Identify Multiple Seeds for Influence Maximization by Statistical Physics Approach and Multi-hop Coverage

Fuxuan Liao (s2060004@jaist.ac.jp)
Japan Advanced Institute of Science and Technology
Yukio Hayashi
Japan Advanced Institute of Science and Technology

Research Article

Keywords: Influence maximum problem, Multiple seeds, Vertex cover problem, L-hop coverage, Overlapping phenomena, SIR model, Statistical physics approach

Posted Date: April 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1510470/v1

RESEARCH

Identify Multiple Seeds for Influence Maximization by Statistical Physics Approach and Multi-hop Coverage

Fuxuan Liao* and Yukio Hayashi

*Correspondence:
s2060004@jaist.ac.jp
Division of Transdisciplinary Sciences, Japan Advanced Institute of Science and Technology, Nomi, Japan
Full list of author information is available at the end of the article

Abstract

Finding the influential nodes as seeds in a real network is an important problem which relates to wide applications. However, some conventional heuristic methods do not consider the overlap phenomenon. In order to avoid the overlap of spreading, we propose a new method in combing the statistical physics approach and multi-hop coverage. We also propose a faster epidemic model which does not need the averaging of stochastic behavior. Through the computer simulation, the obtained results show that our method can outperforms other conventional methods in the meaning of stronger spreading power per seed.

Keywords: Influence maximum problem; Multiple seeds; Vertex cover problem; L-hop coverage; Overlapping phenomena; SIR model; Statistical physics approach

Introduction

Influence maximum problem (IMP) is an optimization problem for finding a small subset of influential nodes as k seeds which maximize the influence represented by the number of activated nodes from the seeds in a social network, $k \geq 1$ is a constant number. The problem has many applications such as the viral marketing [1], brain activation [2], information dissemination [3], and halting global epidemic outbreaks in contact networks [4]. For the maximization, a diffusion model is studied to simulate information propagation from active individuals. The typical models are called Independent Cascade (IC) and Linear Threshold (LT) models [5]. Note that a special case of IC with a constant infection probability on every links is the susceptible-infected-recovered (SIR) model [6] as mentioned later. However, the IMP is NP-hard [7] for both IC and LT models [6]. Thus, many researchers have designed heuristic methods for finding single or multiple seeds by using local or global network properties with spreading power, such as degree centrality [8, 9], k-core [10], local centrality [11], local structure centrality [12], and collective influence [13]. However, there are various problems for these heuristic methods. For example, degree centrality is a straightforward and efficient method, however it considers only the power of direct infections. When two hubs are adjacent to each other, the spreading areas overlap heavily. Although some well-known global methods such as betweenness centrality (BC) [14] and closeness centrality (CC) [15] can give better results [16] for finding multiple seeds, they are unsuitable for very large-scale social networks because of the high computational complexity [17] of $O(\mid V \mid \times \mid E \mid)$ for BC or $O(\mid V \mid^2)$ for CC. Where V and E denote sets of nodes and edges, respectively.

Thus, the design of more effective method is still an open issue especially for finding multiple seeds.

On the other hand, we consider avoiding the overlap by using l-hop coverage to find multiple seeds. The l-hop coverage means that seeds infect their l-hop neighbors. The set cover, dominating set, and the vertex cover problems are corresponded to 1-hop coverage. Similarly, IMP can be corresponded to an extension of the minimum set cover or domination set problem [7] by using l-hop coverage (l is greater than 1), in which seeds infect their connecting neighbors, next neighbors or next-next neighbors, and so on. The minimization is corresponding to that it is effective to maximize the number of activated nodes in the propagation from seeds chosen as few as possible. Moreover, the minimum set cover problem can be reducted to the minimum vertex cover problem [7]. However, the minimum vertex cover problem is NP-hard [7]. In order to efficiently estimate the set of the minimum vertex cover with global spreading power, we focus on collective computation by local interactions through message-passings based on statistical physics [18].

In this paper, we propose a new method to approximately solve the IMP problem in combining the statistical physics approach for the minimum vertex cover [18] and l-hop coverage. Usually, the SIR model [6] is applied to perform the spreading process from multiple seeds, however many trials of spreading is necessary for the averaging of stochastic behavior. When a network is very large, the conventional SIR model requires a lot of time in the averaging of stochastic behavior. To reduce the calculation time, we propose a faster SIR model inspired from the collective influence [13]. It does not need the averaging of behavior from samples with initial random settings, therefore it is expected to be the number of samples times faster than the conventional SIR model.

