Targeted Influence Maximization Based on Cloud Computing Over Big Data in Social Networks

SHIYU CHEN[1], **XIAOCHUN YIN**[2], **QI CAO**[3], **QIANMU LI**[1,4], **AND HUAQIU LONG**[5]

[1]School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2]Facility Horticulture Laboratory of Universities in Shandong, Weifang University of Science and Technology, Shouguang 262700, China
[3]Academy of Science and Technology Strategic Consulting, Chinese Academy of Sciences, Beijing 100190, China
[4]Jiangsu Zhongtian Technology Company Ltd., Nantong 226463, China
[5]Intelligent Manufacturing Department, Wuyi University, Jiangmen 529020, China

Corresponding author: Qi Cao (caoqi@casisd.com) and Qianmu Li (qianmu@njust.edu.cn)

ABSTRACT This paper focuses on the targeted influence maximization based on cloud computing in social networks. Most of existing influence maximization works assume that the influence diffusion probabilities on edges are fixed, and identify the Top-k users to maximize the spread of influence assuming the knowledge of the entire network graph. However, in real-world scenarios, edge probabilities are typically different based on various topics, and may be affected by information received. Meanwhile, obtaining complete network data is difficult due to privacy and computational considerations. Moreover, existing influence maximization algorithms considering target users do not discuss cloud computing which lead to low computational efficiency when dealing with big datasets in social networks. To this end, this paper proposes a targeted influence maximization solution based on cloud computing. First, a new topic-aware model called tag-aware IC model is presented, which takes into account users' interests, characteristics of the item being propagated, and the similarity between users and the related information. Then, efficient algorithms with approximation guarantee are provided using a bounded number of queries to the graph structure. These methods aim to find a seed set that maximizes the expected influence spread over target users who are relevant to given topics. Finally, empirical studies of the proposed algorithms are designed and performed on real datasets. The experimental results show that the techniques in this paper achieves speedup and savings in storage compared with the state-of-the-art methods.

INDEX TERMS Social network, influence maximization, cloud computing.

I. INTRODUCTION

Online social networks play a critical role in the spread of information, ideas, and influence among people in their modern life [1]–[6]. The new generation of social networks contains billions of nodes and edges. They have become one of the most representative sources of big data. Managing and mining big social data is a challenge for academic and industrial. Social networks have been actively used as a propagation and marketing platform. For instance, in viral marketing, a marketer tries to select a set of customers (called seed set)

with great influence on a promoted product. With a fixed budget on the number of selections, a marketer aims to maximize the number of customers finally adopt the product. This is the classical influence maximization problem [7]–[13].

Since the influence maximization problem is NP-hard [14], Kempe *et al.* [15] first propose a greedy algorithm to solve the problem, which returns a seed set with a $(1 - 1/e - \epsilon)$ approximation ratio to the optimal solution. However, the greedy solution still takes a prohibitively long time to finish. To address the efficiency issue, most researches propose a variety of scalable influence maximization algorithms [16]–[22]. Most of scalable approaches require complete knowledge of the network, and focus on finding global

---

The associate editor coordinating the review of this manuscript and approving it for publication was Francesco Piccialli.

influential nodes over the entire social networks to maximize the expected influence spread. However, in practice, it is difficult and costly to collect the entire network information in social networks with big data. Therefore, this paper combines influence maximization with cloud computing to improve the performance.

In recent years, topic-aware influence maximization is emerging due to practical applications demands. It is extended from the classical influence maximization problem. Topic-aware influence maximization algorithms assume that the influence diffusion probabilities between two users are different based on various topics. However, there are still important challenges not solved in real applications of topic-aware influence maximization. One of them is that the influence strength between two users in real world may depend on different topics, and may be affected by various information received. In an online social network, e.g., Weibo, each user is associated with several tags, which represent one's interests. These tags can be obtained based on hashtags and representative keywords from the contents. Moreover, each information is associated with tags that can be obtained by language processing algorithms. Let us consider an example. There is a user x of Weibo, y is a friend of x. In real life, x and y are both college teachers for computer science. At one day, x bought a programming book and sent a content about the book with pictures and comments. Meanwhile, y received advertisements about two books. One is about this programming book and the other one is about literature. y was interested in the programming book instead of the literature one because it was related to his career. Finally, y bought the programming book because he was influenced by the friend x.

Motivated by the above observations, this paper focuses on targeted influence maximization with partial network information based on cloud computing. The goal is to find the Top-k influential nodes who are relevant to given topics for maximizing the influence spread within target customers. First, a new topic-aware model, which is extended from the IC model, is proposed. Then, as the optimization problem is NP-hard, efficient algorithms with approximation guarantee are provided. The main idea is to adopt the topic aggregation strategy and a bounded number of queries for constructing a targeted sketch under the proposed model. Based on this, the greedy algorithm is used on the maximum coverage problem to find the Top-k seed users. Finally, experimental evaluations are conducted on several real-world social networks. The experimental results verify the effectiveness and efficiency of proposed algorithms.

The rest of the paper is organized as follows. Section 2 presents some background and related works. Section 3 introduces the problem studied in this paper and proposes a new topic-aware model. Section 4 designs the algorithms for targeted influence maximization using a targeted sketching technique. Section 5 demonstrates the efficiency and effectiveness of proposed algorithms on several real social networks. Section 6 summarizes the paper and gives relevant conclusions.

II. RELATED WORK

The influence maximization is a NP-hard problem that has been extensively studied. Kempe *et al.* [15] are the first to formulate influence maximization as a combinatorial optimization problem. They present a greedy approach that yields $(1-1/e-\epsilon)$ approximate solutions. This method evaluates the influence spread of any seed set by the Monte Carlo (MC) simulations, which leads to be computationally expensive. To improve computational efficiency, most researches propose a variety of scalable influence maximization algorithms. These algorithms can be classified into three categories: the greedy-based approach, the heuristic-based approach and the sketch-based approach.

A. GREEDY-BASED APPROACH

The first category of approaches focuses on the optimization for MC simulations. Methods such as CELF [30], CELF++ [31] and UBLF [32], reduce the number of estimation of influence function by a lazy evaluation technique. The technique of [33] utilizes community structure to reduce the number of nodes to be estimated. Although improving the efficiency of the greedy algorithm [15], these approaches still require significant computational overhead when selecting the seed set on graphs with billions of edges. Therefore, the study of influence maximization starts to focus on the heuristic-based and the sketch-based approaches.

B. HEURISTIC-BASED APPROACH

The second category, i.e., the heuristic-based approach, relies on approximate scoring mechanisms to estimate the influence spread of the seed set instead of running heavy MC simulations. This makes influence maximization algorithms more scalable on larger graphs. Many methods generate the seed set according to a ranking metric [34]–[37]. For instance, degree discount [34] adopts a simple discount metric for influence estimation. Group-PageRank [35] and IRIE [36] use the PageRank metric. Methods such as SP1M [39], PMIA [40], IPA [41], LDAG [40], SIMPATH [42] and EASYIM [43], leverage the idea that influence of a node can be estimated using a function of the number of simple paths starting at that node. Although improving practical efficiency, the heuristic-based approach loses theoretical guarantee compared with the greedy-based approach.

