Influence Spread in Geo-Social Networks: A Multiobjective Optimization Perspective

Liang Wang, Zhiwen Yu, *Senior Member, IEEE*, Fei Xiong, Dingqi Yang, Shirui Pan, *Member, IEEE*, and Zheng Yan, *Member, IEEE*

***Abstract*—As an emerging social dynamic system, geo-social network can be used to facilitate viral marketing through the wide spread of targeted advertising. However, unlike traditional influence spread problem, the heterogeneous spatial distribution has to incorporated into geo-social network environment. Moreover, from the perspective of business managers, it is indispensable to balance the tradeoff between the objective of influence spread maximization and objective of promotion cost minimization. Therefore, these two goals need to be seamlessly combined and optimized jointly. In this paper, considering the requirements of real-world applications, we develop a multiobjective optimization-based influence spread framework for geo-social networks, revealing the full view of Pareto-optimal solutions for decision makers. Based on the reverse influence sampling (RIS) model, we propose a similarity matching-based RIS sampling method to accommodate diverse users, and then transform our original problem into a weighted coverage problem. Subsequently, to solve this problem, we propose a greedy-based incrementally approximation approach and heuristic-based particle swarm optimization approach. Extensive experiments on two real-world geo-social networks clearly validate the effectiveness and efficiency of our proposed approaches.**

***Index Terms*—Complex network, influence spread, optimization.**

Manuscript received December 24, 2018; accepted March 14, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002000, in part by the National Natural Science Foundation of China under Grant 61872033, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JQ6034, in part by the Fundamental Research Funds for the Central Universities under Grant 31020180QD139, and in part by the European Research Council under Grant 683253/GraphInt. This paper was recommended by Associate Editor J. Liu. *(Corresponding authors: Liang Wang; Fei Xiong.)*

L. Wang and Z. Yu are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 10072, China (e-mail: liangwang0123@gmail.com).

F. Xiong is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: xiongf@bjtu.edu.cn).

D. Yang is with the eXascale Infolab, University of Fribourg, 1700 Fribourg, Switzerland.

S. Pan is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia.

Z. Yan is with the Centre for Artificial Intelligence, University of Technology Sydney, Sydney, NSW 2007, Australia.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2906078

I. Introduction

With the rapidly increasing popularity of online social medias [1], for example, Twitter and Facebook, viral marketing has become one of the most cost-effective marketing tool to promote products and services through the word-of-mouth effects. A fundamental problem in viral marketing is influence maximization (IM) [2] which strives to identify a seed set of k influential nodes, called seed users, over one social network G to trigger a maximum expected number of nodes influenced. In the past decade, IM problem has received considerable attention in both academic and industrial communities [3], [4].

In most of existing works concerning IM problem, two assumptions are enclosed: 1) the benefit achieved by influencing any user is equal and 2) all the initial seed users can be recruited with the same cost (unit one). The proliferation of position enabled devices imports spatial information into the traditional social networks, that is, geo-social networks [3], [5]. Based on the exposed locations and associated semantic meaning, users' profile can be captured to improve the performance of marketing campaigns. However, to launch a location-aware promotion in geo-social networks, the first assumption may not be true due to users' diverse spatial distributions. For instance, users who are close to a target location have a greater probability to adopt the promoted product [6], [7], for example, a restaurant or a gym, and so on. Therefore, it is necessary to differentiate these potential customers from other users, and attach more importance to them; otherwise, it may direct "wrong audiences" who are not profitable.

Moreover, due to the different influence capability of users in social networks, the recruited seed users often incur different recruiting costs. It goes without saying that the cost to incentivize high-profile individuals should be substantially higher than common users. For example, the famous sport star Cristiano Ronaldo' worth per tweet is estimated as \$1 613 309 [8]. Thus, if we follow the second assumption, it may result in infeasible scheme with unaffordable budget. In summary, to achieve a successful promotion in geo-social networks, it is essential to recognize targeted customers from available users, and consider seed users' different recruiting cost resulted from their ranks.

Fig. 1 illustrates our motivation. A geo-social network consists of a group of nodes and edges (e.g., friend relationship). Each user has his/her geographical location distribution and personal preference to locations of different categories (e.g., soccer, fitness, etc.). For ease of exposition, the recruiting cost

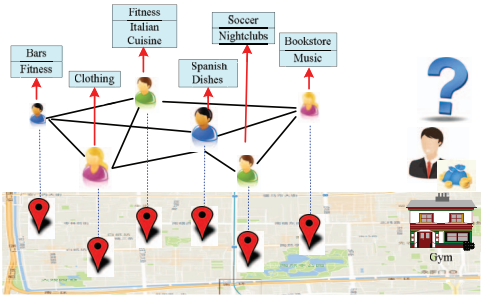


Fig. 1. Toy example for motivation.

of a user is proportional to the size of its icon. Given a gym club located at q, its managers plan to promote the business on geo-social networks. Intuitively, only the users who reside near q and are interested in fitness may be potential consumers. In other words, the targeted users are determined considering the factors of both geographical location and advertising topic. In view of the diversity in benefits and incentive costs, it is natural for business owners to seek a subset of seed users with maximum influence spread and minimum promotion cost.

However, this problem is intractable in practice. The reason lies in the fact that influence spread and promotion costs are like two sides of one coin, both of them cannot be optimized simultaneously [9]. Thus, business managers have to balance the tradeoff between these two objectives, and make a more suitable decision. By imposing a pregiven promotion budget, several research efforts transfer it into a single optimization problem with a bounded constraint. Unfortunately, subjectively determining one appropriate budget without *a priori* knowledge is not an easy task, and remains an open question. Therefore, it is indispensable to provide a wide range of choices (i.e., a set of Pareto-optimal solutions where each one represents a tradeoff between these two objectives) for decision makers. The importance has been seen in real-world scenarios, and recognized by management science literatures [10]. Unfortunately, there is little work focusing on a comprehensive analysis with respect to the involved goals, and most existing works treat all the users equally (i.e., having equal influence benefit and incentive cost). To bridge this gap, in this paper, we aim at developing a framework for influence spread and promotion cost optimization in geo-social networks.

As a result, the problem we studied is inherently more complicated than the traditional IM problem, and calls for more sophisticated algorithms. In response to the concerns mentioned above, we make the following contributions.

1) To approach realistic applications better, we identify and formulate the targeted influence spread-promotion cost (TIS-PC) optimization problem in geo-social networks.
2) We propose a conceptual framework to solve our TIS-PC problem. Based on the technique of reverse influence sampling (RIS), we devise a similarity matching-based weighted RIS (SMW-RIS) strategy, and transform the original problem into a weighted coverage problem.
3) To thoroughly disclose the full view of feasible Pareto-optimal solutions, we propose two optimization algorithms, that is, greedy-based incrementally search algorithm (GIS-TIM) and heuristic-based particle swarm optimization (PSO) algorithm IS-MOPSO+, respectively.
4) We conduct extensive experiments on two real-world data sets, and show the efficiency and effectiveness of our proposed methods.

II. Related Work

A. Influence Maximization in Social Networks

A large amount of literature has studied IM problem [2]. Kempe *et al.* are the first formally define influence propagation model, and prove the hardness of IM problem. Moreover, they propose a greedy-based search algorithm with a $(1 - 1/e - \varepsilon)$ approximation ratio [2]. As a seminal paper, it motivates a plethora of research to study IM problem in the past decade [11], [12]. To improve the runtime efficiency, Leskovec *et al.* [13] utilized the submodularity property, and devise a lazy-forward heuristic algorithm CELF. Based on it, Goyal *et al.* [14] proposed an improved version, namely CELF++, to further promote query efficiency. Chen *et al.* [11] proposed a heuristic-based approach which is scalable to millions of nodes and edges.

Borgs *et al.* [15] made a breakthrough for IM problem on IC model. By employing a random reverse reachable (RR) set, they propose a near-linear time approach RIS, which can return a $(1-1/e-\epsilon)$ approximate-ratio solution with a probability of $1-n^{-l}$. Subsequently, Tang *et al.* [16] devised more efficient algorithms, TIM and IMM, by incorporating novel heuristics. Most recently, Nguyen *et al.* [17], [18] developed SSA and D-SSA algorithms to improve IMM approach in terms of empirical efficiency and theoretical threshold. Although the heterogeneous distribution of users' influence weights or incentive cost is considered in existing works [12], [19], [20], they usually optimize influence spread to achieve just one solution, without comprehensively investigating the relationship between those two competitive objectives.

Several research study the profit maximization problem in social networks [9], [21], [22]. From the perspective of products' pricing strategy, Zhu *et al.* [9] demonstrated that influence and profit are like two sides of one coin. Wei and Lakshmanan [21] employed an unbudgeted greedy framework to maximize expected profit by incorporating prices and valuations. Xu *et al.* [22] recognized the most valuable customers for profit maximization. Yang and Liu [23] studied an IM-cost minimization problem in social networks based on multiobjective PSO algorithm. However, it adopts one simple influence spread model in which promotion information could only diffuse from a seed user within 2-hops in social network, instead of the whole network.

B. Influence Maximization in Geo-Social Networks

By incorporating spatial dimension, Li *et al.* [24] searched a seed set to maximize the influence propagation in a predefined region. Wang *et al.* [3] proposed distance-aware IM (DAIM) problem in geo-social networks, and propose two novel index-based approaches to support online query. Li *et al.* [5] identified a geo-social influence spanning maximization problem,

which attempts to find the maximum geographic spanning region with a predefined regional acceptance rate. The most relevant work to ours is [3]. However, since it assumes that all the initial seed nodes are acquired with same cost, the problem is still a classical IM problem with heterogeneous weights in nature. In particular, the proposed approaches are unable to search a group of Pareto-optimal solutions, but only one solution with k seed nodes. Thus, it cannot be directly applied to address our problem.