The organization of this paper is as follows. In section "Heuristic methods", we explain the conventional heuristic methods for a IMP. In section "Methodology", we briefly review the statistical physics approach and propose our method. The conventional SIR model and faster SIR model are introduced in the subsection "Message-passing for the SIR model". Through computer simulation, the spreading power of our method and other heuristic methods are compared in the subsection "Performance on faster MP-SIR model". Conclusion are given in the last section.

Heuristic methods

In considering a IMP, we explain the following widely-used heuristic methods, whose spreading power are compared with that by our method in the next section.

High Degree

The High degree (HD) method selects k nodes in decreasing order of degrees as the influential seeds [8, 9]. It needs only the local topological properties from the connecting nearest neighbors. Therefore, it is simple and efficient for finding seeds.

k-core

In k-core method [10], seeds are ranked according to their k_s values, which are calculated through the k-shell decomposition. In the k-shell decomposition, nodes are removed iteratively. Firstly, leaves with $k_s = 1$ are removed. This pruning is

repeated until there is no leaves. The peripheral k-shell with index $k_s = 1$ consists of a set of removed nodes. Similarly, the next k-shells with index $k_s \leftarrow k_s + 1$ are extracted, the nodes located within the core have the highest k_s values. Actually, in the k-shell decomposition, all nodes are divided into shells. In comparison with the peripheral nodes, the core nodes tend to involve larger spreading from them. Therefore, the node in the core with the largest k_s is defined as a seed.

Local Centrality and Local Structure Centrality

The HD is simple and efficient, however it neglects the global network properties. When the neighbors of a hub are leaves, the peripheral hub has weak spreading power only for a moment. In contrast, betweenness (BC) and closeness (CC) centrality consider the global information, while their calculations are slightly complicated. Thus, Local centrality (LC) considers a trade-off between locality and time-consuming for the calculation [11]. The LC is defined as

$$Q(u) = \sum_{w \in \partial u} \mid \partial Ball(w, 2) \mid,$$

$$C_L(v) = \sum_{u \in \partial v} Q(u),$$

where ∂u denotes the set of the nearest neighbors of node u, $\partial Ball(w, 2)$ denotes a set of nodes within 2 hops from node w as shown in Fig. 1. $\mid \mid$ denotes its size. As a seed, v is selected in decreasing order of $C_L(v)$. Note that LC gives similar spreading power as good as the closeness centrality [11].

In addition, Local structure centrality (LSC) is an extension of LC [12]. The LSC is defined by the linear interpolation of local clustering coefficient C_w [19] and LC with a tunable balance parameter $0 \leq \alpha \leq 1$.

$$Q(u) = \alpha \mid \partial Ball(u, 2) \mid + (1 - \alpha) \sum_{w' \in \partial Ball(u,2)} C_{w'},$$

$$C_L(v) = \sum_{u \in \partial v} Q(u),$$

where $\partial_2 u$ is a set of the next nearest neighbors of node u. As mentioned in [12], we set $\alpha = 0.7$

Collective Influence

Collective influence (CI) aims to find the minimum set of nodes for the IMP as follows [13]. At the origin $\{v_{i \to j}\} = \{0\}$, the stability of nonlinear message-passing equation

$$v_{i \to j} = n_i \left[1 - \prod_{k \in \partial i \backslash j} (1 - v_{k \to i}) \right], \qquad (1)$$

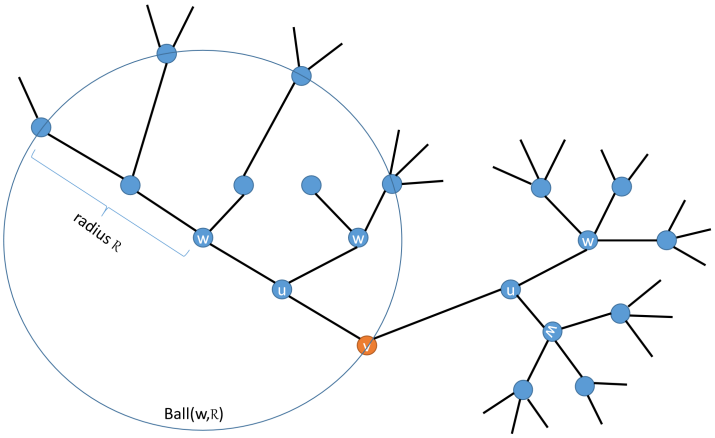


Figure 1 A set of nodes from 2-hops from node w in the process of LC.