C. SKETCH-BASED APPROACH

The third category of approaches overcomes the drawbacks of the greedy-based and the heuristic-based approaches. It achieves a balance between theoretical guarantee and practical efficiency. The sketch-based approach first precomputes a large number of sketches, and then utilizes the sketches to evaluate influence spread. Therefore, it avoids rerunning the MC simulations. In this paper, a sketching technique is proposed to solve targeted influence maximization. Methods such as NewGreedyIC [34], StaticGreedy [44], PrunedMC [45] and SKIM [44], construct sketches by

examining the entire graph for estimating the influence spread. However, the time complexity is still too expensive to run on graphs with millions of nodes and billions of edges. For reducing the time consumption, Borgs *et al.* [23] first propose the concept of the Reverse Reachable (RR) set. The RR set consists of the nodes that can active the randomly sampled node. Moreover, Borgs *et al.* [23] propose a method called Reverse Influence Set (RIS). This method keeps generating RR sets until the total number of edges examined during the generation process reaches a pre-defined threshold. TIM [24] and IMM [25] are techniques that further improve over RIS.

D. TOPIC-AWARE INFLUENCE MAXIMIZATION
In the real world, the influence diffusion probabilities between two users are related to various topics. However, all the above algorithms are not topic-aware. To meet specific applications, topic-aware influence maximization studies are emerging [46]–[48]. The existing topic-aware influence maximization studies can be classified into two categories. The first category considers the users are topic-aware, and the second category formalizes the edges are topc-aware. Aslay *et al.* [46] are the first to study this problem. They design an index-based approach INFLEX with pre-computation and similarity search schemes. INFLEX first samples a set of topic distributions, and pre-computes the seed set under each distribution by CELF++ [31]. At query time, given an online query, the method finds a sufficient set of pre-computed distributions similar to the online query, and combines the corresponding seed sets by using a rank aggregation technique. In [46], Aslay et al. apply maximum-likelihood Dirichlet estimation for sampling topic distributions, Bregman-ball tree for fast similarity search, and Kendall-τ distance-based schemes for seed set aggregation. Chen *et al.* [47] adopt a similar framework and develop optimization techniques, which are suitable for some special graphs with properties like topically-separable and sub-additive. As [46] and [47] have no theoretical guarantee on the influence spread, Chen *et al.* [48] propose the improved method. They develop algorithms having a bounded approximation ratio under PMIA [40]. Specifically, algorithms use the samples to estimate upper and lower bounds for pruning instead of directly answering the query in [46], [47].

Compared with existing topic-aware influence maximization solutions, all users and user-to-user influence strength are both topic-aware in this paper. Moreover, seed nodes are selected depending on multiple factors.

III. PROBLEM FORMULATION
A social network is generally modeled as a graph $\mathcal{G} = (V, E, P)$, where V is the set of nodes (i.e., users) in $\mathcal{G}$, $E \subseteq V \times V$ is the set of directed edges in $\mathcal{G}$. To simulate the information propagation process in the social networks, a number of methods such as Independent Cascade (IC) model and Linear Threshold (LT) model [15] have been proposed.

In this section, IC model is introduced first, and then a new topic-aware model is proposed, which is extended from IC model.

A. IC MODEL
Under the classical IC model [15], each directed edge $e = (u, v)$ has an independent influence probability $P(e)$, which measures the social impact from user u to user v. Each user is either in an active state or inactive state, and an active user can active his inactive neighbors with probability $P(e)$. The dynamic information propagation process under the IC model unfolds in the following discrete steps. Initially, at time step 0, a selected set of seed users S are activated, while setting all other users are inactive. If a user u is first activated at time step t, then for each of his currently inactive outgoing neighbors v, u has probability $P(e)$ to activate v at time step $t + 1$. After time step $t + 1$, u cannot activate any user. Once a user is activated, his active state remains unchanged. The influence diffusion process runs until no more new nodes can be activated.

The IC model is generally considered equivalent to the possible world semantics. Let $G = (V, E_G)$ denote a possible world, where $E_G \subseteq E$. It is one certain instance of the input uncertain graph $\mathcal{G}$, and is obtained by independent sampling of the edges. Therefore, the probability of existence of each possible world G as follows.

$$\text{Prob}[G] = \prod_{e \in E_G} P(e) \prod_{e \in E \setminus E_G} (1 - P(e)) \tag{1}$$

Given a set $T \subseteq V$ of customers, the targeted influence spread in the possible world G is defined as the number $\sigma_G(S, T)$ of target nodes that are reachable from the seed set S in G, i.e.,

$$\sigma_G(S, T) = \sum_{t \in T} I_G(S, t), \tag{2}$$

where the indicative function $I_G(S, t) = 1$ if at least one node in S is reachable to t in G, otherwise, $I_G(S, t) = 0$.

The expected influence spread $\sigma(S, T)$ on the target set T from the seed set S can be calculated as the expectation of the number of reachable target nodes from S, i.e.,

$$\sigma(S, T) = \sum_{G \subseteq \mathcal{G}} \text{Prob}[G] \cdot \sigma_G(S, T) \tag{3}$$

B. TAG-AWARE IC MODEL
In the classical models, e.g., IC and LT model, the influence probability $P(e)$ is normally set to a constant one or $1/d_v$, where d_v is the in-degree of node v. However, in real applications of social networks, the influence probability may depend on various topics, and may be affected by information received. Therefore, a new tag-aware IC model is proposed in this section, which is the extended IC model considering multiple factors.

Let notation C denote the set of all tags present in the social network. These tags contain not only users' interests, but also characteristics of information being propagated.

Appearance of specific tags in the social network affects the corresponding influence diffusion probabilities between two users. Therefore, the influence probability in the classic IC model changes to be a function $P : E \times C \rightarrow (0, 1]$. It assigns a conditional probability to every edge $(u, v) \in E$ given a specific tag $c \in C$. In other words, $P((u, v) \mid c)$ is the probability that a user u will influence his neighbor v, given the tag c.

Similarly, in the tag-aware IC model, a social network is represented by a graph $\mathcal{G} = (V, E, P)$, and each node is associated with active or inactive state. In the new model, the nodes and influence strength on the edges are both topic-aware. In other words, each node $v \in V$ is associated with a set of tags $C_v \subseteq C$, and each special message M is specified with a set of tags $C_M \subseteq C$. Hence, the information diffusion procedure under the tag-aware IC model can be explained as follows.

Given a social network $\mathcal{G}$, a target set T, a message M with a set of tags $C_M \subseteq C$, and a seed set S, the information diffusion process unfolds in discrete steps.

- At time step 0, all nodes in S will become active, and all other nodes will set to be inactive.
- At time step t, for each active node u which is just activated, u will try to activate its every inactive neighbor v. For each tag $c \in C_M$, u will influence its neighbor v with probability $P(e \mid c)$. If v is influenced by u for all tags in C_M, then v will calculate the similarity $\text{Sim}(v_{\text{tag}}, M_{\text{tag}})$ between message M and itself, i.e.,

$$\text{Sim}(v_{\text{tag}}, M_{\text{tag}}) = \frac{|C_v \cap C_M|}{|C_v|} \tag{4}$$

It is the probability that v will continue with the next step, otherwise, v will remain inactive. Next, the following threshold will be computed

$$F(v) = \frac{|\hat{N}_v| + 1}{|N_v| + 1}, \tag{5}$$

where N_v is the set of neighbors of v and $\hat{N}_v$ is the set of active nodes in N_v. If node v is activated, it will continue to activate its neighbors.

- The procedures above run until no more new nodes can be activated.

