, the total influence area index of the top-k nodes is 0.0359 and their total social influence index is calculated as 0.06375. By comparing these indexes with the extremums in Fig. 11, it becomes clear that the top-k nodes selected by using $\alpha = 0.85$ have better performance. For Gowalla, the total influence area index calculated as 0.0348 for top-k nodes selected using $\alpha = 0.75$ and the total social influence index was 0.15637 from the former section. Compared to the extremums of Fig. 11, although the social influence index of the top-k set is lower, it has a higher influence area index, and overall it can be said that the performance of the nodes in this set is also optimal.

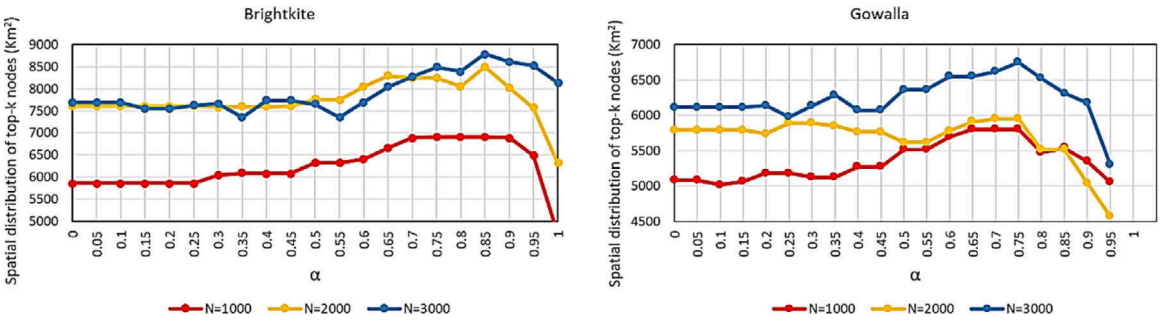

Fig. 7. (Color online) Finding the optimal value of α for Brightkite and Gowalla using the training data.

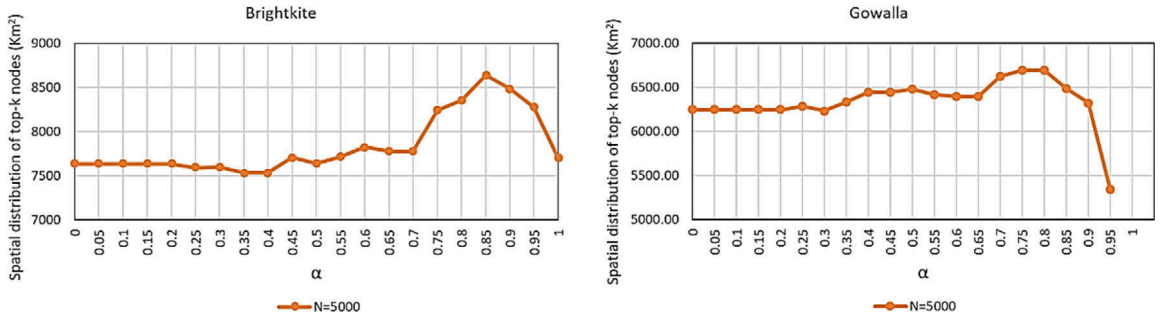

Fig. 8. (Color online) Spatial distribution of the selected nodes compared with different values of α for test data from Brightkite and Gowalla.

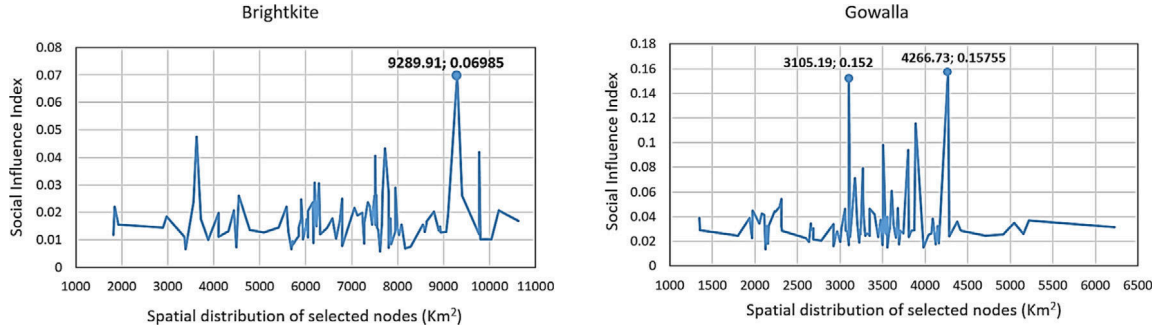

Fig. 9. Comparison of the spatial distribution of the randomly selected nodes' neighbors and their total social influence index for Brightkite and Gowalla.