is determined by the largest eigenvalue of the Jacobian matrix $\left[\frac{\partial v_{i\to j}}{\partial v_{k\to l}}\right]$. In other words, when the largest eigenvalue is less than 1, the spreading is stopped by removing a set of nodes $\{i : n_i = 0\}$ as influences. Thus, by using a greedy algorithm to minimize the eigenvalue, CI is derived [20] through a power method for each node i,

$$CI_{\mathcal{R}}(i) = (k_i - 1) \sum_{j \in \partial Ball(i,\mathcal{R})} (k_j - 1),$$

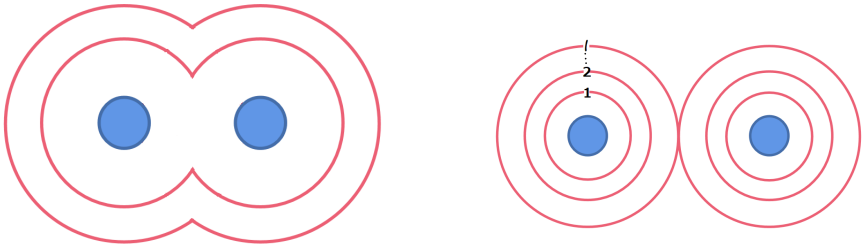
where $\mathcal{R}$ is the radius of the ball. The highest $CI_{\mathcal{R}}(i)$ is selected as a seed. After removing the node i, $CI_{\mathcal{R}}(i')$ is recalculated for the remaining nodes $i' \in V$ in the network. It needs only local topological structure within the ball of the radius $\mathcal{R}$ instead of the whole network.

Methodology

Although the conventional heuristic methods [8, 9, 10, 11, 12, 13] can be applied to find multiple seeds, they do not consider the overlap phenomena. As shown in Figure 2 (a), when hubs are near to each other, HD method is not suitable for finding multiple seeds. Because their spreading areas heavily overlap. In order to avoid the overlap, we consider a new method inspired from a statistical physics approach and l-hop coverage.

Influence maximization by survey propagation

Firstly, we consider the l-hop coverage. The set cover, dominating set, and the vertex cover problems in combinatorial optimization [7] are corresponded to 1-hop coverage. Similarly, finding the optimal multiple seeds can be corresponded to an extension of minimum set cover or dominating set problem by using l-hop coverage (l is greater than 1), in which seeds infect their connecting neighbors, next neighbors or next-next neighbors, and so on. Furthermore, a set cover problem can be reducted



(a)Close case (b)Distant case

Figure 2 Two situations of (a) overlap and (b) avoiding the overlap.

to a vertex cover (VC) problem [7]. However, the minimum VC problem is NP-hard. In order to efficiently estimate a set of nodes as the minimum VC with global spreading power, we focus on collective computation by local interactions through message-passings based on statistical physics approach [18].

We briefly review the approximate algorithm called survey propagation for the minimum VC problem. In the algorithm [18], each node i has one of the three states: covered (state 1), never covered (state 0), or sometimes covered and sometimes not (joker state $*$). Note that the joker state $*$ is between the state 0 and 1. The number of covered states can be regulated in the extended search space by introducing joker state. That is the reason why it is called survey propagation. These probabilities are denoted as $\hat{\pi}^{(1)}_{j\to i}$ (state 1), $\hat{\pi}^{(0)}_{j\to i}$ (state 0), and $\hat{\pi}^{(*)}_{j\to i}$ (joker state $*$), respectively. We should remark that the following message-passing for estimating the minimum VC differs from information spreading on SIR model [6]. For each node i, the message-passing equations [18] are given by

$$\hat{\pi}^{(0)}_i = C_i^{-1} \prod_{j\in\partial i} (1-\hat{\pi}^{(0)}_{j\to i}),$$

$$\hat{\pi}^{(*)}_i = C_i^{-1} e^{-y} \sum_{j\in\partial i} \hat{\pi}^{(0)}_{j\to i} \prod_{j'\in\partial i\backslash j} (1-\hat{\pi}^{(0)}_{j'\to i}),$$

$$\hat{\pi}^{(1)}_i = C_i^{-1} e^{-y} \left[1 - \prod_{j\in\partial i} (1-\hat{\pi}^{(0)}_{j\to i}) - \sum_{j\in\partial i} \hat{\pi}^{(0)}_{j\to i} \prod_{j'\in\partial i\backslash j} (1-\hat{\pi}^{(0)}_{j'\to i}) \right], \qquad (2)$$

where $\partial i\backslash j$ is the set of the nearest neighbors of node i but not including j, e^{-y} is a penalty factor for minimizing the size of VC, y is an inverse temperature parameter. The normalization constant is given by

$$C_i = e^{-y} \left[1 - (1-e^{y}) \prod_{j\in\partial i} (1-\hat{\pi}^{(0)}_{j\to i}) \right]. \qquad (3)$$