The tag-aware IC model assumes that tags are independent of one another. Hence, given a set of tags $C_M \subseteq C$, the information diffusion probability via an edge e can be derived as $P(e \mid C_M) = 1 - \prod_{c \in C_M}(1 - P(e \mid c))$. According to the information diffusion process of tag-aware IC model, all edge probabilities depends on $P(e \mid c)$ and $\text{Sim}(v_{\text{tag}}, M_{\text{tag}})$. Hence, the probability of each possible world $G = (V_G, E_G)$ in the

tag-aware IC model can be calculated as follows.

$$\begin{aligned}
&\text{Prob}[G|C_M] \\
&= \prod_{e \in E_G} P(e|C_M) \prod_{e \in E \backslash E_G} (1 - P(e|C_M)) \\
&\quad \cdot \prod_{v \in V_G} \text{Sim}(v_{\text{tag}}, M_{\text{tag}}) \prod_{v \in V \backslash V_G} (1 - \text{Sim}(v_{\text{tag}}, M_{\text{tag}}))
\end{aligned} \tag{6}$$

Analogously, the expected influence spread $\sigma(S, T, C_M)$ of the seed set S in $\mathcal{G}$, given the target set T and the tag set C_M, is computed as:

$$\sigma(S, T, C_M) = \sum_{G \subseteq \mathcal{G}|C_M} \text{Prob}[G|C_M] \cdot \sigma_G(S, T) \tag{7}$$

C. PROBLEM DEFINITION

Given the social network $\mathcal{G} = (V, E, P)$, the target set T, and the size of the seed set k, the targeted influence maximization aims at finding the Top-k seed nodes, such that the influence spread within the target users is maximized.

$$S^* = \text{argmax}_{|S|=k, S \subseteq V \backslash T} \sigma(S, T, C_M) \tag{8}$$

IV. ALGORITHMS FOR TARGETED INFLUENCE MAXIMIZATION

The goal of this paper is to find a seed set that maximizes the influence inside a target set of users considering various topics. This section presents efficient algorithms for targeted influence maximization. For addressing targeted influence maximization using partial information of the network, some ideas are adopted from [49], [50]. At a high level, the proposed solution consists of two phases: limiting for targeted sketching and seed selection.

A. TARGETED SKETCHING

This section proposes a targeted sketching technique with the approximation guarantee $(1 - 1/e - \epsilon)$. Let G denote a deterministic sub-graph of the input uncertain graph $\mathcal{G}$. It is generated by removing each edge $e \in E$ with probability $(1 - P(e \mid c)) \cdot (1 - \text{Sim}(v_{\text{tag}}, M_{\text{tag}}))$ independently based on tag-aware IC model.

Since a sub-graph G is a combination (i.e., graph union) of each chosen tag's individual uncertain graph, random samples are first built on uncertain graphs by considering each tag separately. First, only the edges e associated with a tag c in $\mathcal{G}$ are kept. Then, for each node, its neighbors are probed based on tag-aware IC model, keeping the edges with probability $P(e \mid c) \cdot \text{Sim}(v_{\text{tag}}, M_{\text{tag}})$. Each node never be probed more than once and each edge is sampled at most one chance. Any remaining edge $e \in E$ is removed with probability $(1 - P(e \mid c)) \cdot (1 - \text{Sim}(v_{\text{tag}}, M_{\text{tag}}))$. Next, when receiving some message, which is associated with a set of tags C_M, the targeted sketching method randomly selects one sample from random sample set of each tag of C_M, and combines them together as the sub-graph G.

Let τ denote a threshold that limits the number of nodes in each connected component. In other words, if there are more than a threshold τ nodes in a connected component, the process of probing the neighbors stops. Note that the probing may stop even before hitting τ nodes if there is no new nodes can be activated. In the targeted sketching algorithm, the value of each connected component is the number of target nodes in that component. The influence spread of the node is computed by adding the values of all of the components containing that node.

The above procedures are repeated θ times and θ subgraphs of $\mathcal{G}$ are obtained.

In [49], [50], it was proved that when θ and τ are sufficiently large, the proposed targeted sketching algorithm returns near-optimal results with high probability.

Theorem 1: If θ and τ satisfy separately:

$$\theta \geq \frac{3 \left(\delta + \log 2 \right) \left(k + 1 \right) \log n}{\epsilon^2},$$

$$\tau \geq \frac{n \log \left(1 / \epsilon \right)}{\epsilon k}$$

Then targeted sketching algorithm returns a $(1 - 1/e - \epsilon)$-approximate solution, with at least $1 - 2e^{-\delta}$ probability.

Proof: Let $\Lambda^{(\theta)}$ and $\mathrm{L}^{(\theta)}$ denote the maximizer of $\sigma^{(\theta)}$ and its maximal value. The cardinality of $\sigma^{(\theta)}$ is k. Similarly, let $\Lambda_\epsilon^{(\theta)}$ and $\mathrm{L}_\epsilon^{(\theta)}$ denote the optimal solution to the targeted influence maximization on $\sigma_\epsilon^{(\theta)}$ and its value, respectively. Moreover, let $\Lambda^{\alpha,(\theta)}$ and $\Lambda_\epsilon^{\alpha,(\theta)}$ denote the α-approximate solutions to the targeted influence maximization on $\sigma^{(\theta)}$ and $\sigma_\epsilon^{(\theta)}$, respectively. The goal is to show that any $\Lambda_\epsilon^{\alpha,(\theta)}$ is also $\Lambda^{\alpha(1-\epsilon),(\theta)}$.

Let us consider $\mathcal{G}_\epsilon^{(1)}, \ldots, \mathcal{G}_\epsilon^{(\theta)}$ as subgraphs of $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(\theta)}$. Then, for any set of nodes S, have $\sigma^{(\theta)}(S) \geq \sigma_\epsilon^{(\theta)}(S)$. The process whereby $\mathcal{G}_\epsilon^{(i)}$ is obtained after removing some nodes and edges from $\mathcal{G}^{(i)}$ is called an ϵ-cutting, and subsequently, $\mathcal{G}_\epsilon^{(i)}$ and $\mathcal{G}^{(i)}$ are referred to as the cut the uncut copies, respectively. Finally, $\emptyset_\epsilon^{(\theta)}(v, S)$ is defined using the ϵ-cut copies $\mathcal{G}_\epsilon^{(1)}, \ldots, \mathcal{G}_\epsilon^{(\theta)}$.