III. Research Background

A. Preliminary

To simulate influence propagation process, in this paper, we focus on the independent cascade (IC) model as it is a widely adopted propagation model in [3], [11], and [24]. Nonetheless, our proposed framework can be easily extended to other models, including linear threshold model, general triggering model [2], etc. Considering a directed graph $G = (V, E)$, where V denotes a set of nodes and E denotes a set of edges (friendships or "follow" relationships between nodes), with $|V| = n$ and $|E| = m$. For any two nodes u and v in V, if $(u, v) \in E$, v is regarded as an outgoing neighbor of u, and u is an incoming neighbor of v. Moreover, the directed edge (u, v) is associated with a propagation probability $p(u, v) \in (0, 1]$ to quantify the influence from node u to v. In practice, the propagation probability $p(u, v)$ is usually set to $1/N_v$, where N_v denotes the in-degree of node v [3], [19], [24]. And the cost of propagation probability computation is $O(|V|^2)$.

1) Influence Maximization: Under IC model, assuming we have a set of seed nodes $S \subseteq V$, influence diffusion happens in a discrete-time stochastic process as follows.

1) At round $t = 0$, all nodes in S are activated and the other nodes are inactivated.
2) At round $t \geqslant 1$, once a node u gets activated, it remains in active state for the subsequent rounds, and it has only one chance to activate its each inactive neighbor node v with influence probability $p(u, v)$.
3) The influence diffusion process terminates until no more nodes can be activated.

After the above propagation process stops, let $I(S)$ represents the number of activated nodes, that is, influence spread. Mathematically, the expected value of $I(S)$, denoted by $\mathbb{E}[I(S)]$, can be calculated by $\sum_{v \in V} p(S \mapsto v)$, where $p(S \mapsto v)$ is the probability that node v can be activated by seed node set S. Given one social network G and an integer k, IM problem strives to seek a seed set S with k influential nodes to maximize $\mathbb{E}[I(S)]$, which can be formalized as follows:

$$S = \arg \max_{S:|S|=k} \mathbb{E}[I(S)] = \arg \max_{S:|S|=k} \sum_{v \in V} p(S \mapsto v). \quad (1)$$

2) Reverse Influence Sampling: Since IM problem is NP-hard, Kempe *et al.* [2] formulated it as a combinatorial optimization problem; as a solution, they propose a greedy-based approach to yield a near-optimal seed set S. However, this approach suffers from a long computation time, especially on the process of estimating influence spread. To attack this efficiency issue, a state-of-the-art approach, namely RIS, is proposed by Borgs *et al.* [15]. Specifically, it employs a concept of random RR set [16], [25], which we will explain as follows: let g be a subgraph instance obtained by removing each edge (u, v) in G with probability $1 - p(u, v)$. Given a sampled instance g and a node v, the RR set $R(g, v)$ represents the set of nodes in g which can reach the selected source node v. If source node v is picked randomly from G, $R(g, v)$ is called a *random RR set* [4]. Fig. 2 shows a toy example of a reverse reachable set. Here, we list three subgraph instances: g_1, g_2, and g_3, where the number associated with each edge indicates the corresponding propagation probability. The RR set for node c in g_1, that is, $R(g_1, c)$, is $\{c, d, e\}$, since these three nodes can reach c in subgraph instance g_1.

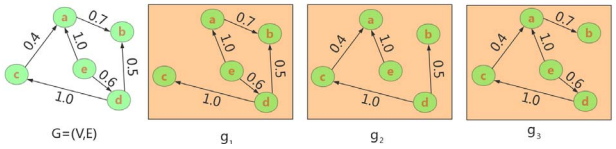


Fig. 2. RR set.

Based on *random RR set*, Borgs *et al.* [15] proposed a lemma to prove that random RR sets can be utilized to estimate the expected influence spread.

Lemma 1 [15]: Given a seed set S, a sampled instance g from G, and a random *RR set* $R(g, v)$. Let p_1 be the probability that S can activate v in a propagation process, and p_2 be the probability that $R(g, v)$ intersect with S, then $p_1 = p_2$.

Intuitively, a node u in RR set $R(g, v)$ must have at least one connected path from u to the given source node v among network G. It indicates that u has a probability to influence v in propagation process. According to Lemma 1, if a node u has a greater influence on other nodes in G, u must have a higher probability to appear in a group of *random RR sets* for different random nodes.

Let $\mathbb{R}$ be a group of *random RR sets*, $\mathbb{R} = \{R_1, R_2, \ldots, R_\theta\}$, $C_{\mathbb{R}}(S)$ be the number of *RR sets* in $\mathbb{R}$ that intersect with S, that is $R_i \cap S \neq \varnothing, 1 \leqslant i \leqslant \theta$. According to Lemma 1, an unbiased estimation of seed set S's expected influence spread can be employed as $(n/|\mathbb{R}|)C_{\mathbb{R}}(S)$, which can be formalized as follows:

$$\mathbb{E}\left[\frac{n}{|\mathbb{R}|} C_{\mathbb{R}}(S)\right] = \mathbb{E}[I(S)]. \quad (2)$$

In other words, if a given seed set S can cover most of the *RR sets*, S is likely to maximize $\mathbb{E}[I(S)]$. Based on this idea, the original IM problem can be converted into a coverage maximization problem.

B. Problem Definition

Consider a marketing promotion query $Q = (q, T^{\#})$, where q and $T^{\#}$ denote a specific location and an advertisement topic, respectively. In geo-social networks, not all the users are potential customers for Q. Intuitively, users who are close to location q and interested in $T^{\#}$ have a greater probability to adopt it [6], [7]. As a result, it is natural to attach more importance to those targeted users. In the following, with respect to these two key factors, that is, spatial proximity and topic interest, we will quantify users' weights for query Q.

1) Spatial Proximity: Following the practice in previous works [3], we employ a widely used decay function as below to measure the probability that user v may visit the promoted

location q, in which the measurement of distance adopts Euclidean distance

$$sp(v, q) = \alpha e^{-\beta \text{dist}(v,q)}. \tag{3}$$

Clearly, according to (3), users who are close to q would have higher probability of visiting q. Considering the skew distribution of users' historical check-ins [24], [26], we determine user v's place as the visited position which is nearest to the promoted location in advertising query. It should be pointed out that our proposed approaches are orthogonal to the choice of user place. Based on the calculation of spatial proximity, our proposed approaches could be easily extended to any choice of user place determination.

2) Topic Interest: Here, we leverage a vector space model to capture advertisement topics and user profiles [27]–[29]. Formally, given a topic space $T = (kw_1, kw_2, \ldots, kw_m)$, where kw_i, $1 \leqslant i \leqslant m$, represents one keyword, for example, restaurant, gym, etc., a promoted advertisement's topic $T^{\#}$ can be formalized as a binary vector $\vec{X} = (x_1, x_2, \ldots, x_m)$. If keyword kw_i is contained in $T^{\#}$, the entry x_i in $\vec{X}$ equals to one; otherwise, it is zero. In other words, advertisement's topic can be formally represented as: $T^{\#} = T * \vec{X}$. With respect to user profile, each user v is associated with a weighted term vector $\vec{Y}$= $(y_1, y_2, \ldots, y_m)$, where y_i indicates user v's preference to keyword kw_i, and $\sum_{i=1}^{m} y_i = 1$. Based on the shared check-ins, weighted term vector can be learned directly from the distribution of location semantic. Hence, the interest degree of an advertisement to user v can be calculated as the dot product of binary vector $\vec{X}$ and weighted term vector $\vec{Y}$

$$ti\big(v, T^{\#}\big) = \vec{X} \cdot \vec{Y} = \sum_{i=1}^{m} x_i * y_i. \tag{4}$$

Therefore, we weight each user v for a given promotion $Q = (q, T^{\#})$, by integrating both spatial proximity and advertisement topic interest

$$w(v, Q) = w\big(v, \big(q, T^{\#}\big)\big) = sp(v, q) * ti\big(v, T^{\#}\big). \tag{5}$$

Based on (5), the potential users can be identified. As discussed above, our main goal is to disseminate the promotion information to targeted users. Next, we will formally define the concept of targeted influence spread.

Definition 1 (Targeted Influence Spread): Given a geo-social network $G = (V, E)$ and a promotion query $Q = (q, T^{\#})$, the targeted influence spread of a set of seed nodes $S \subseteq V$, denoted by $TI_Q(S)$, is computed as $\sum_{v \in V} p(S \mapsto v) w(v, Q)$, where $w(v, Q)$ is the weight of user v for Q.

To initiate promotion campaign, it is necessary to provide seed users with certain incentives. In general, each user's recruiting cost is not the same, which depends on its rank in social network, for example, degree centrality [20]. Here, in order to generalize the related applications, we employ PageRank centrality to indicate each user's recruiting cost, denoted as $c^*(v)$ for user v, rather than real price (e.g., \$10 000). Note that, to enhance the adaptability, we conduct normalization operation upon the obtained centrality values as follows:

$$c(v) = \frac{c^*(v) - c^*_{\min}}{c^*_{\max} - c^*_{\min}} \tag{6}$$

where $c^*_{\min}$ and $c^*_{\max}$ indicate the minimum and maximum centrality of all nodes in V, and $c(v)$ is the normalized centrality value. Thus, the promotion budget is set as the sum of all seed users' dimensionless centralities, that is, $C(S) = \sum_{v \in S} c(v)$.