For each link $i \to k$, the probability is also given by

$$\hat{\pi}^{(0)}_{i\to k} = C^{-1}_{i\to k} \prod_{j\in\partial i\setminus k} (1 - \hat{\pi}^{(0)}_{j\to i}), \qquad (4)$$

$$C_{i\to l} = e^{-y} \left[1 - (1 - e^{y}) \prod_{j\in\partial i\setminus k} (1 - \hat{\pi}^{(0)}_{j\to i}) \right]. \qquad (5)$$

Equations (2)$\sim$(5) are calculated after T round iterations (until convergence). The node i with the largest $\hat{\pi}^{(1)}_i$ is selected as the VC. Then it is removed and recalculate the $\hat{\pi}_i$ until all nodes are covered in the following decimation process on the l-hop coverage. Our method is proposed as

Step 1 By using Eqs.(2)$\sim$(5), the probability $\hat{\pi}^{(1)}_i$ of node i is calculated for estimating the minimum VC.

Step 2 As the decimation process, the node j with the highest $\hat{\pi}^{(1)}_j$ is selected as a seed, the chosen node j and its $\partial Ball(j, l-1)$ are removed from the network. We emphasize that the $\partial Ball(j, l-1)$ is represented the l-hop coverage. The number of seeds are updated as $N_s \leftarrow N_s + 1$ (initially set as $N_s = 0$).

Step 3 Repeat Steps 1 and 2 until all nodes have been removed in the network. Finally, the size of multiple seeds is obtained as N_s.

Table 1 shows that the solution by our method seems to be nearly optimal, while that by 2-approximation method [21] almost double size of the optimal solution. Note that the survey propagation is a statistical physics approach, the 2-approximation method is a computer science approach with guaranteed accuracy of the size at most twice. For comparing with the survey propagation, we also apply the belief propagation algorithm [22] to estimate the VC from the feedback vertex set (FVS). Because the minimum FVS can be reducted to the minimum VC. However, after removing the FVS, the remaining part of network becomes trees (forest). As shown in Fig. 3, we apply the well known method [23] to divide the tree into odd and even layer, and select one of the layers (odd or even) whose size is smaller as VC. Table 1 shows that the size of VC by the survey propagation is slightly better than that by the belief propagation. When the inverse temperature parameter is set as $y = 7$, the result of the minimum VC is the best of the minimum size. Moreover, Table 2 shows that $T = 50$ round gets the best result for the minimum VC. Therefore, we apply the survey propagation with $y = 7$ and $T = 50$ in the following part.

Inverse temperature parameter y	0	0.5	1	2	3	5	7
|VC| by the survey propagation	3520	3510	3517	3510	3508	3511	**3507**
|VC| / N	0.462	0.460	0.461	0.460	0.460	0.461	**0.460**
|VC| by the belief propagation	3520	3514	3516	3519	3515	3523	3524
|VC| / N	0.462	0.461	0.461	0.461	0.461	0.462	0.462
|VC| by the 2-approximation				5498			
|VC| / N				0.721			

Table 1 |VC| by the survey propagation plus assignment for the remaining tree except FVS versus |VC| by the belief propagation and 2-approximation under different inverse temperature parameter y for the social network lastFM with N=7624. The bold numbers are the best results with the minimum VC.