The proof follows [49] closely. In particular, it suffices to show the existence of a set $\mathcal{L}$, $\mathrm{card}(\mathcal{L}) = k$ satisfying $\sigma_\epsilon^{(\theta)}(\mathcal{L}) \geq (1 - \epsilon) \mathrm{L}^{(\theta)}$. Because if there exists such a set $\mathcal{L}$, then for any α-approximate solution $\Lambda_\epsilon^{\alpha,(\theta)}$, have (recall $\forall_S \sigma^{(\theta)}(S) \geq \sigma_\epsilon^{(\theta)}(S)$):

$$\begin{aligned}
\sigma^{(\theta)} \left(\Lambda_\epsilon^{\alpha,(\theta)} \right) &\geq \sigma_\epsilon^{(\theta)} \left(\Lambda_\epsilon^{\alpha,(\theta)} \right) \\
&\geq \alpha \sigma_\epsilon^{(\theta)} \left(\Lambda_\epsilon^{(\theta)} \right) \geq \alpha \sigma_\epsilon^{(\theta)} (\mathcal{L}) \geq (1 - \epsilon) \alpha \mathrm{L}^{(\theta)}
\end{aligned}$$
(9)

implying that $\Lambda_\epsilon^{\alpha,(\theta)}$ is also $\Lambda^{\alpha(1-\epsilon),(\theta)}$. To show the existence of such a set $\mathcal{L}$, a probabilistic argument is used to construct a random set L, satisfying $\mathbb{E} \left\{ \sigma_\epsilon^{(\theta)}(\mathrm{L}) \right\} \geq (1 - \epsilon) \mathrm{L}^{(\theta)}$. The set L is constructed is as follows: Starting from $\Lambda^{(\theta)}$, remove ϵk nodes randomly, and replace them with ϵk nodes chosen uniformly at random from V. To show that

$\mathbb{E} \left\{ \sigma_\epsilon^{(\theta)}(\mathrm{L}) \right\} \geq (1 - \epsilon) \mathrm{L}^{(\theta)}$, consider

$$\mathrm{L}^{(\theta)} = \sum_{v \in V} \emptyset^{(\theta)} \left(v, \Lambda^{(\theta)} \right), \qquad (10)$$

$$\mathbb{E} \left\{ \sigma_\epsilon^{(\theta)}(\mathrm{L}) \right\} = \sum_{v \in V} \mathbb{E} \left\{ \emptyset_\epsilon^{(\theta)}(v, \mathcal{L}) \right\} \qquad (11)$$

The inequality, $\mathbb{E} \left\{ \sigma_\epsilon^{(\theta)}(\mathrm{L}) \right\} \geq (1 - \epsilon) \mathrm{L}^{(\theta)}$, would follow if for any node $v \in V$ have,

$$\mathbb{E} \left\{ \emptyset_\epsilon^{(\theta)}(v, \mathrm{L}) \right\} \geq (1 - \epsilon) \emptyset^{(\theta)} \left(v, \Lambda^{(\theta)} \right) + \epsilon \geq \emptyset^{(\theta)} \left(v, \Lambda^{(\theta)} \right)$$
(12)

It only remains to verify the truth of the former inequality, $\mathbb{E} \left\{ \emptyset_\epsilon^{(\theta)}(v, \mathrm{L}) \right\} \geq (1 - \epsilon) \emptyset^{(\theta)} \left(v, \Lambda^{(\theta)} \right) + \epsilon$. First, $\mathbb{E} \left\{ \emptyset_\epsilon^{(\theta)}(v, \mathrm{L}) \right\}$ represents the probability of node v being connected to one of the nodes in the random set L averaged over the θ copies $\mathcal{G}_\epsilon^{(1)}, \ldots, \mathcal{G}_\epsilon^{(\theta)}$. Consider each of the θ copies in uncut sketch, $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(\theta)}$, and the connections between node v and the optimal set $\Lambda^{(\theta)}$ in these uncut copies. If these connections remain unchanged in the ϵ-cut copies $\mathcal{G}_\epsilon^{(1)}, \ldots, \mathcal{G}_\epsilon^{(\theta)}$, then with probability at least $(1 - \epsilon)$ they remain unchanged after ϵk nodes in $\Lambda^{(\theta)}$ are randomly replaced. However, if any of these connections are affected by the ϵ-cutting, then this is an indication that v belongs to a connected component of size τ. This connected component is large enough to contain one of the ϵk nodes of L with probability at least ϵ. Indeed, the probability that one of the τ nodes is chosen is at most ϵ:

$$\left(1 - \frac{\tau}{n} \right)^{\epsilon k} = \left(1 - \frac{\log(1/\epsilon)}{\epsilon k} \right)^{\epsilon k} \leq e^{-\log\left(\frac{1}{\epsilon}\right)} = \epsilon.$$

Based on the above analysis, the algorithm of targeted sketching is as follows:

Step 1: Input a social network $\mathcal{G}$, a set C_M of tags.

Step 2: Keep only the edges associated with a tag $c \in C$ from $\mathcal{G}$.

Step 3: Probe every node by trying to activate their neighbors with probability $\mathrm{P}(e \mid c) \cdot \mathrm{Sim}\left(v_{\mathrm{tag}}, M_{\mathrm{tag}}\right)$.

Step 4: Stop probe if there are no more new nodes can be activated or if the size of the connected component exceeds τ.

Step 5: Remove any remaining edge $e \in E$ with probability $(1 - \mathrm{P}(e \mid c)) \cdot \left(1 - \mathrm{Sim}\left(v_{\mathrm{tag}}, M_{\mathrm{tag}}\right)\right)$.

Step 6: Retain all nodes of $\mathcal{G}$.

Step 7: Repeat Step 2-6 and obtain random samples of every tag c.

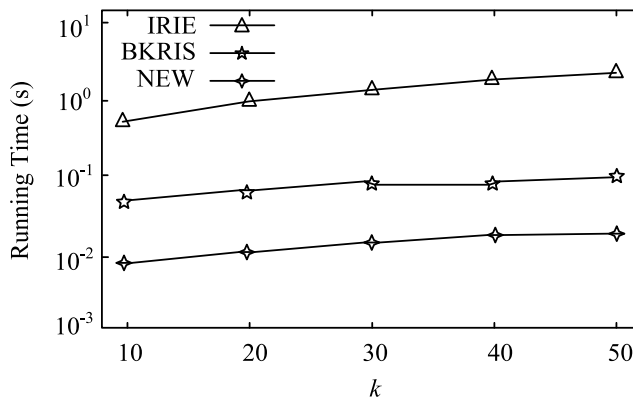**Step 8:** Randomly select one sample from random sample set of a tag $c \in C_M$, and combine them together.

Step 9: Repeat Step 8 θ times and obtain θ sub-graphs of $\mathcal{G}$.
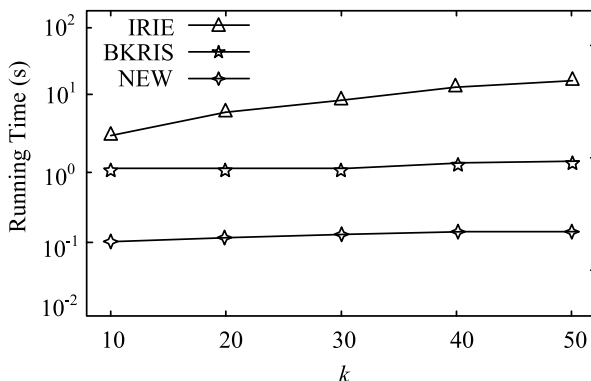

B. SEED SELECTION

This section provides the algorithm for targeted influence maximization on the sampled graphs. The implementation of
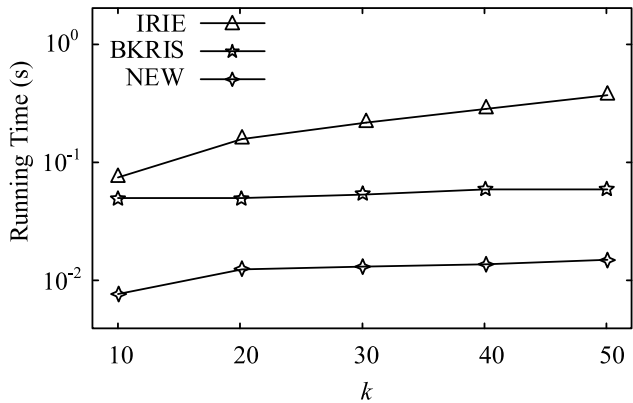
(a) Comparison on NetHEPT dataset.