However, influence spread and promotion cost could not be optimized simultaneously. If one objective is improved, the other will be degenerated accordingly. In most cases, with respect to different purposes, it is difficult to precisely determine a suitable preference weight between these two competing objectives. As a result, it is indispensable to uncover the full view of targeted influence spread and promotion cost goals for managers to make a more suitable decision. We now define our TIS-PC optimization problem as follows.

Definition 2 (TIS-PC Optimization Problem): Given a geo-social network $G = (V, E)$, each node $v \in V$ is attached with check-in history, the TIS-PC optimization problem attempts to search a set of optimal solutions S with respect to maximizing targeted influence spread and minimizing promotion cost, for a given promotion query $Q = (q, T^{\#})$. Formally, the TIS-PC optimization problem can be represented as follows:

$$\arg_S \begin{cases} \max \ : \ TI_Q(S) = \sum_{v \in V} p(S \mapsto v) w(v, Q) \\ \min \ : \ C(S) = \sum_{v \in S} c(v). \end{cases} \tag{7}$$

Due to latent influence overlap, targeted influence spread function $TI_Q(S)$ has been proven to be monotonic and submodular in [3]. Theoretically, there exists one upper bound (i.e., saturation point) for $TI_Q(S)$, which could be formulated as $TI_Q(\bigcup_{i=1}^{n} v_i)$. Specifically, when targeted influence spread arrives at its saturation, TI_Q could not be further improved. With regard to promotion cost objective, its loose upper bound could be intuitively derived as $\sum_{i=1}^{n} c(v_i)$, which is the sum of all nodes' recruiting cost. However, on the basis of the saturation influence spread, we could derive its tight upper bound, such as $\sum_{v \in S^*} c(v)$, where S^* is one user set whose influence spread has just reached saturation in influence spread.

Remarks: TIS-PC problem generalizes IM problem in the following ways: 1) TIS-PC problem has heterogeneous distribution over both users' weight and recruited costs instead of being equal to 1 [15], [16], thus the scale of desired seed users is arbitrary, rather than constant, for example, k and 2) TIS-PC problem is dual-objectives optimization [30], [31], which strives to discover Pareto-optimal solutions as much as possible, rather than searching just one solution. And the desired Pareto solutions would cover the whole feasible region with respect to these two objectives. Furthermore, without the size restriction of seed user, such as k in classical IM problem, the variable space has dramatically extended from C_n^k to $\sum_{i=1}^{n} C_n^i$. In other words, not just investigating the combinations of length k nodes, it is required to examine all the combinations of nodes in V. For the enormous searching space, it is fairly intractable to efficiently solve this problem. Therefore, TIS-PC problem is more realistic and complicated than the IM problem. In addition, some research efforts focus on discovering a part of Pareto-optimal solutions, which are usually located at the middle of Pareto front, namely knee solutions [32]. However, considering the complex characteristic of decision space and business managers' different promotion

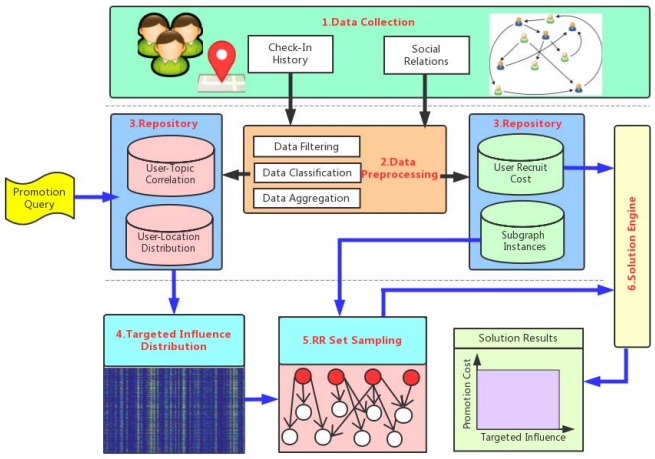


Fig. 3. Conceptual framework for TIS-PC problem.

purposes, it is better to provide a wide range of choices to assist the decision making.

IV. Conceptual Framework

A conceptual framework of TIS-PC is presented in Fig. 3, which consists of six components. Specifically, the data collection module is responsible for collecting users' social relations and historical check-ins. Each check-in is associated with a user ID, a time stamp, a location and its semantic meaning (e.g., bar, restaurant, and so on). The data processing module is used to formulate and aggregate gathered raw data. Based on preprocessed data, the hidden knowledge is extracted in an offline manner and recorded into repository, which includes user-topic correlation, user-location distribution, and user recruit cost evaluation. In addition, the subgraph instances $SG = \{g_1, g_2, \ldots, g_n\}$ are also generated and stored in the repository in advance. While the computation cost of producing subgraph instances is as follows: its time complexity and space complexity are $O(|SG|)$ and $O(|SG| * |V|^2)$, respectively, where $|SG|$ denotes the scale of desired subgraph instances. When a promotion query Q arrives, by leveraging user-topic correlation and user-location distribution, the module of targeted influence distribution is in charge of deriving promotion query's weight distribution over all users. And then, according to the obtained weight distribution, the RR set sampling module samples a subset of source nodes $\tilde{V}$ from V, and produces random RR set by selecting subgraph instance in SG and source node in $\tilde{V}$. After obtaining a group of random RR sets $\mathbb{R}$, the original problem could be transformed into a weighted coverage maximization problem with cost constraints. Toward the dual-objective optimization, the core module, that is, solution engine, is used to retrieve a solution space, and return a set of Pareto solutions.

V. Approaches for TIS-PC Problem

A. User-Topic Correlation Calculation

In this part, based on users' historical check-in records, the correlation between each user v and topic space T can be derived. In essence, the role of user-topic correlation is to characterize and quantify user's preference for different advertisement keywords kw_i, $1 \leqslant i \leqslant m$, in T. And then, for a given promoted advertisement topic $T^{\#}$ combined by any keywords, it facilitates the computation of the topic interest degree of users. In the following, we will elaborate user-topic correlation calculation in detail.

By aggregating all users' historical check-ins, a latent topic space T which contains explicit keywords can be constructed by using topic modeling technique, such as [27]–[29]. For the built topic space $T = \{kw_1, kw_2, \ldots, kw_m\}$, we extract each user's activity semantic, and count the frequency in existing keywords. Based on the obtained empirical distribution over the topic space, each user's profile can be captured. Afterward, by the means of normalization, the correlation between user v and topic space T can be qualified by a weighted term vector $\vec{Y}$, such as $\vec{Y} = (y_1, y_2, \ldots, y_m)$ and $\sum_{j=1}^{m} y_j = 1$, in which y_j indicates v's preference degree for keyword kw_j. For ease of retrieval, we employ a matrix UT to denote the user-topic correlation, in which the row and column denote registered users and keywords, respectively. And each entry $ut(i, j)$ indicates user v_i's preference to advertisement keyword kw_j.

B. Similarity Matching-Based Weighted RIS Sampling

In classical IM problem, all nodes are treated equally. While in our problem scenario, the targeted users are more relevant to promotion query Q, and thus they should be sampled with a higher probability. So, the uniform RIS sampling technique in [4], [16], and [25] cannot be directly applied into our problem, and the unbiased estimation of expected influence spread derived by (2) no longer holds. In this paper, we propose an SMW-RIS sampling approach. The main difference of SMW-RIS sampling from traditional RIS is that it chooses source nodes in accordance with their weight distribution rather than uniformly random selection in RIS. Specifically, one node v is sampled with the probability of $w(v, Q)/\Gamma$, where Γ is the sum of all nodes' weight for query Q: $\Gamma = \sum_{v \in V} w(v, Q)$. Moreover, to ensure that the sampled source nodes will follow the users' weight distribution, we devise similarity-based matching to coordinate the sampling process. Therefore, the expected targeted influence spread can be estimated by Lemma 2.

Lemma 2: Given a set of seed nodes S, the expected targeted influence spread can be estimated as

$$\mathbb{E}\big[TI_Q(S)\big] = \Gamma * \mathbb{E}\left[\frac{C_{\mathbb{R}}(S)}{\theta}\right] \tag{8}$$

where θ denotes the cardinality of random RR sets, $\theta = |\mathbb{R}|$.

Proof: Let $g \sim G$ represents that subgraph instance g is constructed from a geo-social network G

$$\begin{aligned}
\mathbb{E}\big[TI_Q(S)\big] &= \sum_{v \in V} p(S \mapsto v) * w(v, Q) \\
&= \sum_{v \in V} P_{g \sim G}\big[\exists u \in S \text{ and } u \in R(g, v)\big] * w(v, Q) \\
&= \Gamma * P_{g \sim G, v \in V}\big[\exists u \in S \text{ and } u \in R(g, v)\big] \\
&= \Gamma * P_{g \sim G, v \in V}\big[S \cap R(g, v) = \varnothing\big] \\
&= \Gamma * \mathbb{E}\left[\frac{C_{\mathbb{R}}(S)}{\theta}\right].
\end{aligned} \tag{9}$$

■

In IM problem, based on Chernoff bound, most existing works derive a tight bound of θ with greedy-based algorithm [16], [19]. However, our problem differs from the previous works in two aspects: 1) we need a larger feasible search space to discover Pareto-optional solutions for business owners as much as possible, rather than one single solution and 2) the imported spatial proximity and topic interest constraints can remarkably shrink the scale of candidate targeted users, compared with all other users. Consequently, more RR set samplings are required in our TIS-PC problem. In this paper, we adopt a theoretical bound of sampling θ for IM problem in [16], and adjust it to accommodate our problem scenario.