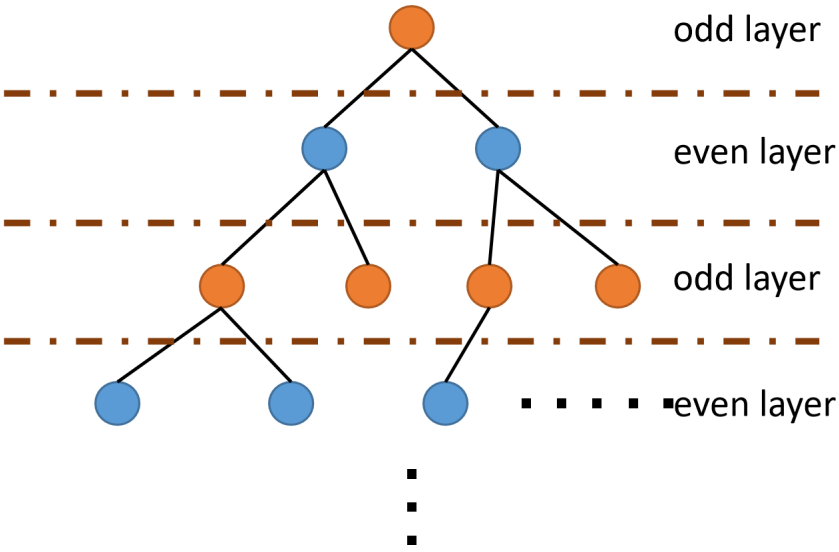


Figure 3 Well-known assignment method [23] for finding VC in a tree.

Round T	5	10	20	50	100	200
|VC| by the survey propagation	3511	3512	3519	**3508**	3516	3516
|VC| / N	0.461	0.461	0.462	**0.460**	0.461	0.461

Table 2 |VC| by the survey propagation under different round T for the social network lastFM with N=7624.

Message-passing for the SIR model

Let us consider the averaging behavior in a stochastic SIR epidemic model (we call it AVG-SIR) [6] with three states S: susceptible (inactive) nodes represents the individuals susceptible to the disease, I: infected (active) nodes denotes the individuals that have been infected and are able to spread the disease to susceptible individuals, and R: recovered stands for individuals that have been recovered and will never be infected again [6]. At each time step, in the spreading process, an infected node changes the states of its neighbors from S to I with probability $\beta = \lambda \frac{\langle k \rangle}{\langle k^2 \rangle}$ [24], and then changes its own state from I to R with recovery probability $\mu = 1$. Usually, the conventional AVG-SIR model is applied to perform the spreading process from a set of nodes as multiple seeds, however many trials of spreading is necessary for the averaging of stochastic behavior. It means that if the size of the network is very large, AVG-SIR model requires a lot of time for the averaging. We set sample size = 1000. In order to reduce the calculation time, we consider the following message-passing equations inspired from that in CI.

$$P_i^I(t+1) = P_i^S(t) \left[1 - \prod_{j \in \partial i} (1 - \beta P_j^I(t)) \right], \qquad (6)$$

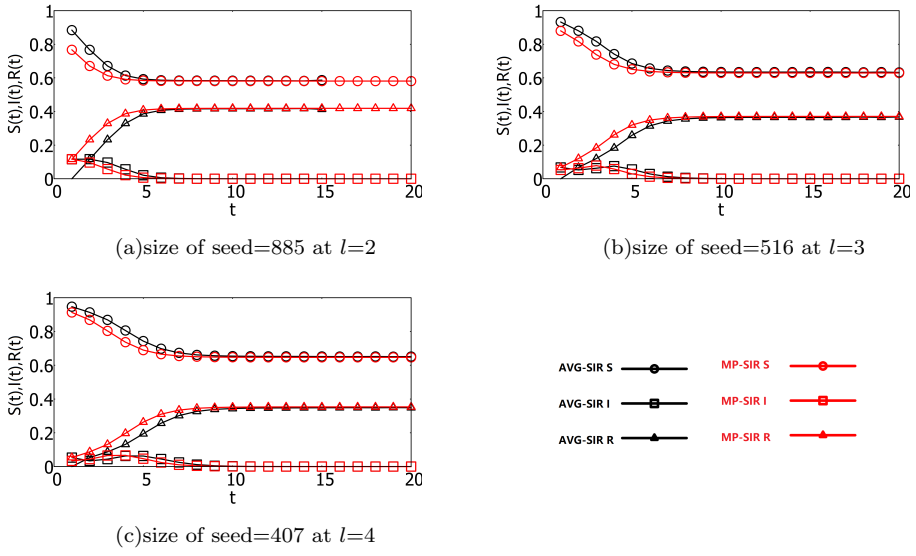
$$P_i^R(t+1) = P_i^R(t) + P_i^I(t),$$

$$P_i^S(t+1) = 1 - P_i^I(t+1) - P_i^R(t+1),$$

where $P_i^I(t+1)$, $P_i^R(t+1)$, and $P_i^s(t+1)$ denote the probabilities of states I, R, and S for node i at time $t+1$, respectively. We call it MP-SIR model.

