PLOS ONE

RESEARCH ARTICLE

TSSCM: A synergism-based three-step cascade model for influence maximization on large-scale social networks

Xiaohui Zhao[1,2], **Fang'ai Liu**[1]*, **Shuning Xing**[1], **Qianqian Wang**[1]

1 School of Information Science & Engineering, Shandong Normal University, Jinan, China, **2** School of Mathematical Science, Shandong Normal University, Jinan, China

* lfa@sdnu.edu.cn

Citation: Zhao X, Liu F, Xing S, Wang Q (2019) TSSCM: A synergism-based three-step cascade model for influence maximization on large-scale social networks. PLoS ONE 14(9): e0221271. https://doi.org/10.1371/journal.pone.0221271

Editor: Gaoxi Xiao, Nanyang Technological University, SINGAPORE

Received: February 25, 2019

Accepted: July 14, 2019

Published: September 3, 2019

Data Availability Statement: The data underlying the results presented in the study are available from http://networkrepository.com.

Funding: This work was supported by the following grants: National Natural Science Foundation of China 61772321, Natural Science Foundation of Shandong Province ZR2016FP07, and CERNET Innovation Project NGII20170508.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Identification of the most influential spreaders that maximize information propagation in social networks is a classic optimization problem, called the influence maximization (IM) problem. A reasonable diffusion model that can accurately simulate information propagation in social networks is the key step to efficiently solving the IM problem. Synergism of neighbor nodes plays an important role in information propagation dynamics. Some known diffusion models have considered the reinforcement mechanism in defining the activation threshold. Most of these models focus on the synergetic effects of nodes on their common neighbors, but the accumulation of synergism has been neglected in previous studies. Inspired by these facts, we first discuss the catalytic role of synergism in the spreading dynamics of social networks and then propose a novel diffusion model called the synergism-based three-step cascade model (TSSCM) based on the above analysis and the three-degree influence theory. Finally, we devise an algorithm for solving the IM problem based on the TSSCM. Experiments on five real large-scale social networks demonstrate the efficacy of our method, which achieves competitive results in terms of influence spreading compared to the four other algorithms tested.

Introduction

The problem of finding the optimal set of influencers, whereby viruses, information, and epidemics propagate through network edges via interactions between individual constituents, has broad applications in a variety of network dynamics areas [1–8]. Viral marketing can inexpensively achieve large-sale product adoption through advertising with a small group of influential customers [1–3]. The financial crisis that resulted from the cascading bankruptcy of major financial institutions in 2008 caused estimated US economic losses as high as $22 trillion [9]. The immunization of structurally important persons can efficiently halt global epidemic outbreaks. The above applications have important characteristics in common, such as budget restrictions and intervention time constraints, and require efficient real-time applications of large-scale data. These features can be simplified to an optimization problem, called the

influence maximization (IM) problem. IM, first studied by Domingos and Richardson et al. [10,11], is a fundamental research problem in social networks. The issue is described as follows: an online social network can be modeled as a graph with vertices representing users and edges representing the links between users. The cascade process on the network is conducted under a specified diffusion model. The IM problem is defined as finding k seed nodes in the network as the source of information propagation such that under the specified diffusion model, the scale of the cascade is maximized. The IM problem is NP-hard. Kempe, Kleinberg and Tardos [12] proposed a greedy algorithm based on Monte Carlo simulation for solving the optimization problem. The performance of the greedy algorithm reached 63% of the optimal solution, but it is not applicable to large-scale networks because it is time consuming. A series of improved algorithms, including CELF [13], NewGreedy [14] and Mixgreedy [15], were proposed to overcome the inefficiency of the greedy algorithm. Unfortunately, although these algorithms are hundreds of times more efficient than the greedy algorithm, their computational complexity remains too high to be applied to growing networks because Monte Carlo simulations are performed to approximate the influence spread of a given seed set.

Many heuristic algorithms have recently emerged. These algorithms can be grouped into two categories: algorithms based on network topology and algorithms based on propagation path. The first group of algorithms mainly use centrality measures, including high degree [15–17], random selection [18], betweenness centrality [17,19], random walk [20], k-shell [21] and community detection [22]. Chen et al. [23] proposed a degree discount algorithm based on degree centrality. In this algorithm, when a node is selected as a seed node, the degree of its neighbor nodes is discounted. Cao et al. [24] designed a core-covering algorithm based on k-shell and the influence radius. Zhu et al. [25] solved the IM problem by developing a structural hole-based algorithm, called SHIM. These algorithms, which are based exclusively on topology, decrease the running time by several orders of magnitude, but they are unstable under different networks and diffusion models. The second group of algorithms are path based, and influence spreading can be efficiently approximated without Monte Carlo simulation. In some applications, we must identify super blockers. Giant components are fragmented by removing key nodes, and propagation is blocked. The mining of the optimal immunization set is based on a diffusion model, such as the susceptible-infected-recovered model [26], the susceptible-infected-susceptible model [26], the independent cascade model [10], the linear threshold model [10] and other cascade models [4,27–30]. Wattes [4] proposed a simple model of global cascades on random networks to explain global cascades that are triggered by small initial shocks. Wei Wang et al. [27] employed the nonredundant information memory characteristic in their social contagion model, which better captured the dynamics of social contagions in the real world, and discussed the cascading process in multiple networks [28,29]. Flaviano and Hernan A [31] mapped the information spread on social networks onto an optimal percolation and presented an algorithm, called collective influence (CI), based on the weak connection between nodes to identify the minimal set of influencers. Based on this information, the authors leverage the behavior of users in real networks, including Twitter, Facebook, APS and LiveJournal, and use the CI algorithm to locate influential spreaders. The experimental results show that the optimal seed set is much smaller than those obtained by other measures. Sen et al. [32] explored CI in the linear threshold model and proposed a method based on the subcritical path to locate influential spreaders. Andrey Y. and David [33] found that the optimal deployment of the seed set resulted from the interaction between network topology and propagation dynamics. They introduced an effective framework for optimizing the maximization or minimization of propagation. Qin et al. [34] devised a diffusion model, called the three-step cascade model (TSCM), that limits the propagation to three-layer neighbors, and they experimentally verified that the model is suitable for simulating information propagation on Sina

Weibo, a social site similar to Twitter. Then, they proposed an algorithm for solving the IM problem based on the TSCM. The above study draws on the three-degree influence theory [35], which we also consider in this work. The above studies show that a reasonable diffusion model that can accurately simulate information propagation on social networks is the key to effectively solving the IM problem. Most existing diffusion models have one commonality: the information spread between a pair of activated and inactivated nodes is independent of the states of their neighbors, and the accumulation of synergism has been neglected in these threshold models. Therefore, the scale of information diffusion is sensitive to the average degree of the network. Some studies show that parameter uncertainty may greatly affect influence maximization performance, and the interaction of combined nodes produces a collective influence that is larger than the sum of the individual nodes, which is called synergism [36–40]. Synergism is a ubiquitous phenomenon in social systems. Many studies have found that synergism enhances the transmission probability between a pair of nodes and promotes explosive spreading [29,41]. For example, in terms of information spread in social networks, a message transmitted by a group of connected users is more credible than a message transmitted by an individual [42,43]. Therefore, in this paper, we first discuss the catalytic role of synergism on the spreading dynamics in social networks and then propose a novel diffusion model called the synergism-based three-step cascade model (TSSCM) based on the above analysis and three-degree influence theory [35]. Finally, we develop an algorithm for solving the IM problem with the TSSCM. Experiments on five real networks demonstrate the efficacy of our method.

Synergism-based three-step cascade model

Definition of TSSCM

Without loss of generality, we define an unweighted, undirected graph $G = (V,E)$, where V is the set of vertices and E is the set of edges. An online social network can be modeled by the graph. A node $v \in V$ represents an individual in the social network, and an edge $e(u,v) \in E$ denotes that information can spread between u and v. The topology is represented by the adjacency matrix $\{A_{ij}\}_{N\times N}$, where $A_{ij} = 1$ if i and j are connected, and $A_{ij} = 0$ otherwise.