Lemma 3: Given a promotion query $Q = (q, T^{\#})$, $\epsilon > 0$, and $\delta \in (0, 1)$, if the size of RR set samplings θ satisfies the following equation:

$$\theta \geqslant (8 + 2 * \epsilon) * \Gamma * \frac{ln2/\delta + ln\binom{n}{k_0}}{\epsilon^2 \mathrm{OPT}_{k_0}}. \tag{10}$$

Then, we have the following equation holds with at least $1 - \delta$ probability, where OPT_{k_0} denotes the optimal targeted influence spread with k_0 seed nodes

$$\left| \Gamma * \frac{C_{\mathbb{R}}(S)}{\theta} - \mathbb{E}\big[TI_Q(S)\big] \right| \leqslant \frac{\epsilon}{2} * \mathrm{OPT}_{k_0}. \tag{11}$$

Proof: We follow the procedure in [19] to prove the correctness of the lemma. Consider a promotion query $Q = (q, T^{\#})$, let $\rho = P_{g \sim G, v \in V}[S \cap R(g, v) = \varnothing]$, and then $C_{\mathbb{R}}(S)$ can be regarded as the sum of θ independent identically distributed Bernoulli variables with mean ρ

$$\begin{aligned} &P\left[\left| \Gamma * \frac{C_{\mathbb{R}}(S)}{\theta} - \mathbb{E}\big[TI_Q(S)\big] \right| \geqslant \frac{\epsilon}{2} * \mathrm{OPT}_{k_0}\right] \\ &= P\left[\left| \Gamma * \frac{C_{\mathbb{R}}(S)}{\theta} - \rho * \Gamma \right| \geqslant \frac{\epsilon}{2} * \mathrm{OPT}_{k_0}\right] \\ &= P\left[|C_{\mathbb{R}}(S) - \rho * \theta| \geqslant \frac{\epsilon * \theta}{2 * \Gamma} * \mathrm{OPT}_{k_0}\right] \\ &= P\left[|C_{\mathbb{R}}(S) - \rho * \theta| \geqslant \frac{\epsilon * \mathrm{OPT}_{k_0}}{2 * \Gamma * \rho} * \rho * \theta\right]. \end{aligned} \tag{12}$$

Let $\varepsilon_0 = [(\varepsilon * \mathrm{OPT}_{k_0})/(2 * \Gamma * \rho)]$. According to Chernoff bounds, and the fact that $\rho = [(\mathbb{E}[TI_Q(S)])/\Gamma] \leqslant [(\mathrm{OPT}_{k_0})/\Gamma]$, the above equation will be transformed into the following representation:

$$\begin{aligned} &\text{right hand side of (12)} \leqslant 2 * \exp\left(\frac{-\varepsilon_0^2}{2 + \varepsilon_0} * \rho * \theta\right) \\ &= 2 * \exp\left(-\frac{\epsilon^2 * \mathrm{OPT}_{k_0}^2}{8 * \Gamma^2 * \rho + 2 * \Gamma * \epsilon * \mathrm{OPT}_{k_0}} * \theta\right) \\ &\leqslant 2 * \exp\left(-\frac{\epsilon^2 * \mathrm{OPT}_{k_0}^2}{8 * \Gamma * \mathrm{OPT}_{k_0} + 2 * \Gamma * \epsilon * \mathrm{OPT}_{k_0}} * \theta\right) \\ &= 2 * \exp\left(-\frac{\epsilon^2 * \mathrm{OPT}_{k_0}}{8 * \Gamma + 2 * \Gamma * \epsilon} * \theta\right) \\ &\leqslant \delta / \binom{n}{k_0}. \end{aligned} \tag{13}$$

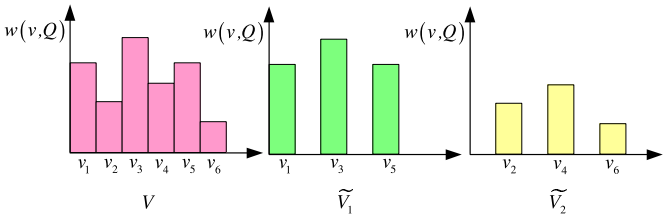


Fig. 4. Matching comparison of different sampled source node sets.

Finally, we have

$$\theta \geqslant (8 + 2 * \epsilon) * \Gamma * \frac{ln2/\delta + ln\binom{n}{k_0}}{\epsilon^2 \mathrm{OPT}_{k_0}}. \tag{14}$$

■

Note that, the role of k_0 in Lemma 3 can be regarded as an upper bound of predefined size of nodes for each solution S. With respect to OPT_{k_0}, in order to find adequate Pareto-optimal solutions, just a loose lower bound of OPT_{k_0} is adequate. Hence, we directly choose *Top-*k_0 users in V, and take the summation of their weights as the estimated OPT_{k_0}.

We utilize matching degree calculation to guide the sampling process. Based on a distance measurement, the matching degree $Sd(\tilde{V}, V)$ between sampled source nodes' weight distribution and existing nodes' can be computed as

$$Sd\big(\tilde{V}, V\big) = \sum_{v \in V} d(v) \tag{15}$$

where the distance function $d(v)$ is defined as follows:

$$d(v) = \begin{cases} w(v, Q), & \text{if } v \in \tilde{V} \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

During sampling process, we search a set of candidate nodes $\tilde{V}$ which can maximize the matching degree of $Sd(\tilde{V}, V)$, where $|\tilde{V}| = \theta$. That is, we strive to achieve a set of θ nodes $\tilde{V}$ which include more source nodes with high weight in V

$$\arg: \max_{\tilde{V}} Sd\big(\tilde{V}, V\big). \tag{17}$$

Concretely, based on the weight distribution of V, a group of sampled source node sets $\tilde{\mathbb{V}} = \{\tilde{V}_1, \tilde{V}_2, \ldots, \tilde{V}_\gamma\}$ is generated. Note that γ could be determined from 1 to $|\tilde{V}|$, when it equals to 1, it is degraded into the conventional RIS. And then, by leveraging matching calculation by (15), the one having maximum value will be chosen as the final sampled node set $\tilde{V}$. As shown in Fig. 4, compared with $\tilde{V}_2$, $\tilde{V}_1$ should be determined as the final sampled node set $\tilde{V}$, as more high-weighted source nodes are sampled in $\tilde{V}_1$. For each node $v \in \tilde{V}$, we randomly select a subgraph instance g from G to generate random RR set $R(g, v)$. Finally, a group of random RR set $\mathbb{R} = \{R(g, v), v \in V\}$ can be obtained. The pseudocode of SMW-RIS is shown in Algorithm 1.

Lemma 4: The time complexity of SMW-RIS algorithm is $O((\theta+1)*|V|+\theta*|E|)$, the space complexity is $O((\theta+1)*|V|)$, where θ denotes the number of sampled source nodes, and $|V|$ and $|E|$ denote the number of nodes and edges in network G. Please refer to the supplementary material, available online, for detailed proofs.

Algorithm 1: SMW-RIS Algorithm

Input: Promotion Query: Q; Social network: $G = (V, E)$; Parameters: ϵ, δ, k_0, γ;
Output: RR Set $\mathbb{R}$;

```
Calculate each node's weight for Q: w(v, Q);
Determine sampling threshold θ based on Eq. (10);
for i=1 to γ do
    for j=1 to θ do
        Choose v_j as source node with probability w(v_j,Q)/Γ;
        Ṽ_i = Ṽ_i ∪ v_j;
    end
    Calculate similarity degree for Ṽ_i: Sd(Ṽ_i, V);
end
Select final sampled node set Ṽ based on Eq. (17);
foreach node v in Ṽ do
    Randomly select a subgraph g from G;
    Generate RR set: R(g, v); ℝ = ℝ ∪ R(g, v);
end
```

C. Greedy-Based Approximate Optimization Approach

Based on sampled random RR set, our problem can be transformed into a discrete coverage maximization problem. Once a user is added into seed node set S, the expected targeted influence spread and total promotion cost will increase accordingly. The increment of promotion cost is directly determined by newly added user's recruiting cost, while the increment in targeted influence spread depends on not only influenced user's weight but also the selecting order. Thus, we consider to loose the original multiobjective optimization into a single-objective optimization problem with constraints, that is, budget constraint targeted influence spread maximization (BTIM) problem. In other words, we strive to optimize the goal of the targeted influence spread, under the constraint of a pregiven promotion budget. And then, by utilizing greedy-based search strategy [33], it can cautiously grow with respect to these two involved objectives, that is, incrementally choose seed users.

Specifically, the BTIM problem is formalized as follows:

$$\arg : \max_{S} \mathbb{E}\big[TI_Q(S)\big] \tag{18}$$

subject to $\sum_{v \in S} c(v) \leqslant B_0$, where B_0 indicates a pregiven budget bound. Generally, the greedy-based search strategy could return a near-optimal solution S with certain approximate-ratio for IM problem [4], [15], [16]. In each iteration, if we can effectively balance between these two involved goals, the iterative trace could be regarded as a group of "optimal solutions." Based on these achieved solutions, it is possible to approximate Pareto-optimal solutions.