In Fig. 4, we show the spreading power on AVG-SIR and MP-SIR models for three different sizes of seed 885 ($l = 2$), 516 ($l = 3$), and 407 ($l = 4$) with $\beta = 0.12$. The rate of seeds are $N_s/N = 0.12$ ($l = 2$), 0.07 ($l = 3$), and 0.05 ($l = 4$), respectively. Here $S(t) = \sum_{i=1}^{N} P_i^S(t)/N$, $I(t) = \sum_{i=1}^{N} P_i^I(t)/N$, $R(t) = 1 - S(t) - I(t)$, and N_s denotes the size of seeds. Note that $N_s/N \leq 20\%$ is realistic [10]. In Fig. 4 (a)(b)(c), $I(t)$ monotonically increases, decreases, and finally converges to zero. $R(t)$ monotonically increases and converges to 0.4. $S(t)$ monotonically decreases and converges to 0.6. The black lines with circle, square, and triangle marks denote the probabilities of state S, I, and R on AVG-SIR model. The red lines with circle, square, and triangle marks denote the probabilities of state S, I, and R on MP-SIR model, respectively. Although the size of seeds is different, the red and black lines of each state S, I, and R on MP-SIR and AVG-SIR models are almost coincided. In addition, $R(t)$ converges to 0.4 in (a), 0.37 in (b), and 0.35 in (c) for $t^* > t_c$ (t_c: it is defined at the convergent time, when all infected nodes are recovered.). Moreover, as l increases, t_c also increases gradually ($t_c = 5$ in (a), $t_c = 7$ in (b), and $t_c = 8$ in (c)). Note that $S(t^*) + R(t^*) = 1$ because of $I(t^*) = 0$. Even if the red and black lines for each state S, I, and R on MP-SIR and AVG-SIR models are almost coincided, the MP-SIR is approximately the number of samples times faster than the AVG-SIR (since the MP-SIR does not need the averaging). Besides, before the convergent time t_c (early spreading), the red and black lines on MP-SIR and AVG-SIR are slightly different. Since there are some gap between the highest and

the lowest on AVG-SIR model as shown in Table 3. In other words, as the reason why the difference appears, $I(t)$ and $S(t)$ are underestimated on AVG-SIR because of the lowest value. Although the gap between the lowest and the highest $R(t)$ is not large, the number $N \times R(t)$ of accumulated infection nodes is large enough because of $N = 7624$.



(a)size of seed=885 at l=2

(b)size of seed=516 at l=3

(c)size of seed=407 at l=4

Figure 4 Time evolution of the fractions of three states S, I, and R, in the conventional AVG-SIR and faster MP-SIR models.

t	2	3	4	5	6	7
lowest $R(t)$	0.11608	0.22258	0.31269	0.36791	0.39178	0.39821
highest $R(t)$	0.11608	0.24278	0.34693	0.40424	0.42195	0.42799
average $R(t)$	0.11608	0.23283	0.32923	0.38586	0.40707	0.41360
variance of $R(t)$	0	1.5401e-06	4.1998e-06	4.5865e-06	4.2386e-06	4.2224e-06

Table 3 Gap of the lowest and the highest accumulated infection $R(t)$ in samples for early spreading on AVG-SIR model.

Performance on faster MP-SIR model

We compare the spreading power from multiple seeds chosen by our method and the conventional HD, k-core, LC, LSC, and CI methods for 8 social networks. The typical result for a social network called LastFM is shown below. Other similar results are shown in the supplement for the remaining 7 real networks.