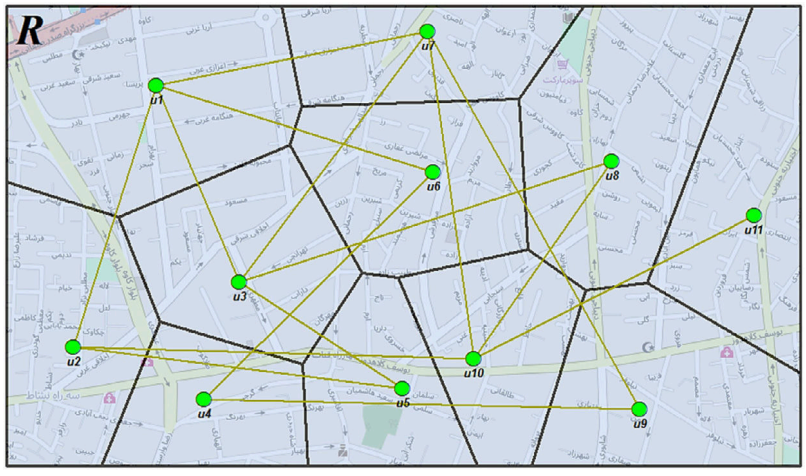

Fig. 10. A sample social network and influence area of each node created with Thiessen polygons.

Since one of the main applications of the proposed method is to consider the area where an event occurs, the diffusion process of information to the most possible geographical regions should be computed and timely evaluated. It appears from our approach that the nodes with higher socio-spatial centrality (SSC), i.e., the influential users as identified by the proposed method can diffuse

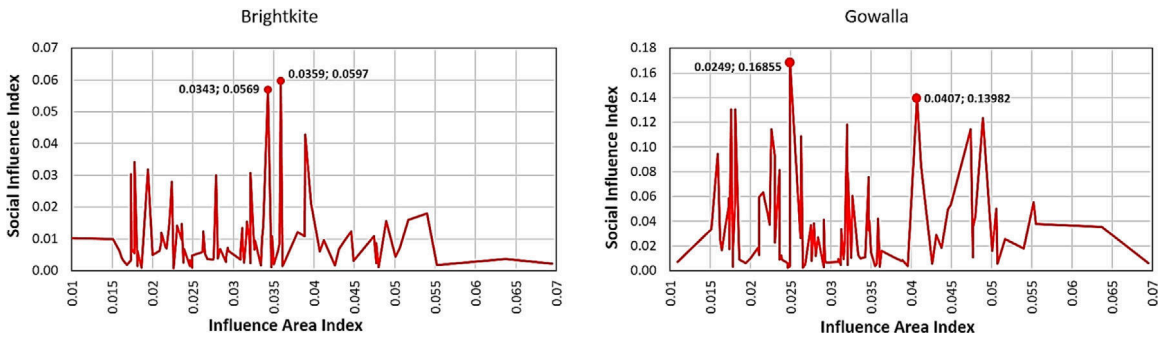

Fig. 11. Changes of the total influence area of neighbors versus the total social influence index for randomly selected subsets of size k of two real datasets.

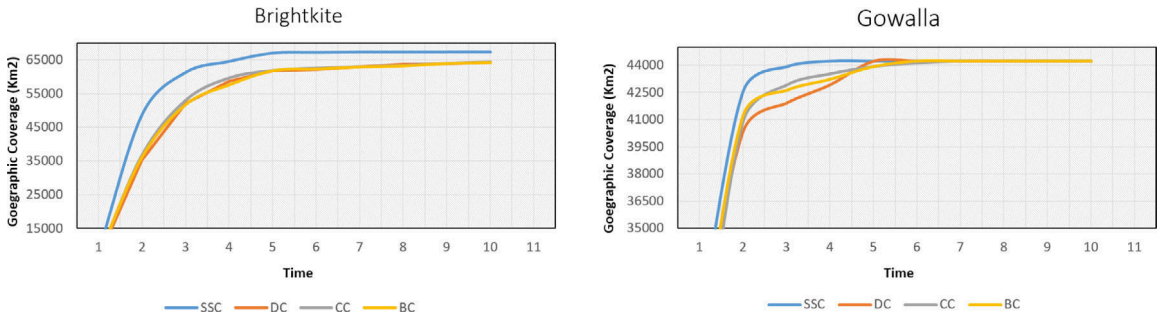

Fig. 12. (Color online) Diagram for the geographic distribution of information over time. Influential users selected by the socio-spatial centrality (SSC) can diffuse information in a vaster area in the first moments of an incident, compared with other centrality metrics like DC, CC, and BC.

information faster and vaster. In order to do so, influential nodes can be conventionally identified using some centrality metrics like degree centrality (DC), betweenness centrality (BC) and closeness centrality (CC). In order to compare these metrics to our own algorithm, ICM first models the spatial distribution of information among the neighbor nodes of influential users and then the SSC method is applied. Computational times are derived from the three centrality measures and the SSC method and are depicted in Fig. 12. The figures show that the SSC method can diffuse information in a vaster area in the first moments of an incident.