(b) Comparison on Epinions dataset.

(c) Comparison on DBLP dataset.

FIGURE 1. Running time vs. Seed size.

FIGURE 2. Speedup ratio.

component already contains some nodes in S, then the value of that component is set to zero. The details of the algorithm is described as follows.

Step 1: Input the θ sub-graphs of $\mathcal{G}$, an positive integer k, a set C_M of tags, a target set T.

Step 2: Find the connected components of θ sub-graphs.

Step 3: For every connected component in each of the θ sub-graphs, initialize the current value of the component as the number of target nodes in that component.

Step 4: Initial a node set $S^* = \emptyset$.

Step 5: Identify the node v that maximizes $\sigma(S \cup \{v\}, T, C_M) - \sigma(S, T, C_M)$ by adding the current values of the connected components containing v.

Step 6: Insert v into S^* and set the current value of the connected components containing v to zero.

Step 7: Repeat Step 5-7 until k seed nodes are found.

V. EXPERIMENTAL ANALYSIS

This section experimentally evaluates the targeted influence maximization algorithms against the-state-of-the-art techniques. All the experiments are conducted on a machine with an Intel Core 2.2GHz CPU. All algorithms are implemented in C++.

A. EXPERIMENTAL SETTINGS

1) DATASETS

To demonstrate the effectiveness and efficiency of the proposed algorithm, five real world networks are employed in the experiments. The details datasets is in Table 1. NetHEPT and DBLP are citation networks. Epinions, LiveJournal and Friendster are online social networks. All datasets are publicly available: NetHEPT can be obtained from [51], and all the other four datasets can be obtained from [52]–[55]. Note that Friendster is the largest dataset, which has more than 1.8 billion edges.

2) PROPAGATION MODEL

The popular IC model is used in the experiments to demonstrate the performance of the proposed algorithm.

this algorithm is based on the structure of θ sub-graphs of $\mathcal{G}$ for a specific set of information tags. Since the targeted influence spread on the sub-graphs is submodular, the seed selection phase applies a greedy algorithm to find the seed set with an approximation guarantee $(1-1/e - \epsilon)$ to the optimal solution. First, the connected components of each of the θ sub-graphs are found by using a graph search (e.g. BFS). Then, the number of target nodes in each connected component is counted as the value of that component. Therefore, the main idea of the proposed method is that finding a seed set S that maximizes the total value of all connected components containing at least one seed node. If a connected
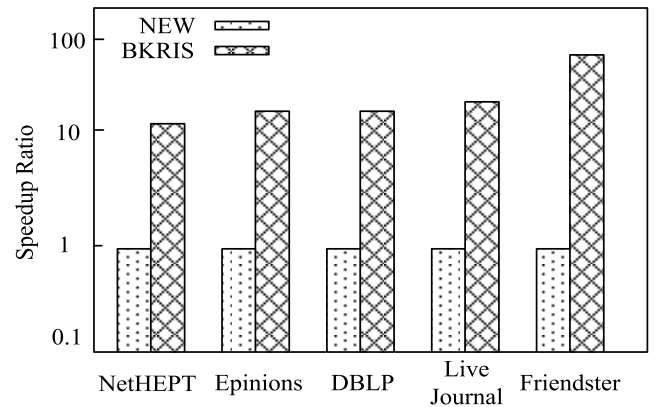
(a) Comparison on NetHEPT dataset.

(b) Comparison on Epinions dataset.

(c) Comparison on DBLP dataset.

(d) Comparison on LiveJournal dataset.
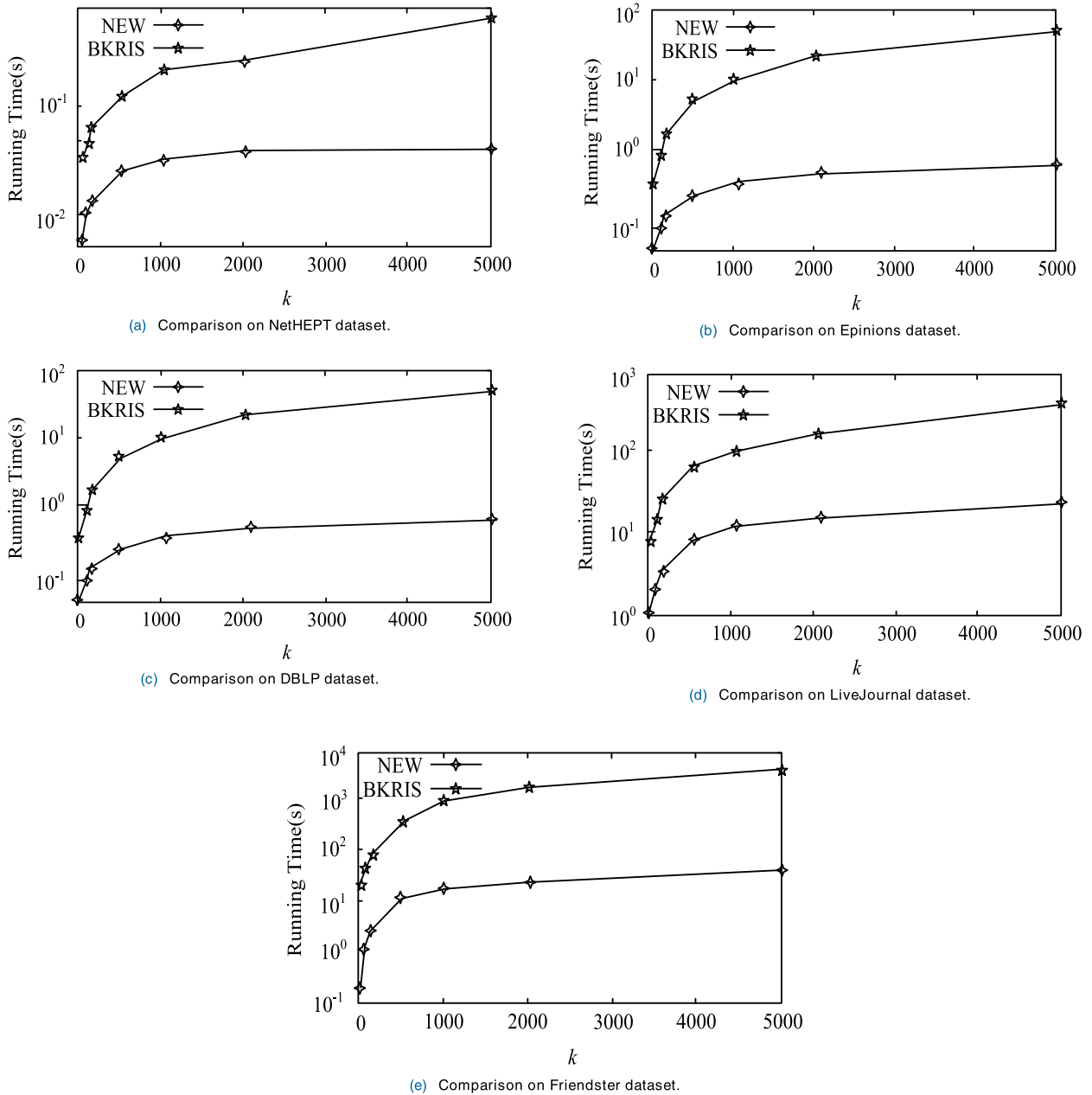

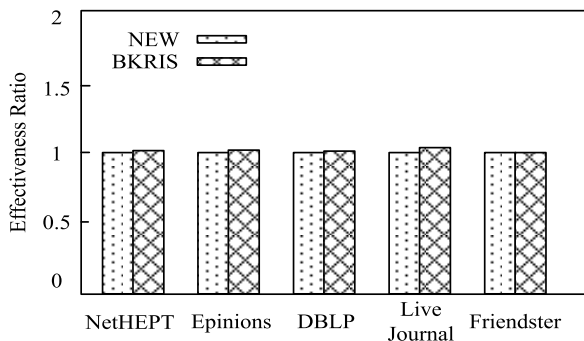(e) Comparison on Friendster dataset.