As shown in Fig. 5, the Pareto frontier denotes a set of Pareto-optimal solutions $\mathbb{S} = \{S_1, S_2, \ldots, S_n\}$ with respect to involved optimized objectives: targeted influence spread and promotion cost. While the curve in pink represents the greedy-based optimization strategy's iterative search trace, and the point A in green indicates a final returned solution for BTIM problem. Based on hill-climbing method, we

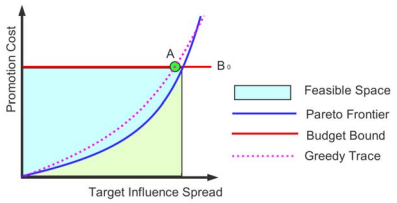


Fig. 5. Greedy-based search trace and Pareto frontier.

devise a GIS-TIM algorithm. By driving the search curve growing with small increment on both multiobjective dimensions, we can move toward Pareto frontier. Moreover, since targeted influence spread may reach saturation, the greedy-based approach would generate many non-Pareto solutions. Thus, budget bound constraint is given in advance to terminate iteration process. As a result, BTIM problem could only retrieve a feasible space in blue restricted by the budget constraint. In realistic application, it can also be regarded as business owners' maximum expenditure.

We will elaborate GIS-TIM's workflow in detail. At each iteration, GIS-TIM examines and picks every remaining node that has not been chosen into seed set S, that $v \in \mathbb{R}/S$. For each node $v \in \mathbb{R}/S$, it first check whether the budget constraint is satisfied after recruiting user v into S. If the total recruit cost is not more than the given budget bound, that is, $C(S \cup v) \leqslant B_0$, user v can be regarded as a valid candidate. While the users who could not meet the hard budget constraint will be skipped immediately. And then, for each valid candidate v, an utility function is constructed to balance these two involved objectives

$$\text{utility}(v) = \big[TI_Q(S \cup v) - TI_Q(S)\big]/c(v). \tag{19}$$

Substantially, in order to maximize the integrated utility function, GIS-TIM algorithm strives to cautiously determine a tradeoff between these two objectives, that is, incrementally increase the influence spread and promotion cost in each step. Among all the valid users, the one having maximum utility is chosen in current iteration. For example, consider two nodes, v_1 and v_2, the incremental benefit, that is, expected targeted influence spread, of these two nodes are 2.7 and 1.9, respectively, such as $TI_Q(S \bigcup v_1) - TI_Q(S) = 2.7$ and $TI_Q(S \bigcup v_2) - TI_Q(S) = 1.9$. While their recruiting costs are 0.7 and 0.4, thus their utility values can be derived based on (19): utility$(v_1) = 3.86$ and utility$(v_2) = 4.75$. Since node v_1's utility is less than v_2's, it is better to choose v_2 in current iteration. Note that, since using an integrated utility function, it no longer has the submodularity nature. Finally, the seed user set S will be updated, that $S = S \cup v$. The procedure will continue until budget constraint B_0 is exhausted. The pseudocode of GIS-TIM is shown in Algorithm 2.

Lemma 5: The time complexity of GIS-TIM algorithm is $O([(\partial * (2 * |V| - \partial - 1))/2] + \sum_{i=|V|-\partial+1}^{|V|} i * \log_2 i)$, where $\partial = (B_0/rc_{\text{ave}})$ denotes the average number of iterations, rc_{ave} indicates the average recruiting cost of all users in V. $\sum_{i=|V|-\partial+1}^{|V|} i * \log_2 i$ refers to the operation in line 10. The space complexity of GIS-TIM algorithm is $O(\theta * |V|)$. Please refer to the supplementary material, available online, for detailed proofs.

Algorithm 2: GIS-TIM Algorithm

Input: Budget Constraint: B_0, Nodes set: V;
Output: Optimization trace $\mathbb{S}$;
Initializing Seed Set $S = \varnothing$;
while $\sum_{v \in S} c(v) < B_0$ **do**
  Initializing valid candidate node set $V_{valid} = \varnothing$;
  for *each node* $v \in V/S$ **do**
    if $c(S \cup v) \leqslant B_0$ **then**
      Merge v into valid candidate nodes V_{valid};
      Calculate node v's utility based on Eq. (19);
    end
  end
  Search node $v \in V_{valid}$ with maximum utility: $arg \underset{v}{:} max\ utility(v)$;
  Merge v into S and record the trace of S into $\mathbb{S}$;
end

D. Heuristic-Based Particle Swarm Optimization Approach

In order to exploit the whole problem space, we turn to a heuristic-based evolutionary method, PSO [34]. It drives a population of candidate solutions (i.e., particles) to move around in the search space based on particle's position and velocity. Note that, as a stochastic algorithm, it may produce good solutions but do not come with a guarantee on the quality of their solution.

By leveraging the classical PSO algorithm, we devise our influence spread multiobjective PSO (IS-MOPSO+) algorithm. First, a group of initializing solutions are randomly generated in feasible solution space. For each solution S_i, $1 \leqslant i \leqslant$ pop, where pop denotes the population size, we use $S_{(i,j)}$ to denote whether the node $v_j \in \mathbb{R}$ is selected into seed set S_i: if node v_j is selected in solution S_i, $S_{(i,j)} = 1$; otherwise, it is zero. And the fitness of each solution S_i is calculated according to (7). Next, based on the concept of dominated relationship, we conduct a pairwise comparison operation to pick out the desired Pareto solution in the current iteration. Given two solution S_i and S_j, if two involved objective functions of S_i are both better than S_j's, it is considered that S_j is dominated by S_i, that is $S_i \succ S_j$. Note that the picked out solutions are recorded into an external storage set, and serve as global best particle p_g. Moreover, in each iteration, if particle S_i's current position is better than its historical position p_l, then the value of p_l will be replaced by current position. And then, each particle's position will be updated according to (20) as below, where b_1 denotes an inertia weight of particle's velocity *vel* (in the first iteration, *vel* is initialized to zero), b_2 and b_3 are two positive constants, and Z_1 and Z_2 are random values in the range of [0, 1]. Following the previous work [34], the parameters of b_1, b_2, and b_3 are set as 0.729, 1.495, and 1.995, respectively

$$p^* = p + b_1 * v + b_2 * Z_1 * (p_l - p) + b_3 * Z_2 * (p_g - p). \tag{20}$$

The procedure will continue until it reaches a predefined maximum iterations. To improve its performance, we develop three strategies in our IS-MOPSO+ algorithm.

1) In order to accelerate convergence, we devise a hybrid strategy to generate initial particle population. First, a part of initial solutions are randomly chosen from nodes in $\mathbb{R}$, instead from V. The reason is that nodes in V but not in $\mathbb{R}$ have no chance to influence source nodes. Second, a portion of initial solutions are imported from the results obtained by GIS-TIM. In this way, initialized particles can directly move close to Pareto frontier in GIS-TIM, thus the performance of IS-MOPSO+ can be improved. Here, we set a parameter r to represent the scale of imported GIS-TIM algorithm's achieved solutions.
2) To improve the search efficiency, a micro-clusters-based mechanism is proposed by exploiting the correlation between candidate nodes. Concretely, based on the overlapped influence spread calculation, the implicit correlation in candidate nodes could be extracted. Mathematically, the correlation could be quantified by the overlapping targeted influence spread between these two involved nodes, such as v_i and v_j

$$rd = \frac{TI_Q(v_i \bigcup v_j)}{TI_Q(v_i) + TI_Q(v_j)} \tag{21}$$

where $\max\{TI_Q(v_i), TI_Q(v_j)\} \leqslant TI_Q(v_i \bigcup v_j) \leqslant TI_Q(v_i) + TI_Q(v_j)$. And then, on the basis of correlation calculation *rd*, all candidate nodes could be grouped into different micro-clusters, where the nodes in the same micro-cluster have larger overlapping ratios than those in other ones. In the subsequent population evolution, the partitioned micro-clusters could be leveraged from the following two aspects. First, in the process of population initialization, instead of randomly select nodes from all the candidate ones, we choose it with respect to the constructed micro-clusters. Concretely, for one initialized solution, the contained nodes should come from different micro-clusters. Second, the particles fly across the micro-cluster space, that is, updating their current locations by a varied velocity value. And both the representations of particle location and velocity are built in micro-cluster space, rather than the dimension of all candidate nodes. In this way, the comparison operation could be implemented on lower-dimensional space, since the scale of micro-clusters is far less than the candidate nodes.
3) After position updating, the newly generated position may become a continuous value, not 0-1 discrete value. Based on the mechanism formulated in (22), we can revise newly generated solution into correct format. The basic idea is that if the entry $S_{(i,j)}$ in S_i is close to 0 or 1, the probability that it is really 0 or 1 is largest, otherwise it would be opposite value. The pseudocode of IS-MOPSO+ is presented in Algorithm 3. And the workflow of IS-MOPSO+ is demonstrated

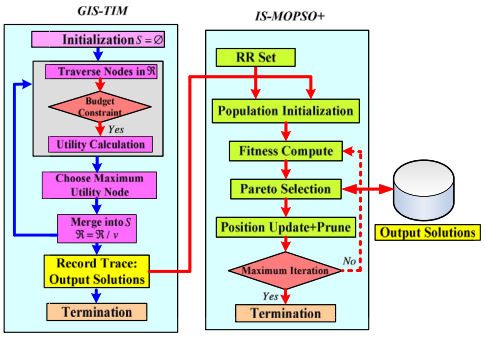


Fig. 6. Workflow of GIS-TIM and IS-MOPSO+ algorithm.