7. Conclusion

The research developed in this paper addresses the issue of identifying influential nodes in the location-based social networks. We introduce a model and a series of indexes to resolve the socio-spatial influence maximization problem. Regarding the importance of neighbor nodes in the diffusion of influence, the line graph is considered as an appropriate framework to access and analyze the neighboring nodes directly. The line graph is derived by converting each node with degree 3 and above into a complete subgraph. In order to identify the influential nodes in the location-based social networks, a local social and spatial index and two global indexes have been introduced. By a linear combination and ponderation of social and spatial dimensions, a socio-spatial influence index is derived for all complete subgraphs. Therefore, by considering these values as the subgraph weights, a weighted farness index is defined and calculated for all subgraphs. The distance between the subgraphs and the query point also computed as the query point index. Assuming the values of these three indexes as the components of the feature vector, all subgraphs can be mapped to a 3D space. By defining the notion of 'ideal' point, whose components include the best values of the features, the distance between the subgraphs and the 'ideal' point are calculated. Finally, using the *kNN* algorithm, the top-k subgraphs and their corresponding nodes in the initial graph are chosen as the top-k nodes.

The proposed method has been implemented on two real datasets from location-based social networks, Brightkite and Gowalla. Formerly, three subsets of the main networks are selected randomly as training data and the appropriate value of α is calculated using these sampled data for each real dataset. The results are evaluated in a different subset of data from each network, named the test data. After confirmation of the results, the proposed method is implemented on the real datasets. In order to evaluate the results, two methods are also used. The evaluations show the proper functioning of the selected top-k nodes to meet the needs for high social influenceability and the condition for optimal spatial distribution of the neighbor nodes.

So far, the networks are considered static, and the location of the users in the network is also calculated from the average locations visited by them, but social networks, and in particular, location-based social networks have a dynamic nature and existing connections between nodes as well as the location of individuals are constantly changing. Nodes may also act as influential at a time and lose this feature another time. Therefore, it is necessary to identify the influential nodes immediately over the time using the dynamic modeling of the networks.

Conflict of interest

None.

Declaration of competing interest

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

References

[1] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440.