FIGURE 3. Running time.

FIGURE 4. Effectiveness ratio.

The probability of each edge is set as follows. For each edge (u, v), p(u, v) is assigned as $1/indeg(v)$, where $indeg(v)$ is the

number of incoming neighbors of node v. This setting is the most widely used in the literature.

3) ALGORITHMS COMPARED

In this paper, the proposed algorithm is compared with BKRIS [56], D-SSA [57] and IRIE [36]. BKRIS and D-SSA are the best current algorithms using a sketching technique. IRIE is one of the fastest heuristic method. For BKRIS, when estimating the sample size, the parameters ϵ and ℓ are set as 0.5 and 1 respectively; bk is set as 16 by default for all the datasets. For D-SSA, the parameters ϵ is set as 0.1 and δ is set as $1/n$, which is the default settings suggested by authors. For IRIE, the internal parameters are set as recommended in the original paper. For the proposed
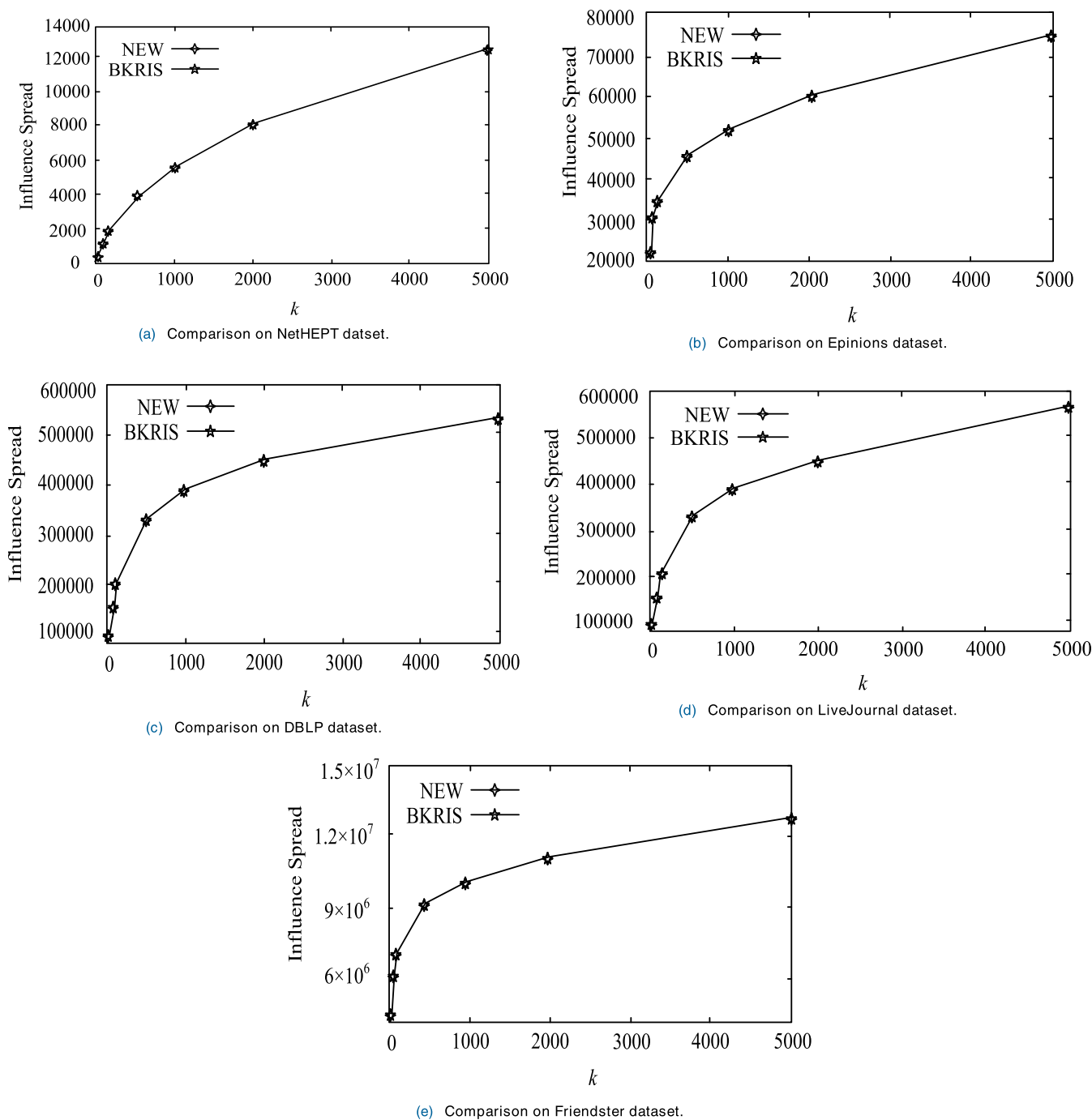
(a) Comparison on NetHEPT datset.

(b) Comparison on Epinions dataset.

(c) Comparison on DBLP dataset.

(d) Comparison on LiveJournal dataset.

(e) Comparison on Friendster dataset.

FIGURE 5. Influence spread.

algorithm, the parameters ρ and σ are set as 0.02 and 100 respectively.

4) PARAMETERS SETTINGS

In the experiments, the seed set size k is varied from 10 to 5000. To demonstrate the scalability of all these algorithms, the seed size $k = 5000$ is only considered. 10,000 round Monte Carlo simulations are conducted on each returned seed set to calculate its influence spread, in order to verify the

algorithms' effectiveness. On each setting and dataset, each algorithm is ran five times and the average is took as the final results.

B. COMPARING IN RUNNING TIME

This section examines the performance of running time of each algorithm. The experiments are conducted over three datasets, NetHEPT, Epinions and DBLP. Due to the scalability issue of IRIE, k is only varied from 10 to 50. The results

TABLE 1. Datasets characteristics.

Dataset	|V|	|E|	Type
NetHEPT	15,233	62,774	undirected
Epinions	75,879	508,837	directed
DBLP	317,080	2,099,732	undirected
LiveJournal	4,847,571	68,993,773	directed
Friendster	65,608,366	1,806,067,135	undirected

are shown in Fig. 1. For most of the cases, IRIE can return a result with close influence to BKRIS and the new algorithm. So the influence spread is not shown here. The experimental results in Fig. 1 show that the new algorithm developed in this paper consistently outperforms IRIE, and BKRIS always runs faster than IRIE. As BKRIS consistently outperforms IRIE, thus in the following experiments, BKRIS is only compared with to evaluate the performance of the proposed algorithm.