Many studies have found that synergism enhances the transmission probability and promotes explosive spreading [40]. Therefore, in our model, the probability that a seed node activates its neighbors is proportional to the number of other activated nodes connected to the seed node. Furthermore, some real information diffusion findings have supported the hypothesis that influence gradually dissipates and ceases to have a noticeable effect on people beyond the social frontier of three degrees of separation, which is called intrinsic decay [34,35,44,45]. Many research results on real social networks have confirmed this theory. Qin et al. [34] analyzed Sina Weibo retweet activities and illustrated that the retweet trees are small and shallow, and the average number of retweets decreases as the cascade depth increases. More than 96% of retweets are within three steps, and no retweet tree has deeper than 11 steps. Leskovec et al. [44] crawled blog links and found that more than 98.8% of the linked trees of all blogs have depths of less than three. Goel et al. [45] described the diffusion patterns arising from seven online social networks, including communications platforms, networked games and microblogging services, and found that most adoptions occur within a few steps of the seed node, even for the largest cascades observed. Based on the above research results, we consider the diffusion process within three steps and propose TSSCM.

In TSSCM, we suppose that node u can influence node v only if the distance from u to v is no greater than three. When u attempts to activate v, the activation probability $\alpha(u,v)$ is dependent not only on the number of activated neighbors connected to u but also the cascade depth

$d(d = 1,2,3)$.

$$\alpha(u, v) = p(u, v)l(d) \tag{1}$$

$$p(u, v) = 1 - (1 - \beta)^{1+\frac{m}{k-1}} \tag{2}$$

$$l(d) = \frac{1}{d}(d = 1, 2, 3) \tag{3}$$

where β is the basic spreading probability, m and k represent the number of activated neighbors connected to u and the degree of node u, respectively, $p(u,v)$ represents the synergism spreading probability, and $l(d)$ is the information decaying ratio, which is inversely proportional to d.

Eq (2) indicates that the larger the value of $\frac{m}{k-1}$, the higher the synergism spreading rate. Our model reduces to the classic decade model for $d = 1$ and $m = 0$, where $\alpha(u,v) = \beta$. If $m>0$, then $\alpha(u,v)>\beta$, which means synergism promotes information spread. In addition, $k>1$ for nonleaf nodes; thus, the synergistic ability of any activated neighbor of an active node is less than that of itself. This assumption is based on real disease propagation, where the probability that a susceptible node is infected by an infected direct neighbor is always greater than the probability of becoming infected from an infected indirect neighbor [36,37].

For $d>1$, we describe TSSCM as follows. Let $S_0 \subseteq V$ be the seed set. All nodes in S_0 are activated in the first time step. In the cascade steps $0 \leq t \leq 3$, $S_t \subseteq V$ is the set of nodes that are activated at step t. At step t+1, each node $u \in S_t$ attempts to activate its neighbor node $v \notin S_t$ with probability $\alpha(u,v)$. If such activation is successful, then v changes state from inactive to active and remains in the active state. Each activated node has only one chance to activate its neighbors during the step immediately following its initial activation. The above cascade process is repeated until no nodes in the network can be activated or $t = 3$.

As shown in Fig 1, at step t, node 1 is a seed and the other nodes are inactive. For node 1, there are no active neighbors; thus, $m = 0$, $\alpha(1,2) = \alpha(1,3) = \alpha(1,4) = 1 - (1 - \beta)^{1+\frac{0}{3-1}} = \beta$, which means node 1 activates its neighbors with probability β. At step t+1, node 3 is activated by node 1. Because it has an active neighbor, node 3 will activate its neighbors, node 5 and node 6, with a large probability $\alpha(3,5) = \alpha(3,6) = 1 - (1 - \beta)^{1+\frac{1}{2}}$. Unlike other diffusion models, TSSCM accumulates synergism, i.e., active neighbors of an active node cooperate to spread information. This phenomenon is common in real social systems, such as microblogging retweeting [44], opinion propagation [30], and animal invasion [46].

Influence maximization problem under TSSCM

Given a seed set S, we use $\sigma_{TSSCM}(S)$ to represent the influential spread of S, which can be quantified as the number of activated nodes under TSSCM when the propagation process ends. The IM problem under TSSCM is defined as follows.

Definition 1 Given a network $G = (V,E)$, the IM problem under TSSCM aims to find a subset $S^* \subseteq V$, $|S^*| = k$ such that

$$S^* = \arg_s \max \sigma_{TSSCM}(S) \tag{4}$$

Kempe et al. [12] have proved that the IM problem is NP-hard using the maximum coverage problem. Inspired by this proof we consider a similar reduction method and prove that the IM problem under TSSCM is NP-hard.

Theorem 1: The IM problem under TSSCM is NP-hard.

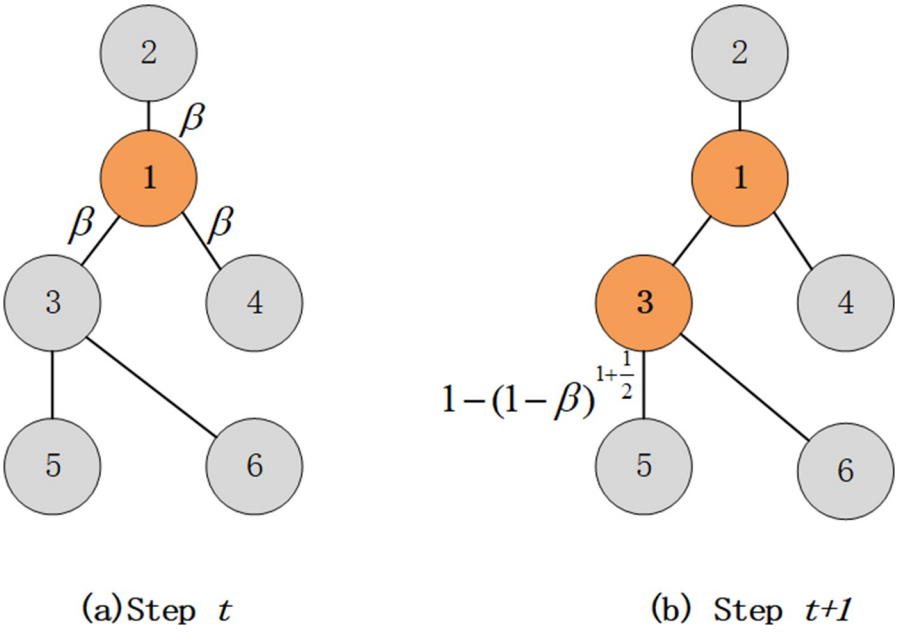

Fig 1. Illustration of TSSCM.

https://doi.org/10.1371/journal.pone.0221271.g001

Proof: The problem can be viewed as a Maximum Coverage problem, which is defined as follows:

Given a ground set $U = \{u_1,u_2,\ldots,u_n\}$ and a collection of subsets $S = \{S_1,S_2,\ldots,S_m\}$, where $S_i \subseteq U$ and $\bigcup_{i=1,2,\ldots,m} S_i = U$, we want to find k of the subsets $S' = \{S_1,S_2,\ldots,S_k\}$, $k<n<m$, where the union of $S_i \subseteq S'$,is equal to U. We show that the above description can be viewed as a special instance of the IM problem under TSSCM.