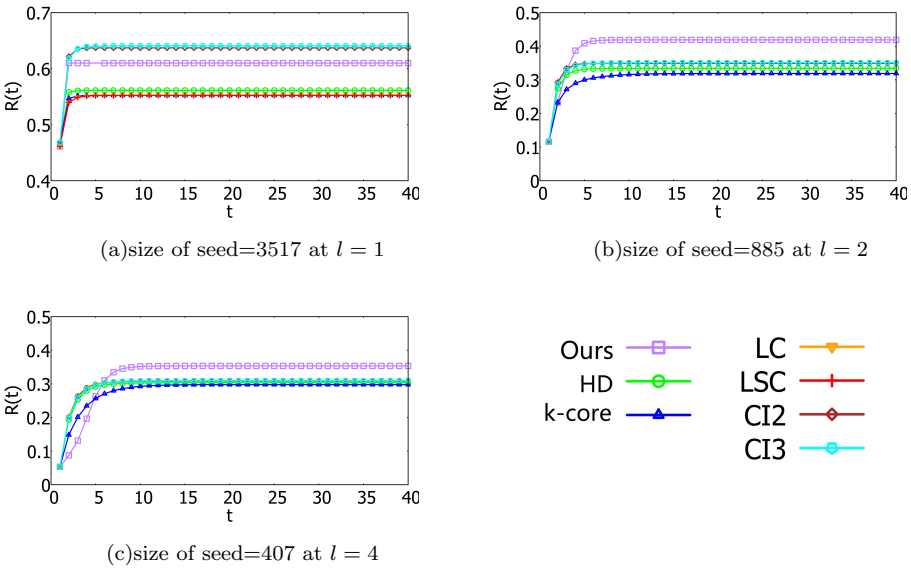
Figure 5 shows the time evolution of accumulated infections $R(t)$. As shown in Fig. 5 (a), purple line with square mark (the minimum VC is chosen as seeds) is lower than brown line with diamond mark (CI_2) and cyan line with pentagon mark (CI_3). Although the reason of lower performance is discussed in Fig. 6 later, it is considered as that the minimum VC does not consider the multi-hop coverage and can not avoid the overlap. As shown in Fig. 5 (b)(c), brown lines with diamond mark (CI_2), cyan lines with pentagon mark (CI_3), orange lines with inverse triangle mark (LC), and red lines with cross mark (LSC) are higher than green lines with circle mark (HD) and blue lines with triangle mark (k-core). We remark that the CI, LC, and LSC

have more spreading power than the HD and k-core, because CI, LC, and LSC not only consider the nearest neighbors of seeds but also the next nearest neighbors, or next-next nearest neighbors, and so on. Remember that, N_s=3517, 885, and 407 (N_s/N =0.46, 0.12, and 0.05). In particular, the purple lines with square mark (our method) is the highest above the green lines with circle mark (HD), blue lines with triangle mark (k-core), orange lines with inverse triangle mark (LC), red lines with cross mark (LSC), brown lines with diamond mark (CI_2), and cyan lines with pentagon mark (CI_3) on faster MP-SIR model. Although the reason of higher line is discussed in Fig. 6 later, it is considered as that seeds chosen by our method are located away from each other as illustrated in Fig. 2(b). Moreover, after the convergent time t_c, the gap between purple line with square mark and other lines in Fig. 5 (b) is larger than ones in Fig. 5 (c). Because as the number of seeds becomes smaller, the spreading power per seed becomes larger. Besides, as l increases, t_c also increases. while the size of seeds decreases.

Figure 6 shows the distribution of distance $d_{i,j}$ of each pair of seeds i,j on 2- or 4-hop coverage. The peaks of two purple lines are righter than the peak of other color lines. It means that seeds chosen by our method are located more far away from each other than ones by the conventional methods. Since the larger distance of two seeds reduces the overlap, our method have more spreading power than the conventional methods. Moreover, the peak of purple line with filled square marks (at distance $d = 6$) is righter than the peak of purple line with square marks (at distance $d = 5$). It indicates that as l(-hop) increases, the distance of seeds increases. However, there is a limitation of larger l as mentioned later with Table 4.

With different spreading rates $\beta = \lambda\frac{\langle k\rangle}{\langle k^2\rangle}$, we investigate the performance of our method for finding multiple seeds. As shown in Fig. 7, the horizontal axis indicate the infection parameter λ from 2 to 10 (β from 0.12 to 0.6). Note that the case of λ=2 corresponds to Fig. 5. The vertical axis $R(t_c)$ is the accumulated infections at the t_c. As shown in Fig. 7 (a), because of the overlap phenomena (l=1 does not consider the multi-hop coverage), purple lines with square mark is not the best. When $l > 1$, purple lines with square mark (our method) are always higher than others (by the conventional methods). However, we can see that the difference between our method (purple line with square mark) and others (brown lines with diamond mark, cyan lines with pentagon mark, orange lines with inverse triangle mark, and red lines with cross mark, green lines with circle mark, and blue lines with triangle mark) becomes gradually smaller as spreading rate increases with the parameter value of λ. Because as spreading rate increases, seeds infect more nodes. Note that a higher spreading rate than percolation threshold [24] is realistic [25].

Furthermore, from Table 4, we can see the spreading power per seed ($N \times R(t_c)/N_s$) chosen by our method is greater than ones by the conventional methods (each of the best performance is emphasized by bold in comparison with the methods at l-values). In particular, as the coverage distance l increases, the spreading power per seed chosen by our method becomes larger. Thus seeds chosen by our method on the larger coverage distance l have better spreading power as l increases, although l is limited as smaller than the $D-1$ of the network. Because when l is larger than $D-1$, all nodes are removed after the first seed are chosen.