Next, the experiment is conducted on all the datasets to compare the effectiveness of BKRIS and the new algorithm. The value of k is set as 5000. Fig. 2 shows that the new algorithm outperforms BKRIS on running time by a wide margin. It is up to two orders of magnitude speedup on all datasets.

Fig. 3 reports the running time on all the datasets by varying k from 10 to 5000. As shown in Fig. 3, the new algorithm significantly outperforms BKRIS under all settings, achieving the speedup of two orders of magnitude.

C. COMPARING IN INFLUENCE SPREAD

In this section, the accuracy of BKRIS and the proposed algorithm is compared by showing the influence spread of the obtained seed set. In Figure 4, the effectiveness ratio is evaluated on all the datasets with k to be 5,000. The effectiveness ratio of BKRIS is rather close to 1 for all the datasets. This shows that two methods have the similar performance in terms of influence spread. Compared with Figure 4, these results show that the new algorithm is superior to BKRIS. Because the new algorithm can achieve the similar influence spread with much less running time. For instance, on the largest dataset Friendster, BKRIS takes more than 10 times time compared with the new algorithm to achieve the same result.

Fig. 5 reports the influence spread on all the datasets by varying k from 10 to 5000. As shown in Fig. 5, two algorithms achieve the similar influence spread.

VI. CONCLUSION

This paper considers the targeted influence maximization based on cloud computing. The goal is to find the optimal seed set that maximizes the topic-aware influence within target customers. In this paper, a new topic-aware model, in which all users and user-to-user influence strength both are topic-aware, is proposed. Since the targeted influence maximization is NP-hard problem, the algorithms with

approximation guarantee are provided. As obtaining full knowledge of the network structure is very costly in practice, a targeted sketching technique is presented based on partial network information. First, random samples are built for each topic under the proposed tag-aware IC model. Then, a targeted sketch is obtained by combining the corresponding samples for given topics using the topic aggregation technique. Based on this, the greedy algorithm is employed to find the Top-k seed users. Performance is evaluated on several real complex social networks with billions of edges and hundreds of topics. The experimental results confirm the effectiveness and efficiency of proposed algorithms.

REFERENCES

[1] H. Song and R. Srinivasan, *Smart Cities: Foundations, Principles, and Applications*. Hoboken, NJ, USA: Wiley, 2017, pp. 1–906.

[2] M. A. Rahman, M. M. Hasan, A. T. Asyhari, and M. Z. A. Bhuiyan, "A 3D-collaborative wireless network: Towards resilient communication for rescuing flood victims," in *Proc. IEEE 15th Int. Conf. Dependable, Autonomic Secure Comput., 15th Int. Conf. Pervas. Intell. Comput., 3rd Int. Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 385–390.

[3] J. Zhou, J. Sun, P. Cong, Z. Liu, X. Zhou, T. Wei, and S. Hu, "Security-critical energy-aware task scheduling for heterogeneous real-time MPSoCs in IoT," *IEEE Trans. Services Comput.*, to be published, doi: 10.1109/TSC.2019.2963301.

[4] A. Ramlatchan, M. Yang, Q. Liu, M. Li, J. Wang, and Y. Li, "A survey of matrix completion methods for recommendation systems," *Big Data Mining Anal.*, vol. 1, no. 4, pp. 308–323, 2018.

[5] C. Zhang, M. Yang, J. Lv, and W. Yang, "An improved hybrid collaborative filtering algorithm based on tags and time factor," *Big Data Mining Anal.*, vol. 1, no. 2, pp. 128–136, 2018.

[6] X. Wang, L. T. Yang, X. Xie, J. Jin, and M. J. Deen, "A cloud-edge computing framework for Cyber-Physical-Social services," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 80–85, Nov. 2017.

[7] H. Song and G. Fink, *Security and Privacy in Cyber-Physical Systems: Foundations, Principles, and Applications*. Chichester, U.K.: Wiley, 2017, pp. 1–472.

[8] L. Qi and Y. Chen, "A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems," *World Wide Web*, to be published, doi: 10.1007/s11280-019-00684-y.

[9] Y. Jiang, H. Song, Y. Yang, H. Liu, M. Gu, Y. Guan, J. Sun, and L. Sha, "Dependable model-driven development of CPS: From stateflow simulation to verified implementation," *ACM Trans. Cyber-Phys. Syst.*, vol. 3, no. 1, pp. 1–31, Aug. 2018, doi: 10.1145/3078623.

[10] J. Zhou, X. S. Hu, Y. Ma, J. Sun, T. Wei, and S. Hu, "Improving availability of multicore real-time systems suffering both permanent and transient faults," *IEEE Trans. Comput.*, vol. 68, no. 12, pp. 1785–1801, Dec. 2019.

[11] X. Wang, W. Wang, L. T. Yang, S. Liao, D. Yin, and M. J. Deen, "A distributed HOSVD method with its incremental computation for big data in Cyber-Physical-Social systems," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 2, pp. 481–492, Jun. 2018.

[12] L. Yu, W. Shuai, and M. S. Khan, "A novel deep hybrid recommender system based on auto-encoder with neural collaborative filtering," *Big Data Mining Anal.*, vol. 1, no. 3, pp. 211–221, 2018.

[13] L. Ren, X. Cheng, X. Wang, J. Cui, and L. Zhang, "Multi-scale dense gate recurrent unit networks for bearing remaining useful life prediction," *Future Gener. Comput. Syst.*, vol. 94, pp. 601–609, May 2019, doi: 10.1016/j.future.2018.12.009.

[14] L. Sun, W. Huang, P. S. Yu, and W. Chen, "Multi-round influence maximization," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2018, pp. 2249–2258.

[15] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2003, pp. 137–146.

[16] L. Qi, Q. He, F. Chen, W. Dou, S. Wan, X. Zhang, and X. Xu, "Finding all you need: Web APIs recommendation in Web of things through keywords search," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 5, pp. 1063–1072, Oct. 2019.

[17] M. A. Rahman, M. Y. Mukta, A. Yousuf, A. T. Asyhari, M. Z. A. Bhuiyan, and C. Y. Yaakub, "IoT based hybrid green energy driven highway lighting system," in *Proc. IEEE Int. Conf. Dependable, Autonomic Secure Comput., Int. Conf. Pervas. Intell. Comput., Int. Conf. Cloud Big Data Comput., Int. Con. Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech)*, Aug. 2019.

[18] W. Gong and L. Qi, "Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment," *Wireless Commun. Mobile Comput.*, vol. 2018, Apr. 2018, Art. no. 3075849.

[19] J. Zhou, J. Sun, X. Zhou, T. Wei, M. Chen, S. Hu, and X. S. Hu, "Resource management for improving soft-error and lifetime reliability of real-time MPSoCs," *IEEE Trans. Comput.-Aided Design Integr.*, vol. 38, no. 12, pp. 2215–2228, Dec. 2019.

[20] X. Wang, L. T. Yang, H. Li, M. Lin, J. Han, and B. O. Apduhan, "NQA: A nested anti-collision algorithm for RFID systems," *ACM Trans. Embedded Comput. Syst.*, vol. 18, no. 4, pp. 1–21, Jul. 2019, doi: 10.1145/3330139.

[21] Z. Shao, C. Xue, Q. Zhuge, M. Qiu, B. Xiao, and E. H.-M. Sha, "Security protection and checking for embedded system integration against buffer overflow attacks via hardware/software," *IEEE Trans. Comput.*, vol. 55, no. 4, pp. 443–453, Apr. 2006.