We define a directed bipartite graph containing n^2+m nodes. Node v_i and node v_j correspond to S_i and u_j, respectively. For each set S_i, there is a corresponding node v_i, and for each element u_j, there are n corresponding nodes $v_{j_1}, v_{j_1}, \ldots, v_{j_n}$. If $u_j \in S_i$, a direct edge $(v_i, v_{j_l}), l = 1, 2, \cdots, n$ exists with a spreading probability $p = 1$. We define X as a set of k of S_i and T as the union of the elements covered by $S_i \in X$, $T \subseteq U$. If k nodes corresponding to X are selected, they activate the nodes corresponding to elements in T; then, the number of active nodes is $k+n|T|$. Similarly, in the converse direction, the number of active nodes is also $k+n|T|$. In summary, we know that the maximum coverage instance $|T|$ element can be covered by k sets if and only if $k+n|T|$ nodes can be activated by k seeds in the instance, i.e., $\sigma_{TSSCM}(X) = k+n|T|$. In the instance, the longest path of the directed bipartite graph includes two nodes; thus, in TSSCM, $l(1) = 1$ and $l(2) = l(3) = 0$. k active seed nodes correspond to a maximum coverage solution due to information propagation to all other nodes corresponding to the ground set U. Thus, the maximum coverage problem is solved.

The optimal solution of the IM problem under TSSCM can be approximated by the greedy algorithm, $Greedy(G,\sigma_{TSSCM}(S),k)$, as shown in Algorithm 1.

Algorithm 1 $Greedy(G,\sigma_{TSSCM}(S),k)$
Input: G: network; k: size of seed set
Output: seed set S
initialize $S = \emptyset$
while $|S|<k$ do
  select $v = \arg\max_{v \in V}(\sigma(S \cup v)-\sigma(S))$;
  $S = S \cup v$;

```
5: end while
6: return S;
```

The approximation ratio of the greedy algorithm can reach $1-\frac{1}{e}\approx 0.63$

To optimize the global function of the IM problem, Flaviano Morone [30] mapped information spread asymptotically onto the optimal percolation and proposed another topological centrality measure called CI, which is defined as

$$CI_l(i) = (k_i - 1)\sum_{j\in\partial Ball(i,l)}(k_j - 1) \tag{5}$$

where k_i is the degree of node i, and $\partial ball(i,l)$ denotes the set of nodes at a distance l from node i.

The above CI does not consider the spreading rate between two linked nodes. However, in the actual information spreading process, a node receives information transmitted by other neighbor nodes with a certain probability. Clearly, a realistic and efficient algorithm for optimal resource allocation should consider both the topological characteristics and the details of the dynamics; additionally, propagation should be maximized within a limited time window. Because TSSCM is inherently probabilistic, we proposed a measure, called three-layer collective influence with synergism (CI_TLS), that incorporates CI formulation and spreading dynamics with synergism. For node i, $l(i,v)$ denotes the shortest distance from i to v. The spreading influence of node i is confined to a node set that consists of the nodes at a distance l from i, $L(i,l) = \{v|l(i,v) = l, v\in V$. We assign to node i the CI_TLS following Eq (6):

$$CI\_TLS(i) = \sum_{v\in L(i,l)}\sigma(i,v)\ \ l = 1,2,3 \tag{6}$$

$$\sigma(i,v) = 1 - \prod_{\substack{u\in L(i,l-1)\\ A_{uv}=1}}(1-\sigma(i,u)\alpha(u,v)) \tag{7}$$

where $\alpha(u,v)$ is the activation probability defined in Eq (1). $\sigma(i,v)$ is the final activation probability of node v by node i, which is obtained by recursively calculating the influence propagation. The CI_TLS of a node contains rich topological information and propagation dynamics, which can tell us more about the roles of the nodes in the network than a measure that considers only one aspect. In Eq (6), the sum contains the contributions of nodes whose distance to i is less than 3. Therefore, a node located at the center of a cluster with many links would have a large CI_TLS, even if it has a low degree. Thus, these low-degree nodes with the bridging property outrank those with a large degree but mediocre peripheral location. Fig 2 provides an illustrative example. We set node 1 as the initial seed, $\beta = 0.5$. The number next to node v is the value of $\sigma(1,v)$. The arrows denote the direction of information spread. The calculation is as follows.

$$\begin{aligned}
&\sigma(1,6) = 1-[1-\sigma(1,3)\alpha(3,6)][1-\sigma(1,4)\alpha(4,6)]\\
&= 1-[1-0.5*(1-(1-\beta)^{1+\frac{1}{2-1}})*\frac{1}{2}][[1-0.5*(1-(1-\beta)^{1+\frac{1}{2-1}})*\frac{1}{2}]\\
&= 1-\frac{13}{16}*\frac{13}{16}\\
&= 0.3398
\end{aligned}$$

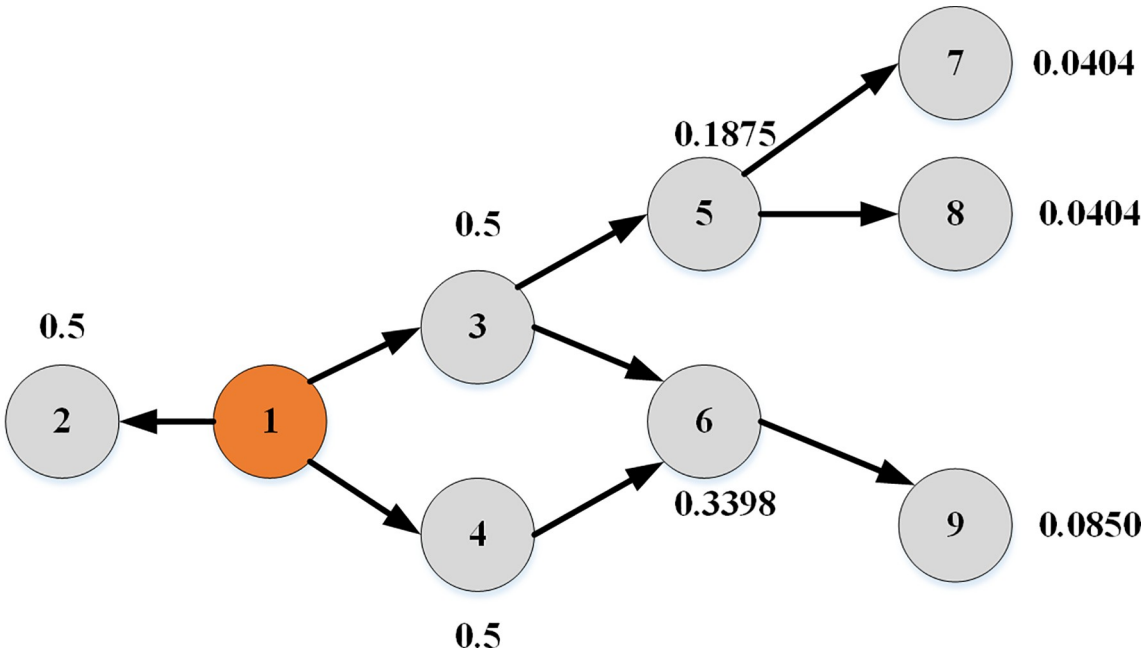

Fig 2. The final activation probabilities of the nodes.

https://doi.org/10.1371/journal.pone.0221271.g002

Node 6 has a larger activation probability than node 5 because node 6 is influenced by two nodes while node 5 is influenced by one node. Similarly, the activation probability of node 9 is larger than those of nodes 7 and 8 because node 6, which precedes node 9, has two activated neighbors, while the preceding node of nodes 7 and 8 has one activated neighbor. Therefore, the synergistic influence on node 9 is greater than that on nodes 7 and 8. $CI\_TLS(1)$ is the sum of $\sigma(1,i)$,i = 2,3,4,5,6,7,8,9, and it is used to measure the spreading influence of node 1 in CL_TSL.