Algorithm 3: IS-MOPSO+ Algorithm

Input: RR Set: $\mathbb{R}$, population size: *pop*, maximum number of iterations: *mt*
Output: A set of Pareto solutions: $\mathbb{S}_{pareto}$

```
Initialize particle population 𝕊, |𝕊| = pop;
Set local best position pos_l as initialized particle;
Initial velocity vel equals to zeros vector;
while not meet iteration mt do
    foreach particle S_i ∈ 𝕊 do
        Calculate fitness of objective functions;
        Update local best position p_l for each particle;
    end
    Determine Pareto solutions in 𝕊;
    Update external storage set 𝕊_pareto;
    foreach particle S_i ∈ 𝕊 do
        Select a particle randomly from 𝕊_pareto as p_g;
        Update and calibrate newly generated position;
    end
end
```

as shown in Fig. 6

$$S_{i,j} = \begin{cases} 1, & S_{i,j} \geqslant 0.5 \\ 0, & S_{i,j} \leqslant 0.5. \end{cases} \tag{22}$$

Lemma 6: The time complexity of IS-MOPSO+ algorithm is $O(mt * ([(\text{pop} * (\text{pop} - 1))/2] + \text{pop} * |S|_{\text{ave}}))$, where $|S|_{\text{ave}}$ represents the average number of nodes contained in particle S across *mt* iterations. And the most computation time is spent on the process of fitness calculation, that is, $\text{pop} * |S|_{\text{ave}}$. Its space complexity is $O((\theta + \text{pop}) * |V|)$. Please refer to the supplementary material, available online, for detailed proofs.

VI. Evaluation and Discussion

We conduct an extensive evaluation on real-world geo-social network datasets to demonstrate the effectiveness and efficiency of our proposed techniques. Our experiments are conducted on a standard server (Windows), with Intel Core i7-6700HQ CPU, 2.60 GHz, and 32-GB main memory.

A. Experimental Setup and Baselines

1) Data Set: We use two real-world geo-social networks in which users share their check-ins [35]. The two networks are directed graphs due to friend relationship, and the detail information is shown as follows: 1) the first one is a small-scale network collected from Foursquare in New York, in which 5100 nodes, 11 933 edges, and 706 344 check-ins are included and 2) another larger-scale network is collected from Gowalla in Boston, in which 145 381 nodes, 546 335 edges, and 8 427 156 check-ins are included. Moreover, a part of check-in records collected from Foursquare are visualized in Fig. 7.

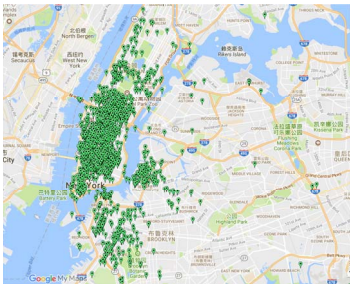


Fig. 7. Spatial distribution of check-ins.

2) Baseline Algorithms: To the best of our knowledge, there is little work focusing on multiobjective optimization of influence spread in geo-social networks. To evaluate the performance of our approaches, we devise three competing baseline algorithms: 1) MODPSO [23]; 2) MOEA/DD [36]; and 3) NSGA-III [37], by adjusting three recent multiobjective optimization methods. As mentioned previously, MODPSO is a state-of-the-art technique to search Pareto solution for IM-cost minimization problem in social networks [23]. While the latter two ones, that is, MOEA/DD and NSGA-III, are recent multiobjective evolutionary algorithms. For the visiting probability $sp(v, q)$ in (3), α and β are empirically set to 1, respectively. For the parameters of required samplings θ, $\epsilon = 0.3$, $\delta = 1/n$, and k_0 is set to $0.1 \times n$.

B. Experimental Results

1) Experiments on Foursquare Dataset: We first conduct experiments on small-scale geo-social networks, by comparing different optimization search techniques. In GIS-TIM algorithm, the upper bound of budget B_0 is set as 10. And in the remaining evolution algorithms (MODPSO, MOEA/DD, NSGA-III, and IS-MOPSO+), the parameters of population size and iterations are set to pop = 150 and *mt* = 60, respectively.

The experimental results are presented as shown in Fig. 8. It is obvious that our proposed GIS-TIM and IS-MOPSO+ algorithms remarkably outperform three baseline algorithms, as they can better approximate the involved two optimization objectives. For example, among the results obtained from these three baseline algorithms, when the goal of targeted influence spread arrives at 40, the incurred promotion cost reaches up to about 30; while our proposed approaches' are just about 2. This verifies the fact that these mainstream evolution algorithms are incapable of attacking the issue of "curse of dimensionality" in larger-scale combinatorial problem, for example, our problem with decision space $\sum_{i=1}^{|V|} C_{|V|}^{i}$. Concretely, they could not rapidly move close to the Pareto front by means of stochastic evolution mechanism

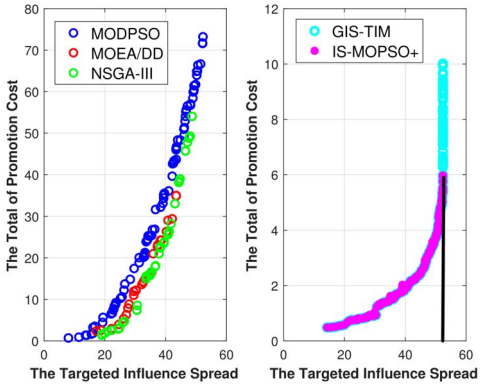


Fig. 8. Results of all achieved solutions.

TABLE I
RESULT ANALYSIS USING WILCOXON SIGNED-RANK TEST

	ER	Δ
GIS-TIM	0.488 ± 0.064	0.7106 ± 0.000
IS-MOPSO+	$\mathbf{0.163 \pm 0.0854}$	$\mathbf{0.5403 \pm 0.0331}$
MODPSO	1.000 ± 0.000	1.1828 ± 0.0246
MOEA	1.000 ± 0.000	1.1678 ± 0.0220
NSGA-III	1.000 ± 0.000	1.1314 ± 0.0182

and random initial inputs. While by adopting an incremental iterative strategy, our GIS-TIM algorithm can directly approximate optimization objectives across the broad decision space.

For the purposes of quantitative evaluation, we independently run each algorithm 30 times, and compare their performance in terms of convergence and diversity indices, where the results are analyzed using Wilcoxon signed-rank tests [38]. Because the true Pareto frontier is unknown, we adopt one index: error ratio (ER) [39] to evaluate the convergence performance among these approaches. Specifically, the measurement ER is formalized as follows:

$$\mathrm{ER} = \frac{\sum_{i=1}^{|\mathbb{S}|} e_s}{|\mathbb{S}|} \tag{23}$$

where e_s is zero if solution S belongs to Pareto solutions and one otherwise, and $|\mathbb{S}|$ denotes the size of returned solutions. Note that, we aggregate all these involved approaches' returned solutions, and develop the common Pareto solutions. For diversity measure, we adopt Δ metric as follows:

$$\Delta = \frac{d_f + d_l + \sum_{i=1}^{|\mathbb{S}|-1} |d_i - \overline{d}|}{d_f + d_l + (|\mathbb{S}| - 1)\overline{d}} \tag{24}$$

where d_i corresponds to the Euclidian distance between consecutive solutions in $\mathbb{S}$, $\overline{d}$ stands for the average of d_i, and d_f and d_l denote the Euclidian distance between the extreme solution of the Pareto solutions $\mathbb{S}$ and the boundary solution in the approximation regarding each of the two objectives, respectively [40]. The results are presented in Table I.

As a deterministic algorithm, the performance of GIS-TIM algorithm should remain unchanged. The fluctuation in ER metric is resulted from the effect of stochastic evolution algorithm IS-MOPSO+, since the Pareto solutions are derived

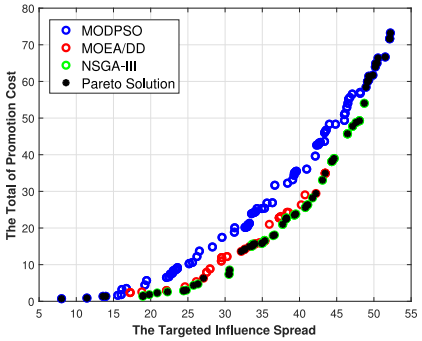


Fig. 9. Results obtained from baseline approaches.

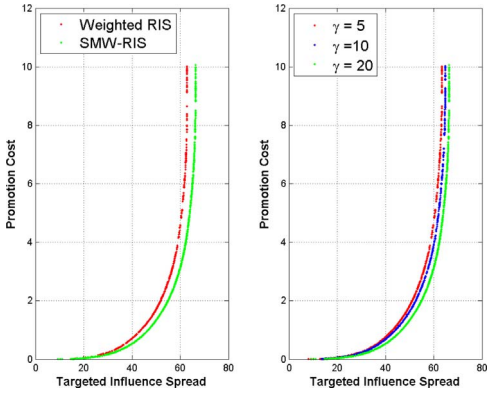


Fig. 10. Results of two sampling strategies.

from all these approaches' results. In particular, it is found that, by feeding the solutions from GIS-TIM, stochastic algorithm IS-MOPSO+'s performance, that is, convergence and diversity, has been improved remarkably, compared to these baseline evolutionary algorithms. Specifically, with respect to the measurement of ER index, it shows that IS-MOPSO+ algorithm can probe more Pareto-optimal solutions, since it benefits from the fed GIS-TIM's results. Furthermore, as implemented Pareto-optimal comparison operation in evolution process, non-Pareto solutions have been eliminated from the traversed solutions in IS-MOPSO+ algorithm. While GIS-TIM algorithm is incapable of distinguishing Pareto-optimal solutions from its resultant solutions. As shown in the right of Fig. 8, when targeted influence spread reaches saturation, for example, the goal of targeted influence spread remains stationary at about 50, while its total promotion cost steadily increases, GIS-TIM will produce many non-Pareto solutions. Moreover, the iterative search process will not stopped until all the nodes are selected into S. It is also why we set a budget bound constraint B_0 as a termination condition. Among these three baseline algorithms, the optimization ability of MODPSO algorithm is the worst, its solutions are far away from the solution curves determined by MOEA/DD and NSGA-III algorithms. And the results are demonstrated in Fig. 9.