[2] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[3] F. Bloch, M.O. Jackson, P. Tebaldi, Centrality measures in networks, Available at SSRN 2749124, 2017.
[4] L. Lü, D. Chen, X.L. Ren, Q.M. Zhang, Y.C. Zhang, T. Zhou, Vital nodes identification in complex networks, Phys. Rep. 650 (2016) 1–63.
[5] M. Gong, J. Yan, B. Shen, L. Ma, Q. Cai, Influence maximization in social networks based on discrete particle swarm optimization, Inform. Sci. 367 (2016) 600–614.
[6] N. Kumar, Y. Chandarana, K. Anand, M. Singh, Using social media for word-of-mouth marketing, in: International Conference on Big Data Analytics and Knowledge Discovery, Springer, Cham, 2017.
[7] F. Huang, X. Li, S. Zhang, J. Zhang, J. Chen, Z. Zhai, Overlapping community detection for multimedia social networks, IEEE Trans. Multimed. 19 (8) (2017) 1881–1893.
[8] T. Yamashita, R. Saga, Cluster-based edge bundling based on a line graph, in: VISIGRAPP (3: IVAPP), 2017.
[9] J. Lee, Z.Y. Zhang, J. Lee, B.R. Brooks, Y.Y. Ahn, Inverse resolution limit of partition density and detecting overlapping communities by link-surprise, Sci. Rep. 7 (1) (2017) 12399.
[10] S. Tabassum, F.S. Pereira, S. Fernandes, J. Gama, Social network analysis: an overview, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8 (5) (2018) e1256.
[11] S.P. Borgatti, M.G. Everett, J.C. Johnson, Analyzing Social Networks, Sage, 2018.
[12] P. Domingos, M. Richardson, Mining the network value of customers, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001.
[13] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003.
[14] G. Nandi, U. Sharma, A. Das, A novel hybrid approach for influence maximization in online social networks based on node neighborhoods, in: Advances in Electronics, Communication and Computing, Springer, Singapore, 2018, pp. 509–520.
[15] L. Cui, H. Hu, S. Yu, Q. Yan, Z. Ming, Z. Wen, N. Lu, DDSE: a novel evolutionary algorithm based on degree-descending search strategy for influence maximization in social networks, J. Netw. Comput. Appl. 103 (2018) 119–130.
[16] Q. Jiang, G. Song, C. Gao, Y. Wang, W. Si, K. Xie, Simulated annealing based influence maximization in social networks, in: Twenty-fifth AAAI Conference on Artificial Intelligence, 2011.
[17] P. Reyes, A. Silva, C. de Villarceaux, Maximum coverage and maximum connected covering in social networks with partial topology information, in: IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2014.
[18] Y. Li, D. Zhang, K.L. Tan, Real-time targeted influence maximization for online advertisements, Proc. VLDB Endowment 8 (10) (2015) 1070–1081.
[19] G. Li, S. Chen, J. Feng, K. Tan, W. Li, Efficient location-aware influence maximization, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, 2014.
[20] S. Bhosale, D. Kulkarni, Influence maximization on mobile social network using location based community greedy algorithm, Int. J. Comput. Appl. 122 (19) (2015).
[21] X. Li, X. Cheng, S. Su, C. Sun, Community-based seeds selection algorithm for location aware influence maximization, Neurocomputing 275 (2018) 1601–1613.
[22] S. Su, X. Li, X. Cheng, C. Sun, Location-aware targeted influence maximization in social networks, J. Assoc. Inf. Sci. Technol. 69 (2) (2018) 229–241.
[23] X. Wang, Y. Zhang, W. Zhang, X. Lin, Distance-aware influence maximization in geo-social network, in: ICDE, 2016.
[24] P. Bouros, D. Sacharidis, N. Bikakis, Regionally influential users in location-aware social networks, in: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2014.
[25] J. Goldenberg, B. Libai, E. Muller, Talk of the network: a complex systems look at the underlying process of word-of-mouth, Market. Lett. 12 (3) (2001) 211–223.
[26] J. Saberian, M.R. Malek, S. Winter, A new framework for solving the spatial network problems based on line graphs, Trans. GIS 18 (5) (2014) 767–782.
[27] A. Arora, S. Galhotra, S. Ranu, Debunking the myths of influence maximization: an in-depth benchmarking study, in: Proceedings of the ACM International Conference on Management of Data, ACM, 2017.
[28] N.D. Roussopoulos, A max {*mn*} algorithm for determining the graph H from its line graph G, Inform. Process. Lett. 2 (4) (1973) 108–112.
[29] D. Liu, S. Trajanovski, P. Van Mieghem, Reverse line graph construction: the matrix relabeling algorithm MARINLINGA versus Roussopoulos's algorithm, 2010, arXiv preprint arXiv:1005.0943.
[30] D. Liu, S. Trajanovski, P. Van Mieghem, ILIGRA: an efficient inverse line graph algorithm, J. Math. Model. Algorithms Oper. Res. 14 (1) (2015) 13–33.
[31] E. Cho, S.A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011.
[32] F. Aurenhammer, The one-dimensional weighted Voronoi diagram, Inf. Process. Lett. 22 (3) (1986) 119–123.

Mohammad Hosseinpour received the M.Sc. degree in Geospatial Information Systems (GIS) from K.N.Toosi University of Technology, Tehran, Iran in 2008. Since 2012, he is a Ph.D. student at K.N.Toosi University of Technology. His doctoral research investigates the use of mathematical spaces like dual space to model and analyze location-based social networks. Clustering of the users and identifying the influential nodes in LBSNs using the advantages of line graphs are some examples of the application of his research.

Dr. M.R. Malek is currently an Associate Professor and the head of Ubiquitous and Mobile GIS Research Lab. at the Geodesy and Geomatics Engineering faculty of K.N.Toosi University of Technology. He has more than 250 peer-reviewed journal articles, book chapters, and international conference papers. He serves in the editorial boards of several national and international journals. Some of Dr. Malek's research interests include Ubiquitous and Mobile GIS, Volunteered GI and Location-Based Social Networks, Spatial analysis and, Uncertainty modeling.

Dr. Christophe Claramunt is a Professor in Computer Science and deputy director of the Naval Academy Research Institute in France. He holds a Ph.D. in Computer Science from the University of Burgundy (France). He has been a senior lecturer in computing at the Nottingham Trent University and a senior researcher at the Swiss Federal Institute of Technology in Lausanne. His research is oriented towards theoretical and pluri-disciplinary aspects of geographical information science. He has widely published and currently serves in the editorial boards of several international GIS journals and is regularly involved in the chairing and program committees of several international GIS conferences.