We propose an adaptive CI_TLS algorithm based on the greedy approach to obtain a scalable algorithm. Define $N(i,4)$ as the set of node i plus those nodes with a short distance to i of less than 4. The details are shown in Algorithm 2.

```
Algorithm 2 CI_TLS(G,k)
Input: G: network; k: size of seed set
Output: seed set S
1: initialize S = ∅
2: Calculate CI_TLS(i) for each node i
2: while |S|<k do
3:    select i = arg max_{i∈V} CI_TLM(i);
4:    S= S∪i;
5: Remove N(i,4) and decrease the degree of N(i,4)'s neighbors by 1.
6: Update CI_TLS(i) for all nodes
7: end while
8: return S;
```

Note that we remove $N(i,4)$ once i is added to the seed set S (line 5 in algorithm 2) because seed i activates $N(i,4)$; thus, the nodes in $N(i,4)$ do not have to be selected in the later calculation. Ideally, $N(i,4)$ can be identified during the computation of $CI\_TSL(i)$ without additional time. The above operation overcomes the defect of the traditional algorithm, where the influence areas of the selected seeds overlap.

Computational complexity analysis of our algorithm

Next, we demonstrate the efficiency of our algorithm by investigating its computational complexity. In a network with N nodes, to compute the CI_TLS of a node, we must iteratively traverse its neighbors within a finite search radius, which costs $O(\langle k \rangle)$, where $\langle k \rangle$ is the average degree of the network. Because $k \ll N$, the result is $O(1)$. CI_TLS must be calculated for every node in the first step. However, during later steps, we have to recalculate the values of only the nodes within l+1 layers of the removed nodes. As verified in reference [31], the computational complexity of the above problem is $O(1)$, compared to $N \to \infty$. Sorting $CI\_TSL(i)$ requires $O(N\log N)$, and we select nodes until the seed set includes k nodes; therefore, the total computational complexity of our algorithm is $O(kN\log N)$, which ensures that our algorithm is scalable to large networks.

Experiments

First, we analyze the retweet data from Sina Weibo and verify that the proposed synergism-based TSSCM can effectively simulate real propagation trends. Then, we compare the spreading influence of the seed sets obtained by the five IM algorithms under TSSCM and the independent cascade model (ICM) to test the effectiveness of the CI_TLS algorithm. Finally, we list the CPU times of the five algorithms.

Analysis of the real diffusion depth

We obtained the post and retweet data from May 3–11, 2014, based on the Sina Weibo API (http://open.weibo.com). In accordance with the breath-first strategy, we crawled 50 messages posted by a user, and for each message, crawled the retweet users and added them to the crawling queue. After processing a user, we removed the user from the queue to reperform the same operational process and loop back and forth. Finally, we randomly selected 2000 retweet trees.

We counted the fraction of retweet trees at each depth. The results are shown in Fig 3. We can observe that the retweet trees are all small and shallow, and the number of retweet trees decreases as the cascade depth increases. Fig 3 shows that most of the cascades are within three steps and that less than 1% of retweet trees have a depth beyond 3. The fraction of retweet trees deeper than 8 is only 0.01% of all trees. Other researchers obtained similar conclusions, such as in references [34,35,44,45].

Datasets used in the experiments

Table 1 lists the empirical networks used to evaluate the effectiveness and efficiency of our proposed algorithm, namely, Blog, DBLP, Email, Epinions, Twitter and Livejournal, all of which can be downloaded from http://networkrepository.com[47].

In Table 1, n and m are the total numbers of nodes and edges, respectively; $k_{\max}$ and $\langle k \rangle$ represent the maximum and average degrees, respectively; C is the clustering coefficient; and β_{th} is the epidemic threshold. In homogenous networks, $\beta_{th} = \frac{1}{\langle k \rangle}$, while in heterogeneous networks, $\beta_{th} = \frac{\langle k \rangle}{\langle k^2 \rangle}$ [26].

Baseline algorithm

We choose four algorithms, namely, CI, degree discount, MaxCoreCover and random, as the baselines for evaluating the performance of our algorithm.

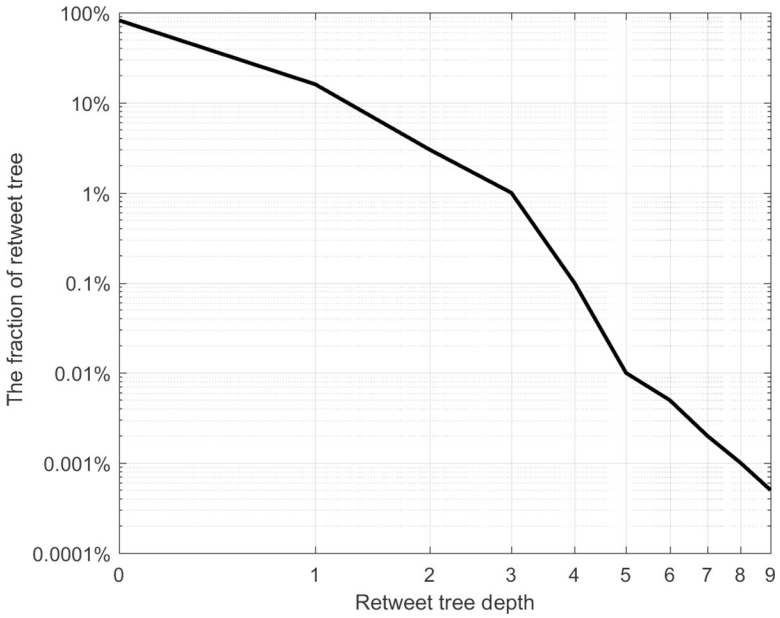

Fig 3. The fractions of retweet trees.

https://doi.org/10.1371/journal.pone.0221271.g003

CI: CI, which is defined as Eq (5), was proposed by Flaviano and Hernan A [30]. CI_l is adaptive and achieves the best performance for l = 3,4. In this paper, we choose CI_4 as a baseline algorithm.

Degree Discount: The degree discount algorithm is a heuristic based on degree centrality [23]. The node with the largest degree is selected as a seed, and the degrees of its neighbors are discounted by 1.

MaxCoreCover: This algorithm, which selects the node with the largest k-shell as a seed, was proposed by Kitsak [21]. When a node is selected, its neighbors can no longer be seeds.

Random: This algorithm randomly selects seed nodes.

Evaluation methodologies

The evaluation indicators we adopt for the IM algorithms are as follows: (a) the spreading influence of the seed set for TSSCM, (b) the spreading influence of the seed set for ICM, and (c) the computational time required by the IM algorithm to find the seed set.

The spreading influence of a seed set, which is used to evaluate the performance of an IM algorithm, is defined as the number of active nodes after the propagation process is complete.

Table 1. The statistical properties of the six empirical networks[a].

Network	n	m	k_{max}	$\langle k \rangle$	C	β_{th}
Blog	10K	326K	3992	64.78	0.0914	0.0018
DBLP	317K	1M	343	6.62	0.6350	0.0834
Email	1K	5K	71	9.62	0.2202	0.0535
Epinions	27K	100K	443	7	0.1351	0.0758
LiveJournal	4M	28M	3k	13	0.2600	0.0534
Twitter	405K	713K	626	3	0.014	0.1874

a. http://networkrepository.com

https://doi.org/10.1371/journal.pone.0221271.t001

The larger the spreading influence is, the more accurate the algorithm. In this paper, we first selected the seed set of each network according to the five algorithms. Then, we compared the spreading influence of different seed sets using five algorithms for each network under TSSCM and ICM. We have shown that the IM problem under TSSCM is NP hard; therefore, we run 10000 Monte Carlo simulations to obtain the results. The five measures are compared in Fig 4, which shows the spreading influence of the seed sets selected by these measures in six real networks. The x-axis represents the number of seeds obtained in the first step, and the y-axis represents the spreading influence of the five algorithms for a network, i.e., the number of active nodes after propagation is complete. The basic spreading probability $\beta = \beta_{th}$, and the values of β_{th} are listed in Table 1.