In the following, we verify the effect of our proposed SMW-RIS sampling strategy by comparing with weighted sampling in [19]. The parameter of γ equals to 20. For the simplicity of explanation, we just conduct GIS-TIM algorithm on

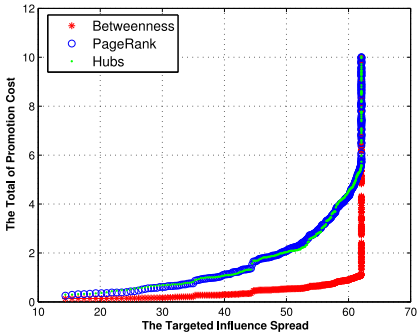


Fig. 11. Impact of different centrality choices.

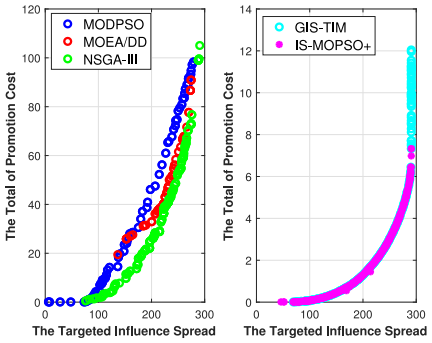


Fig. 12. Results of all achieved solutions.

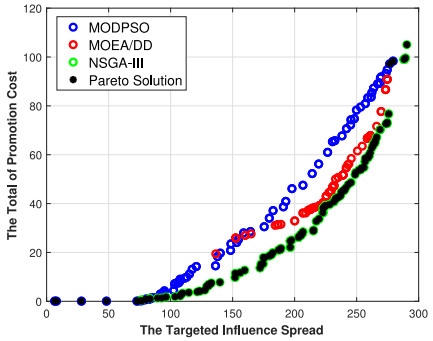


Fig. 13. Results obtained by baseline approaches.

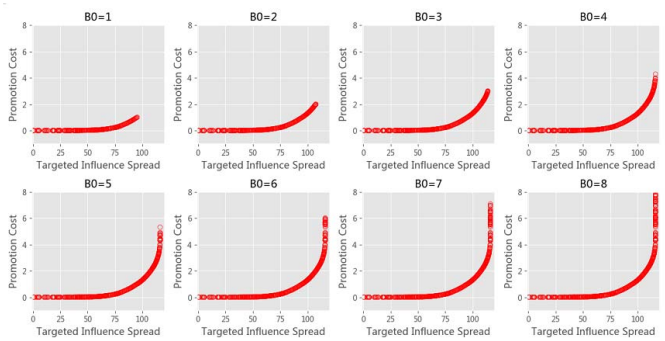


Fig. 14. Impact of promotion budget on GIS-TIM algorithm.

these two strategies. The corresponding results are shown in the left of Fig. 10. Obviously, as the solutions achieved from Weighted RIS strategy is dominated by SMW-RIS strategy's returned solutions, we can draw the conclusion that our proposed SMW-RIS strategy is better than the Weighted RIS strategy. Moreover, we investigate the impact of parameter γ in SMW-RIS sampling algorithm, by varying the value of γ within 5, 10, and 20. The experimental results are reported in the right of Fig. 10. It is clearly that, with the increasing of γ, the performance of achieved Pareto solutions can be improved significantly as the obtained Pareto frontiers become closer to these involved optimization objectives.

In addition, we also examine the impact of centrality choices on the final solution performance, with the GIS-TIM algorithm. Specifically, we conduct experiments under three centrality choices: 1) Betweenness; 2) PageRank; and 3) Hubs, respectively. The experimental results are presented in Fig. 11. Obviously, the output results of betweenness centrality can achieve "less promotion cost," due to the fact that its recruiting cost is generally smaller than the remaining two. But it does not mean that the quality of between centrality's results outperforms the others. As adopted different promotion cost measurements, it could not compare these three measurements' returned results, on the basis of Pareto dominance relationship.

2) Experiments on Gowalla Dataset: In the part, we conduct experiments on larger-scale geo-social networks collected from Gowalla. First, we implement query processing with our proposed approaches and three baseline algorithms. The obtained results are presented as shown in Figs. 12 and 13. Apparently, the results are consistent with the experimental result analysis in Foursquare data set.

Next, we investigate the effect of budget bound B_0 on GIS-TIM algorithm, by varying B_0 from 1 to 8 with 1 increment. The achieved results are reported in Fig. 14. From the chart, the Pareto optimal front could be driven by the parameter B_0, that is, the solution curve grows with the increase of promotion budget B_0. However, since the targeted influence spread has reached saturation when $B_0 \approx 4$, the subsequent iterations cannot continue to optimize it, but only increase promotion costs by recruiting ineffective seed users. In practice, the saturation threshold varies with different queries, that is, promoted location and user-topic correlation distribution, and the structure of networks. So, if we simply tackle TIS-PC problem as a single-objective optimization problem, for example, budgeted max-coverage in [18], it is possible to obtain an inferior solution with limited expected influence spread, but too high promotion cost. And the result once again verifies that it is necessary to explore full view of all Pareto-optimal solutions.

We also investigate the impact of the parameter θ (i.e., the scale of sampled source nodes) on the final solution performance. By varying the size of sampled source nodes $\tilde{V}$ from 100 to 500 with 200 increments, we conduct optimization search by GIS-TIM algorithm. The experimental results are presented as shown in Fig. 15. From the present results, the quality of solutions could be improved significantly by increasing the parameter θ, that is, the results achieved from larger θ are closer to the optimized objectives. The hidden reason is that, with larger θ, the scale of involved nodes in RR Set would increase accordingly, and the candidate nodes could be expanded. While more running time is also required, such as 28.53, 95.71, and 163.84 s, respectively.

By varying the value of budget bound B_0 from 1 to 8, the running time and obtained solutions of GIS-TIM algorithm

TABLE II
RUNTIME EFFICIENCY OF GIS-TIM ALGORITHM

	$B_0 = 1$	$B_0 = 2$	$B_0 = 3$	$B_0 = 4$	$B_0 = 5$	$B_0 = 6$	$B_0 = 7$	$B_0 = 8$
Number of Solutions	259	311	358	384	402	423	446	465
Running Time(Sec.)	162.20	212.96	221.53	241.69	247.60	269.97	272.69	285.36

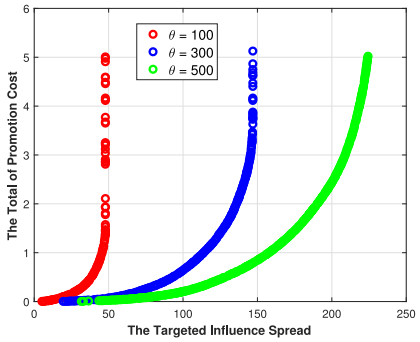


Fig. 15. Impact of different sampled source nodes.

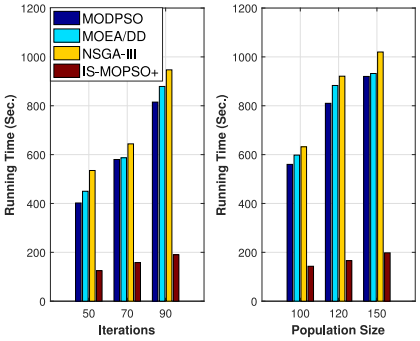


Fig. 16. Runtime efficiency comparison of evolution algorithms.

are reported in Table II. It is obvious that both of computation time and achieved solutions grow with the increasing of B_0. As the budget bound increase, more seed users are required to meet the given budget, thus more iterations will be required accordingly.

Besides, we also examine the runtime efficiency of baseline evolution algorithms and our IS-MOPSO+ algorithm. As shown in Fig. 16, the experimental results are present detailedly. More specifically, the runtime efficiency of our proposed IS-MOPSO+ algorithm has been verified, by varying the parameters of population size and iterations, respectively. Not surprisingly, the computation time grows with the increasing of population size and iteration, as more populations and operations are required. Moreover, the efficiency of IS-MOPSO+ algorithm significantly outperforms other evolution algorithms. The reason is that, by leveraging the micro-cluster-based strategy and the returned solutions of GIS-TIM, its performance in terms of runtime efficiency and convergence has been improved significantly.

VII. CONCLUSION

In this paper, we study one influence spread problem in geo-social networks. Considering the heterogeneity distribution of influenced benefit and recruiting cost, our task is to discover adequate Pareto-optimal solutions to tradeoff between the goals of maximizing targeted influence spread and minimizing promotion cost. Two optimization algorithms, greedy-based incrementally approximate algorithm GIS-TIM and heuristic-based algorithm ISMOPSO+, are proposed. Finally, the efficiency and effectiveness of proposed approaches are demonstrated by extensive experiments on two real-world geo-social networks. In the future, we will apply our proposed approaches into practical applications, and solidly justify its usability. In addition, we will further exploit mechanism to improve the efficient performance without compromising the optimization quality.