[22] X. Tang, K. Li, M. Qiu, and E. H.-M. Sha, "A hierarchical reliability-driven scheduling algorithm in grid systems," *J. Parallel Distrib. Comput.*, vol. 72, no. 4, pp. 525–535, Apr. 2012.

[23] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proc. 25th Annu. ACM-SIAM Symp. Discrete Algorithms*, Dec. 2013, pp. 946–957.

[24] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 75–86.

[25] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1539–1554.

[26] L. Qi, X. Zhang, W. Dou, C. Hu, C. Yang, and J. Chen, "A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment," *Future Gener. Comput. Syst.*, vol. 88, pp. 636–643, Nov. 2018.

[27] W. Wei, M. A. Rahman, I. F. Kurniawan, A. T. Asyhari, S. M. N. Sadat, and L. Yao, "Immune genetic algorithm optimization and integration of logistics network terminal resources," in *Proc. 3rd IEEE Int. Conf. Robotic Comput. (IRC)*, Feb. 2019, pp. 435–436.

[28] H. Liu, H. Kou, C. Yan, and L. Qi, "Link prediction in paper citation network to construct paper correlation graph," *EURASIP J. Wireless Commun. Netw.*, to be published, doi: 10.1186/s13638-019-1561-7.

[29] Y. Xu, L. Qi, W. Dou, and J. Yu, "Privacy-preserving and scalable service recommendation based on simhash in a distributed cloud environment," *Complexity*, vol. 2017, Dec. 2017, Art. no. 3437854.

[30] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2007, pp. 420–429.

[31] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proc. 20th Int. Conf. Companion World Wide Web*, 2011, pp. 47–48.

[32] C. Zhou, P. Zhang, W. Zang, and L. Guo, "On the upper bounds of spread for greedy algorithms in social network influence maximization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2770–2783, Oct. 2015.

[33] Y. Wang, G. Cong, G. Song, and K. Xie, "Community-based greedy algorithm for mining top-K influential nodes in mobile social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2010, pp. 1039–1048.

[34] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 199–208.

[35] Q. Liu, B. Xiang, E. Chen, H. Xiong, F. Tang, and J. X. Yu, "Influence maximization over large-scale social networks: A bounded linear approach," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 171–180.

[36] K. Jung, W. Heo, and W. Chen, "IRIE: Scalable and robust influence maximization in social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 918–923.

[37] X. Wang, L. T. Yang, Y. Wang, X. Liu, Q. Zhang, and M. J. Deen, "A distributed tensor-train decomposition method for Cyber-Physical-Social services," *ACM Trans. Cyber-Phys. Syst.*, vol. 3, no. 4, pp. 1–15, Oct. 2019, doi: 10.1145/3323926.

[38] J. Li, Z. Ming, M. Qiu, G. Quan, X. Qin, and T. Chen, "Resource allocation robustness in multi-core embedded systems with inaccurate information," *J. Syst. Archit.*, vol. 57, no. 9, pp. 840–849, Oct. 2011.

[39] M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," in *Proc. AAAI*, vol. 7, 2007, pp. 1371–1376.

[40] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2010, pp. 1029–1038.

[41] J. Kim, S.-K. Kim, and H. Yu, "Scalable and parallelizable processing of influence maximization for large-scale social networks?" in *Proc. IEEE 29th Int. Conf. Data Eng. (ICDE)*, Apr. 2013, pp. 266–277.

[42] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "SIMPATH: An efficient algorithm for influence maximization under the linear threshold model," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 211–220.

[43] S. Galhotra, A. Arora, and S. Roy, "Holistic influence maximization: Combining scalability and efficiency with opinion-aware models," in *Proc. Int. Conf. Manage. Data*, 2016, pp. 743–758.

[44] S. Cheng, H. Shen, and J. Huang, "Staticgreedy: Solving the scalability-accuracy dilemma in influence maximization," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 509–518.

[45] N. Ohsaka, T. Akiba, and Y. Yoshida, "Fast and accurate influence maximization on large networks with pruned monte-carlo simulations," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 138–144.

[46] C. Aslay, N. Barbieri, and F. Bonchi, "Online topic-aware influence maximization queries," in *Proc. EDBT*, 2014, pp. 295–306.

[47] W. Chen, T. Lin, and C. Yang, "Real-time topic-aware influence maximization using preprocessing," *Comput. Social Netw.*, vol. 3, no. 1, pp. 1–13, 2016.

[48] S. Chen, J. Fan, G. Li, J. Feng, K.-L. Tan, and J. Tang, "Online topic-aware influence maximization," *Proc. VLDB Endowment*, vol. 8, no. 6, pp. 666–677, Feb. 2015.

[49] M. Bateni, H. Esfandiari, and V. Mirrokni, "Almost optimal streaming algorithms for coverage problems," in *Proc. 29th ACM Symp. Parallelism Algorithms Archit. (SPAA)*, 2017, pp. 13–23.

[50] M. Bateni, H. Esfandiari, and V. Mirrokni, "Optimal distributed submodular optimization via sketching," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2018, pp. 1138–1147.

[51] [Online]. Available: http://www.cs.cornell.edu/projects/kddcup/datasets.html

[52] [Online]. Available: http://snap.stanford.edu/data/soc-Epinions1.html

[53] [Online]. Available: https://snap.stanford.edu/data/com-DBLP.html

[54] [Online]. Available: http://snap.stanford.edu/data/soc-LiveJournal1.html

[55] [Online]. Available: https://snap.stanford.edu/data/com-Friendster.html

[56] X. Wang, Y. Zhang, W. Zhang, X. Lin, and C. Chen, "Bring order into the samples: A novel scalable method for influence maximization," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 2, pp. 243–256, Feb. 2017. [Online]. Available: https://snap.stanford.edu/data/com-DBLP.html

[57] 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, Jun. 2015. [Online]. Available: http://snap.stanford.edu/data/soc-LiveJournal1.html
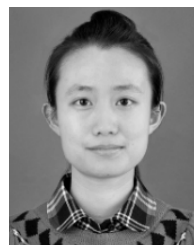
SHIYU CHEN is currently pursuing the Ph.D. degree with the Nanjing University of Science and Technology, China. Her main research interest includes big data mining.

XIAOCHUN YIN received the B.S. degree in education and technology from Qufu Normal University, Qufu, China, in 2004, the M.S. degree in education and technology from Nanjing Normal University, Nanjing, China, in 2007, and the Ph.D. degree from Dongseo University, South Korea, in 2015. She is currently an Associate Professor with the Weifang University of Science and Technology, China. She has published over 20 research articles in international journals and international conferences. Her research interests include network security, the IoT security, authentication protocol, and agricultural intelligence systems.

QI CAO was born in Li'an, Anhui, in October 1984. She is currently an Engineer with the Institute of Science and Technology Strategic Consulting, Chinese Academy of Sciences. Her research interest includes communication and information systems.

HUAQIU LONG is currently pursuing the master's degree with Wuyi University, China. His main research interest includes industrial control network security.

● ● ●

QIANMU LI received the B.Sc. and Ph.D. degrees from the Nanjing University of Science and Technology, China, in 2001 and 2005, respectively. He is currently a Full Professor with the School of Cyber Science and Engineering, Nanjing University of Science and Technology. His research interest includes information security.