As expected, the seed sets obtained by CI_TLS result in the widest information spread, which means the performance of CI_TLS is the best. The trend lines of degree discount and CI are similar to those of CI_TLS because these three algorithms account for the number of neighbors when selecting seed nodes. The performance of CI ranks second among the five measures because CI_TLS considers the effect of synergies between nodes on the propagation probability while CI does not, which validates the rationality and importance of synergy. Degree discount does not consider dynamic attributes, such as the propagation path and the spreading probability between nodes, but considers static attributes, such as node degree, which results in performance that is inferior to those of CI_TLS and CI. However, degree discount performs much better than MaxCoreCover, which indicates that the degree of a node is an important indicator of the node's influence. In addition, the pruning strategy adopted in this measure ensures that the selected seed nodes do not gather in a local area of the network; this strategy is also used in CI_TLS. The random algorithm always has the worst results, indicating that careful seed selection is indeed important for effectively identifying influential nodes in many applications, such as marketing campaigns, epidemic prevention and maximization of information spread.

The spreading probability β is a parameter of TSSCM, and different values of β will result in different diffusion processes. Next, we compare the performance of different algorithms for the DBLP based on TSSCM with different β, where $\beta \in \{0.04,0.05,0.06,0.07\}$.

Fig 5 shows that the performance of CI_TLS is better than that of the other four algorithms over the entire range of β. At the same spreading probability, when the number of seeds is less than 5, the diffusion ranges of these algorithms are similar, except that of Random. Notably, the spreading influence of a few seeds is very limited. As the number of seeds increases, the performance gap of the five algorithms becomes obvious. Next, we analyse the results with different spreading probabilities. As β increases, the superiority of the three degree-based algorithms, CI_TLS, CI, and DegreeDiscount, becomes increasingly obvious, and CI_TLS always performs best. These results indicate that the degree is a key attribute of a node. CI_TLS considers and improves the degree measure by adding spreading distance constraints and a collaborative promotion mechanism; thus, it can mine influential nodes more effectively than can other other methods.

To further verify the performance of the proposed algorithm, we compare the spreading influence of seed sets based on ICM which is widely used in Influence maximization problem, and show the simulate results of the five algorithms in Fig 6. The CI_TLS shows its advantages over the whole range of β, $\beta \in \{0.04,0.05,0.06,0.07\}$. Formula (6) shows that a node with many activated neighbors can easily spread information to its inactive neighbors, that is, a node with many k-step connected neighbors would have a wide propagation range. The synergy mechanism in the CI_TLS makes it possible to effectively find such nodes, which is why the algorithm displays excellent performance.

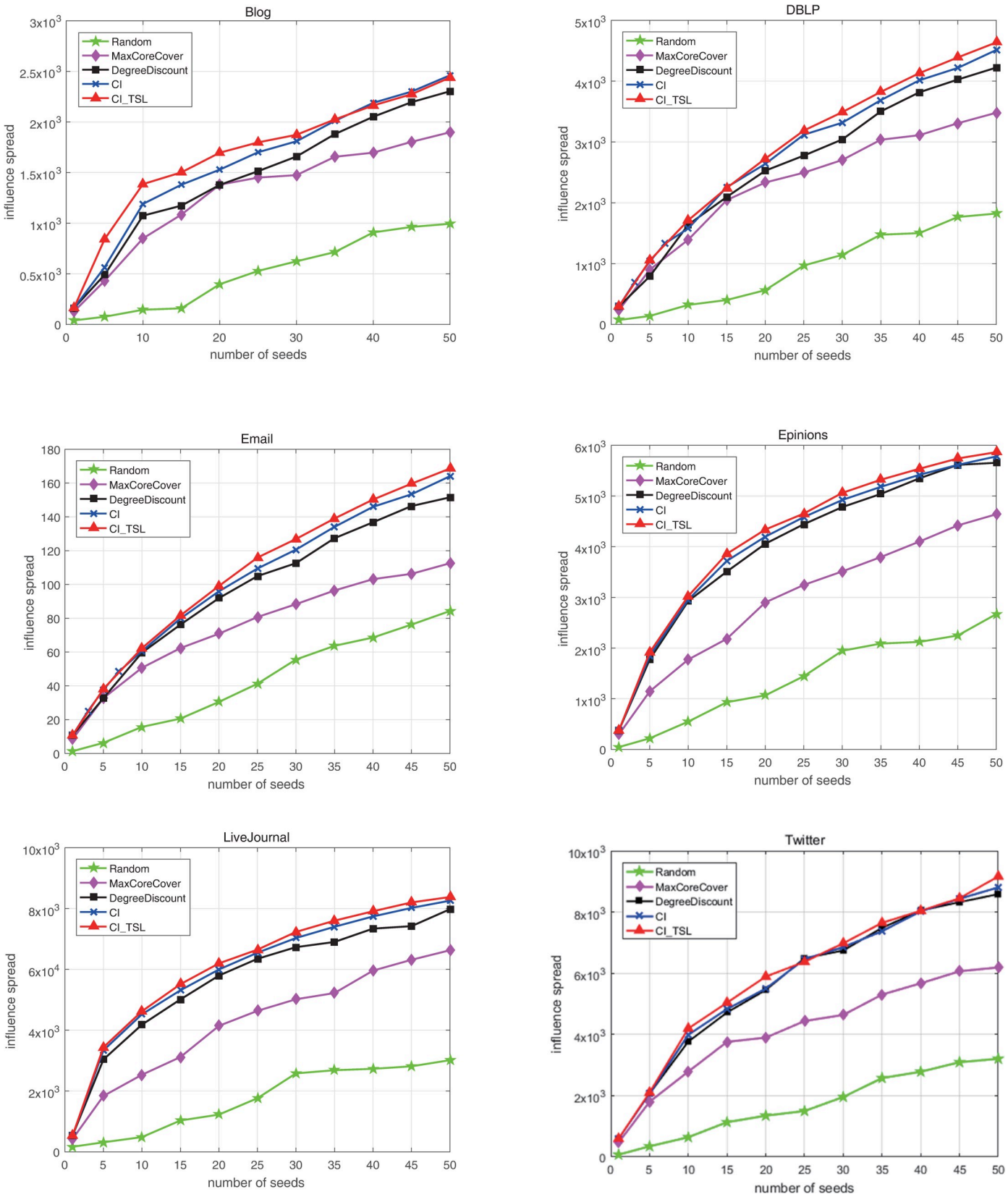

Fig 4. Spreading influence results of five algorithms for six networks.