REFERENCES

[1] Z. Yu, F. Yi, Q. Lv, and B. Guo, "Identifying on-site users for social events: Mobility, content, and social relationship," *IEEE Trans. Mobile Comput.*, vol. 17, no. 9, pp. 2055–2068, Sep. 2018.
[2] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. ACM 9th SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2003, pp. 137–146.
[3] X. Wang, Y. Zhang, W. Zhang, and X. Lin, "Efficient distance-aware influence maximization in geo-social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 3, pp. 599–612, Mar. 2017.
[4] K. Huang, S. Wang, G. Bevilacqua, X. Xiao, and L. V. S. Lakshmanan, "Revisiting the stop-and-stare algorithms for influence maximization," *Proc. VLDB Endowment*, vol. 10, no. 9, pp. 913–924, 2017.
[5] J. Li *et al.*, "Geo-social influence spanning maximization," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1653–1666, Aug. 2017.
[6] L. Wang, Z. Yu, and D. Yang, "Efficiently targeted billboard advertising using crowdsensing vehicle trajectory data," *IEEE Trans. Ind. Informat.*, to be published. doi: 10.1109/TII.2019.2891258.
[7] M. Wang *et al.*, "PINOCCHIO: Probabilistic influence-based location selection over moving objects," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 3068–3082, Nov. 2016.
[8] (2018). *How Much is a Tweet From LeBron James Worth?* [Online]. Available: http://wojdylosocialmedia.com/much-tweet-lebron-james-worth/
[9] Y. Zhu *et al.*, "Influence and profit: Two sides of the coin," in *Proc. IEEE Int. Conf. Data Min.*, 2013, pp. 1301–1306.
[10] H. Campbell and R. Brown, *Benefit-Cost Analysis: Financial and Economic Appraisal Using Spreadsheets*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
[11] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. ACM 15th SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2009, pp. 199–208.
[12] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "A billion-scale approximation algorithm for maximizing benefit in viral marketing," *IEEE/ACM. Trans. Netw.*, vol. 25, no. 4, pp. 2419–2429, Aug. 2017.
[13] J. Leskovec *et al.*, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2007, pp. 420–429.
[14] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "Celf++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proc. ACM Int. Conf. Companion World Wide Web*, 2011, pp. 47–48.
[15] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proc. ACM-SIAM Symp. Discr. Algorithms*, 2014, pp. 946–957.
[16] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2014, pp. 75–86.

[17] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in *Proc. ACM Int. Conf. Manag. Data*, 2016, pp. 695–710.

[18] H. T. Nguyen, T. N. Dinh, and M. T. Thai, "Cost-aware targeted viral marketing in billion-scale networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[19] Y. Li, D. Zhang, and K. L. Tan, "Real-time targeted influence maximization for online advertisements," *Proc. VLDB Endowment*, vol. 8, no. 10, pp. 1070–1081, 2015.

[20] H. Nguyen and R. Zheng, "On budgeted influence maximization in social networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 6, pp. 1084–1094, Jun. 2013.

[21] L. Wei and L. V. S. Lakshmanan, "Profit maximization over social networks," in *Proc. IEEE Int. Conf. Data Min.*, 2012, pp. 479–488.

[22] K. Xu, J. Li, and Y. Song, "Identifying valuable customers on social networking sites for profit maximization," *Expert Syst. Appl.*, vol. 39, no. 17, pp. 13009–13018, 2012.

[23] J. Yang and J. Liu, "Influence maximization-cost minimization in social networks based on a multiobjective discrete particle swarm optimization algorithm," *IEEE Access*, vol. 6, pp. 2320–2329, 2018.

[24] G. Li, S. Chen, J. Feng, K.-L. Tan, and W.-S. Li, "Efficient location-aware influence maximization," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2014, pp. 87–98.

[25] C. Borgs, "Influence maximization in social networks: Towards an optimal algorithmic solution," *CoRR*, vol. abs/1212.0884, pp. 1–19, Dec. 2012.

[26] L. Wang, Z. Yu, B. Guo, T. Ku, and F. Yi, "Moving destination prediction using sparse dataset: A mobility gradient descent approach," *ACM. Trans. Knowl. Disc. Data*, vol. 11, no. 3, p. 37, 2017.

[27] C. Chen *et al.*, "TripImputor: Real-time imputing taxi trip purpose leveraging multi-sourced urban data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 10, pp. 3292–3304, Oct. 2018.

[28] X. Zhang, S. Ji, S. Wang, Z. Li, and X. Lv, "Geographical topics learning of geo-tagged social images," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 744–755, Mar. 2016.

[29] Z. Yu, H. Xu, Z. Yang, and B. Guo, "Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints," *IEEE Trans. Human–Mach. Syst.*, vol. 46, no. 1, pp. 151–158, Feb. 2016.

[30] L. Wang, Z. Yu, Q. Han, B. Guo, and H. Xiong, "Multi-objective optimization based allocation of heterogeneous spatial crowdsourcing tasks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 7, pp. 1637–1650, Jul. 2018.

[31] C. Liu, J. Liu, and Z. Jiang, "A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2274–2287, Dec. 2014.

[32] S. Bechikh, L. B. Said, and K. Ghédira, "Searching for knee regions of the Pareto front using mobile reference points," *Soft Comput.*, vol. 15, no. 9, pp. 1807–1823, 2011.

[33] L. Wang, Z. Yu, D. Zhang, B. Guo, and C. H. Liu, "Heterogeneous multi-task assignment in mobile crowdsensing using spatiotemporal correlation," *IEEE Trans. Mobile Comput.*, vol. 18, no. 1, pp. 84–97, Jan. 2019.

[34] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw.*, Perth, Australia, 1995, pp. 1942–1948.

[35] D.-N. Yang, H.-J. Hung, W.-C. Lee, and W. Chen, "Maximizing acceptance probability for active friending in online social networks," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2013, pp. 713–721.

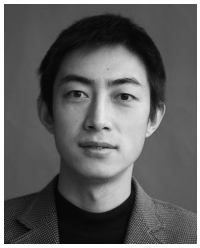
[36] K. Li, K. Deb, Q. Zhang, and S. Kwong, "An evolutionary many-objective optimization algorithm based on dominance and decomposition," *IEEE Trans. Evol. Comput.*, vol. 19, no. 5, pp. 694–716, Oct. 2015.

[37] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints," *IEEE Trans. Evol. Comput.*, vol. 18, no. 4, pp. 577–601, Aug. 2014.

[38] M. Hollander and D. Wolfe, *Nonparametric Statistical Methods*. Hoboken, NJ, USA: Wiley-Intersci., 1999.

[39] S. M. Mousavi, J. Sadeghi, S. T. A. Niaki, and M. Tavana, "A bi-objective inventory optimization model under inflation and discount using tuned Pareto-based algorithms: NSGA-II, NRGA, and MOPSO," *Appl. Soft Comput.*, vol. 43, pp. 57–72, Jun. 2016.

[40] M. Li, S. Yang, and X. Liu, "Diversity comparison of Pareto front approximations in many-objective optimization," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2568–2584, Dec. 2014.



Liang Wang received the Ph.D. degree in computer science from the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China, in 2014.

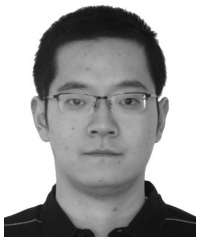
He was a Post-Doctoral Researcher with Northwestern Polytechnical University, Xi'an, China, in 2017, where he is currently an Associate Professor. His current research interests include ubiquitous computing, mobile crowd sensing, and data mining.

Zhiwen Yu (SM'11) received the Ph.D. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2006.

He is currently a Professor with Northwestern Polytechnical University. He was an Alexander von Humboldt Fellow with Mannheim University, Mannheim, Germany, from 2009 to 2010, and a Research Fellow with Kyoto University, Kyoto, Japan, from 2007 to 2009. His current research interests include ubiquitous computing and HCI.

Fei Xiong received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2013.

He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2011 to 2012. He is currently an Associate Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University. His current research interests include Web mining, complex networks, and complex systems.



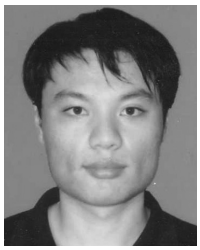
Dingqi Yang received the Ph.D. degree in computer science from Pierre and Marie Curie University (Paris VI), Paris, France, and Institut Mines-TELECOM/TELECOM SudParis, Évry, France.

He is currently a Senior Researcher with the Department of Computer Science, University of Fribourg, Fribourg, Switzerland. His current research interests include ubiquitous computing, social media data analytics, and smart city applications.

Dr. Yang was a recipient of the CNRS SAMOVAR Doctorate Award and the Institut Mines-TELECOM Press Mention in 2015 from the Institut Mines-TELECOM/TELECOM SudParis.

Shirui Pan (M'16) received the Ph.D. degree in computer science from the University of Technology Sydney (UTS), Sydney, NSW, Australia.

He is a Research Fellow with the Centre for Artificial Intelligence, UTS. His current research interests include data mining and artificial intelligence.



Zheng Yan (M'11) received the Ph.D. degree in mechanical and automation engineering from the Chinese University of Hong Kong, Hong Kong, in 2010 and 2014, respectively.

He is currently a Vice-Chancellor Post-Doctoral Research Fellow with the Centre for Artificial Intelligence, University of Technology Sydney, Sydney, NSW, Australia. His current research interests include computational intelligence and model predictive control.