https://doi.org/10.1371/journal.pone.0221271.g004

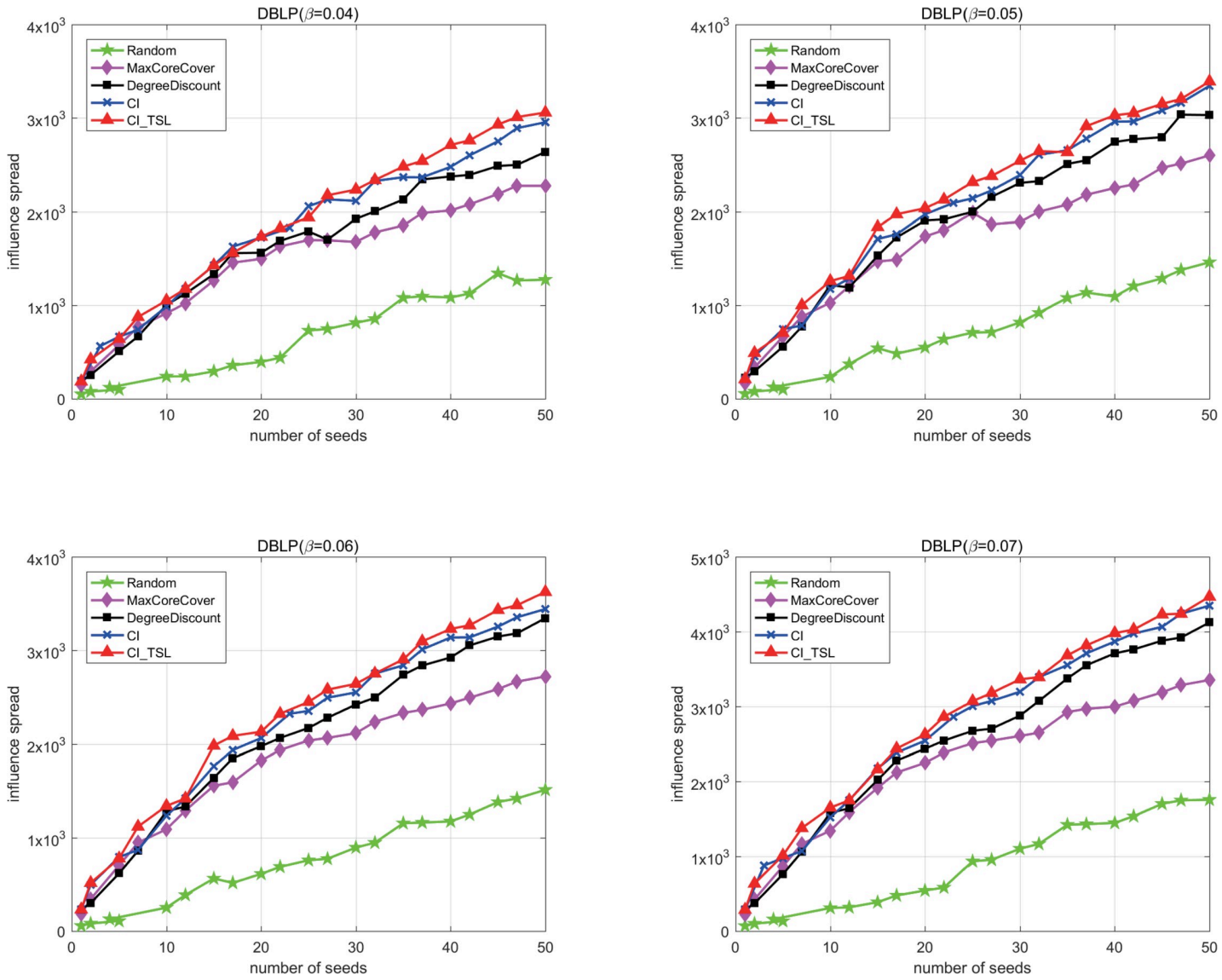

Fig 5. The spreading influence of different algorithms on the DBLP based on TSSCM with different β values.

https://doi.org/10.1371/journal.pone.0221271.g005

Finally, we compare the computational times of the five algorithms. The experiments are run on a server with a 4-core processor and 32 GB RAM using Python. Table 2 shows the computation times required by the five algorithms to find 50 seeds on six networks, i.e., Blog, DBLP, Email, Epinions, LiveJournal and Twitter. The computation time of our algorithm is longer than those of random, MaxCoreCover and DegreeDiscount, but it is almost equal to that of CI on all six real networks. Compared with CI, our method has a computation time increase of less than 2%. The CPU times of these algorithms are compared in Fig 7, where the x axis represents the algorithms and the y axis represents the computation times required to obtain 50 seeds. We can see that our algorithm is suitable for large-scale networks and can effectively mine influential nodes.

Overall, from the results shown in Figs 4, 5, 6 and 7, the proposed algorithm is more efficient in solving IM problems than the other four algorithms.

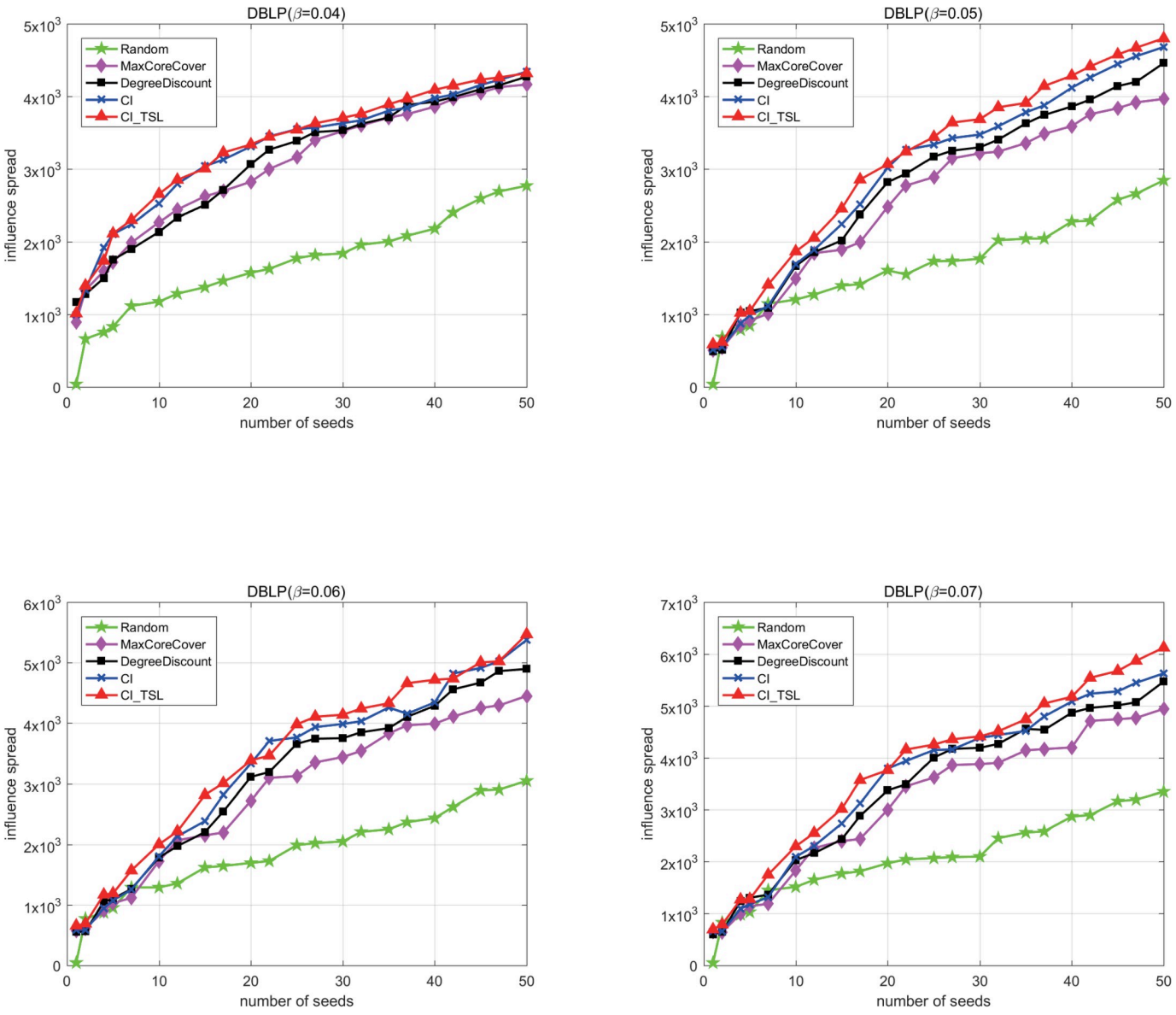

Fig 6. The spreading influence of different algorithms on the DBLP based on ICM with different β values.

https://doi.org/10.1371/journal.pone.0221271.g006

Table 2. The CPU times (in seconds) of five measures for six networks.

Network	Random	MaxCoreCover	DegreeDiscount	CI	CI_TLS
Blog	0.0201	72.5811	3.9826	52.6322	52.7215
DBLP	0.0231	180.9127	20.7354	2803.154	2820.4001
Email	0.0214	4.1752	0.4882	4.8861	4.9652
Epinions	0.0204	34.5144	9.6482	150.4257	152.8561
LiveJournal	0.0258	1975.6480	186.3461	43406.4718	43510.5842
Twitter	0.0226	120.1831	32.0974	3000.1289	3013.0084

https://doi.org/10.1371/journal.pone.0221271.t002

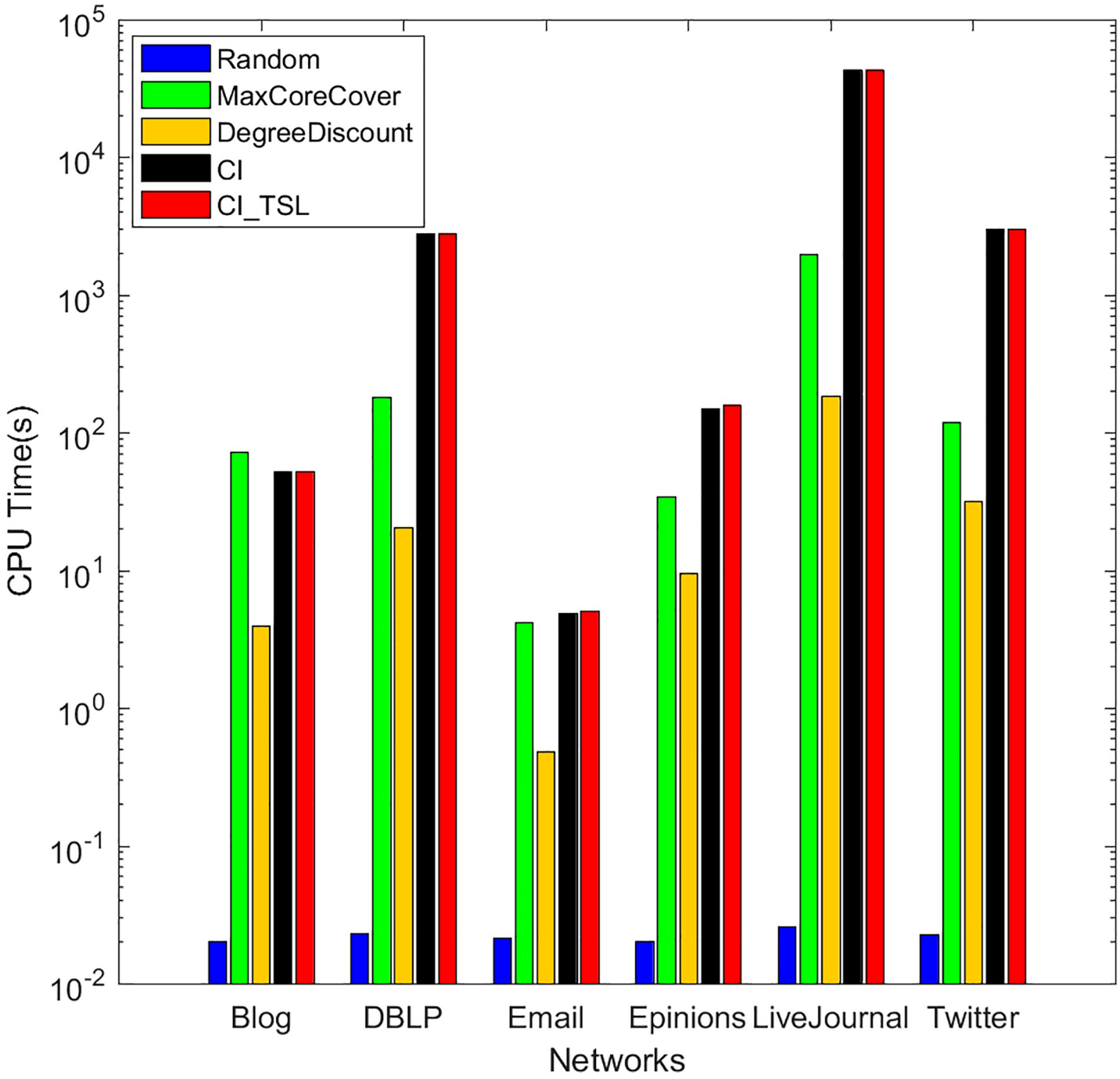

Fig 7. CPU times of five algorithms on six networks.

https://doi.org/10.1371/journal.pone.0221271.g007

Conclusion

Solving the IM problem is important for network analysis, information spreading, and other applications. A new diffusion model based on three degrees of influence theory and the catalytic role of synergism on spreading dynamics, namely, TSSCM, is proposed in this paper. In our model, the probability that a seed node activates its neighbors is proportional to the

number of activated nodes connected to the seed node, which is called synergism. Moreover, our model accurately simulates the cascade process of information transmission with finite steps. Inspired by the CI algorithm, we devised an algorithm for solving the IM problem under TSSCM, namely, CI_TLS. Compared with the CI algorithm, CI_TLS adds only the calculation of the activated neighbors of a node; therefore, the computation time increases only slightly, thereby balancing computational complexity and precision. The experimental results on six networks show that the CI_TLS measure is better than the other four algorithms tested, i.e., random, MaxCoreCover, degree discount and CI, and it achieves the best results for mining influential nodes. The seed sets obtained by CI_TLS result in the widest information spread. With the scale of social networks growing continuously, we can use parallel computing to accelerate the algorithm to effectively and efficiently solve the IM problem in large-scale networks. In many social networks, user behavior is affected by psychological factors, and a diffusion model with user decision making based on game theory would be appropriate [48–50]. Further work could track the IM problem under a diffusion model with psychological game theory. In reality, each individual in a network is always a user in the other networks. Resource diffusion impacts epidemics and information spread [51], and thus, a synergism-based diffusion model in multiple networks would be interesting and important to evaluate in future research.

Acknowledgments

The authors would like to thank the editors and reviewers for their insightful comments, which have helped improve the quality of this paper.

Author Contributions

Data curation: Xiaohui Zhao, Shuning Xing.

Formal analysis: Xiaohui Zhao.

Funding acquisition: Fang'ai Liu.

Investigation: Fang'ai Liu.

Methodology: Xiaohui Zhao, Fang'ai Liu.

Resources: Fang'ai Liu, Qianqian Wang.

Software: Shuning Xing, Qianqian Wang.

Supervision: Fang'ai Liu.

Visualization: Xiaohui Zhao.

Writing – original draft: Xiaohui Zhao.

Writing – review & editing: Xiaohui Zhao.

References

1. Valente T. W., Davis R. L. Accelerating the diffusion of innovations using opinion leaders. The Annals of the American Academy of Political and Social Science. 1999; 556(1),:55–67.
2. Domingos P., Richardson M. Mining knowledge-sharing sites for viral marketing. In Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. 2002:61–70.
3. Iyengar R., Van den Bulte C., Valente T. W. Opinion leadership and social contagion in new product diffusion. Market. Sci.2011; 30:195–212.
4. Watts D. J. A simple model of global cascades on random networks. Proc. Natl. Acad. Sci. USA2002; 99: 5766–5771. https://doi.org/10.1073/pnas.082090499 PMID: 16578874

5. Watts D. J., Dodds P. S. Influentials, Networks and public opinion formation.J. Cons. Res, 2007; 34:441–458.
6. Albert R., Jeong H., Barab´asi A.Error and attack tolerance of complex network, Nature. 2000; 406:378–382. https://doi.org/10.1038/35019019 PMID: 10935628
7. Yan S., Tang S., Fang W., Pei S., & Zheng Z. Global and local targeted immunization in networks with community structure, Journal of Statistical Mechanics Theory & Experiment, 2015; 8:1–11.
8. Morone F, Roth K, Min B, Stanley H.E, Makse H. A. Model of brain activation predicts the neural collective influence map of the brain, Proc Natl Acad Sci USA. 2017; 114(15):3849–3854. https://doi.org/10.1073/pnas.1620808114 PMID: 28351973
9. US Government Accountability Office (2012) Financial Regulatory Reform: Financial Crisis Losses and Potential Impacts of the Dodd-Frank Act (Government Accountability Office, Washington, DC)
10. Li Y, Fan J, Wang Y, et al. Influence Maximization on Social Graphs: A Survey, IEEE Transactions on Knowledge and Data Engineering. 2018; 30(10):1852–1872.
11. Domingos P, Richardson M Mining. The network value of customers. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 2001: 57–66.
12. Kempe D., Kleinberg J., and Tardos É. Maximizing the spread of influence through a social network. In Proc. ACM SIGKDD Int. Conf. Knowl.Discovery Data Mining. 2003:137–146.
13. Leskovec J., Krause A., Guestrin C., Faloutsos C., VanBriesen J., Glance N.S. Cost-effective outbreak detection in networks. In Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2007: 420–429.
14. Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009:199–208.
15. Pastor-Satorras R, Vespignani A. Immunization of complex networks. Phys RevE. 2002; 65:036104–036114.
16. Zhao X, Liu F, Wang J, Tianlai Li, Evaluating Influential Nodes in Social Networks by Local Centrality with a Coefficient. ISPRS International Journal of Geo-Information, 2017; 6(35):1–11.
17. Zhao X, Liu F, Wang Shuning Xing, Qianqian Wang. Identifying Influential Spreaders in Social Networks via Normalized Local Structure Attributes. IEEE Access, 2018; 6:66095–66104.
18. Cohen R, Havlin S, Ben-Avraham D. Efficient immunization strategies for computer networks and populations. Phys Rev Lett. 2003; 91:247901–247910. https://doi.org/10.1103/PhysRevLett.91.247901 PMID: 14683159
19. Holme P, Kim BJ, Yoon CN, Han SK. Attack vulnerability of complex networks. Phys Rev E. 2002; 65:056109–056120.
20. Holme P. Efficient local strategies for vaccination and network attack. Europhys Lett. 2004; 68:908–914.
21. Kitsak M, Gallos L K, Havlin S,Liljeros F, Muchnik L, Stanley H.E et al. Identification of influential spreaders in complex networks. Nature Physics. 2013; 6, (11): 888–893.
22. Shang J, Zhou S, Li X, Liu L, Wu H. CoFIM: A community-based framework for influence maximization on large-scale networks. Knowledge-Based Systems. 2017; 117:88–100.
23. Chen W, Lakshmanan LV, Castillo C. Information and Influence Propagation in Social Networks. Synthesis Lectures on Data Management, 2013; 5:1–10.
24. Cao J, Dong D, Xu S, Zheng X, Liu B, Luo J. A k-core based algorithm for influence maximization in social networks. Chinese Journal of Computers. 2015; 38(2):238–248.
25. Zhu Jianghua, Liu Yong, Yin Xuming. A New Structure-Hole-Based Algorithm for Influence Maximization in Large Online Social Networks. IEEE ACCESS, 2017; 7:23405–23413.
26. Wang X.F., Li X., Chen G.R. The importance and similarity of nodes. In Network Science: An Introduction, Higher Education Press: Beijing, China,2012:157–185.
27. Wang Wei, Tang Ming, Zhang Hai-Feng, and Lai Ying-Cheng. Dynamics of social contagions with memory of nonredundant information. Phys. Rev. E. 2015; 92: 012820–012832
28. Wang Wei, Cai Meng, Zheng Muhua. Social contagions on correlated multiplex networks. Physica A. 2018; 499: 121–128.
29. Wang Wei, Tang Ming. Stanley H. Eugene, and Braunstei Lidia A. Social contagions with communication channel alternation on multiplex networks. Phys. Rev. E. 2018; 98: 062320–062338.
30. Zheng Muhua, Linyuan Lü, and Ming Zhao. Spreading in online social networks: The role of social reinforcement. Phys. Rev. E. 2013; 88:012818–012824.
31. Morone Flaviano, Makse Hernan A. Influence maximization in complex networks through optimal percolation. Nature, 2015; 524: 65–75. https://doi.org/10.1038/nature14604 PMID: 26131931

32. Pei S, Teng X, Shaman J, Morone F, Makse H.A. Efficient collective influence maximization in cascading processes with first-order transitions. Scientific Reports. 2017; 7:45240–45255. https://doi.org/10.1038/srep45240 PMID: 28349988

33. AY Lokhov D Saad. Optimal deployment of resources for maximizing impact in spreading processes. Proceedings of the National Academy of Sciences of the United States of America.2017; 114(39): 8138–8150.

34. Qin Y, Ma J, Gao S. Efficient influence maximization under TSCM: a suitable diffusion model in online social networks. Soft Comput. 2017; 21:827–838.

35. Christakis NA, Fowler JH. Connected: The surprising power of our social networks and how they shape our lives. Little, Brown and Company, USA. 2009:220–256.

36. Centola D M, Macy M W, Eguíluz V M. Cascade dynamics of multiplex propagation. American Institute of Physics. 2005:200–200.

37. Goh KI, Lee DS, Kahng B, Kim D. Sandpile on scale-free networks. Physical Review Letters. 2003; 91 (14):148701–14710. https://doi.org/10.1103/PhysRevLett.91.148701 PMID: 14611564

38. Lockwood J L. Evolution of Concepts Associated with Soilborne Plant Pathogens. Annual Review of Phytopathology. 1988; 26(26):93–121.

39. Chen W, Lin T, Tan Z, Zhao MF, Zhou X. Robust Influence Maximization. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD "16,2016:795–804.

40. Ludlam J J, Gibson G J, Otten W, Gilligan C A. Applications of percolation theory to fungal spread with synergy. Journal of the Royal Society Interface. 2012; 9(70):949–959.

41. Liu Q H, Wang W, Tang M, Zhou T, Lai Y. Explosive spreading on complex networks: The role of synergy. Phys.rev.e, 2017; 95(4):042320–042330.

42. Gleeson J P. High-accuracy approximation of binary-state dynamics on networks. Physical Review Letters. 2011; 07(6):068701–068710.

43. Chen D., Lü L.; Shang M.S., Zhang Y.C.; Zhou T. Identifying influential nodes in complex networks. Phys. A Stat. Mech. Appl. 2012; 391(4):1777–1787.

44. Apte C, Skillicorn D, Liu B, Parthasarathy S. Patterns of Cascading Behavior in Large Blog Graphs. Proceedings of the Seventh SIAM International Conference on Data Mining, April 26–28, 2007, Minneapolis, Minnesota, USA. 2007.

45. Goel S, Watts D J, Goldstein D G. The structure of online diffusion networks. Proceedings of the 13th ACM Conference on Electronic Commerce. Valencia, Spain (2012.06.04–2012.06.08)2012:623.

46. Murray J D. Mathematical Biology. Biomath. 2002; 19(1–2):261–283.

47. Rossi R A, Ahmed N K. The network data repository with interactive graph analytics and visualization. Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015: 4292–4293.

48. Xiong X, Qiao S, Li Y, Zhang H, Huang P, Han N, et al. ADPDF: A Hybrid attribute discrimination method for psychometric data with fuzzinese. IEEE Transactions on Systems Man & Cybernetics Systems. 2018; 6(27):99–113.

49. Xiong X, Li Y, Qiao S, Han N, Wu P, Peng J, et al. An emotional contagion model for heterogeneous social media with multiple behaviors. Physica A Statistical Mechanics & Its Applications. 2018; 49:185–202.

50. Wei Sheng, Shuqing N.Teng, Hui-jia Li. Hierarchical structure in the world's largest high-speed rail network. Plos One. 2019, 3:1–10.

51. Xiaolong C, Wei W, Shimin C, Stanley H.E, Braunstein L A. Optimal resource diffusion for suppressing disease spreading in multiplex networks. Journal of Statistical Mechanics: Theory and Experiment. 2018; 5:053501–053520.