(a)size of seed=3517 at $l = 1$ (b)size of seed=885 at $l = 2$

(c)size of seed=407 at $l = 4$

Figure 5 Time evolution of accumulated infections $R(t)$ on the MP-SIR model for LastFM with $\lambda = 2$ and $l = 1, 2, 4$.

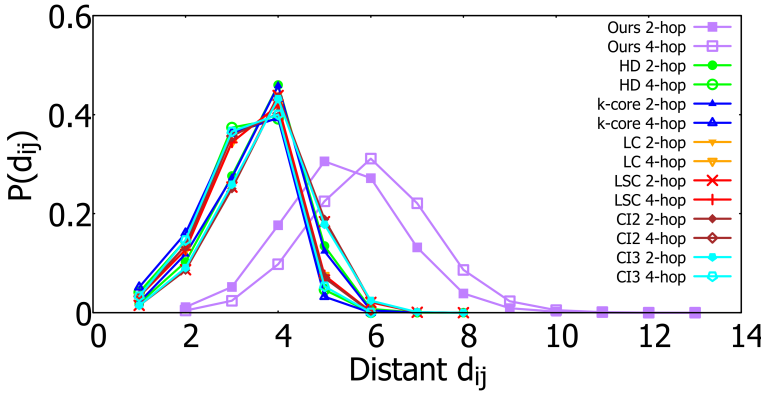
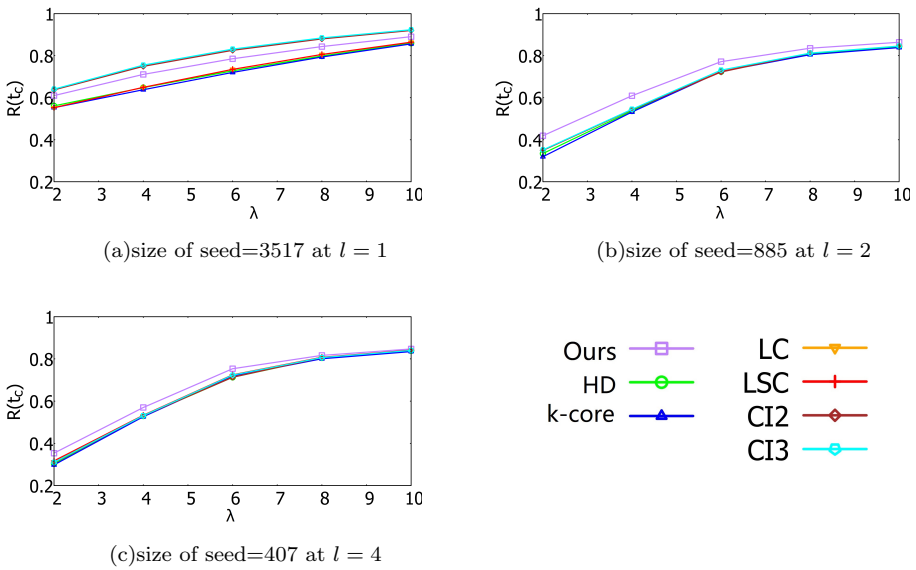


Figure 6 Distribution of distance $d_{i,j}$ of each pair of seeds i,j on 2- or 4-hop coverage ($l = 2, 4$).

	$N \times R(t_c)/N_s$ in AVG-SIR			$N \times R(t_c)/N_s$ in MP-SIR		
l-hops	1	2	4	1	2	4
# of seed	3517	885	407	3517	885	407
Our method	1.3222223	**3.5621469**	**6.5798526**	1.3412023	**3.60896**	**6.62231**
HD	1.2165087	2.8531073	5.4840295	1.242151	2.87991	5.68519
k-core	1.1973728	2.6711864	5.2137592	1.2172916	2.74845	5.5809
LC	1.1977411	2.9990584	5.7027027	1.1993315	3.00862	5.77058
LSC	1.2106251	3.0000389	5.7137027	1.2216768	3.01052	5.78548
CI_2	1.382978	3.0109228	5.5593776	1.4116518	3.00742	5.78242
CI_3	**1.389038**	2.9899919	5.6830467	**1.419038**	3.02844	5.80876

Table 4 Spreading power per seed chosen by our method and the conventional six methods on the AVG-SIR and MP-SIR models, $N \times R(t_c)/N_s$ denotes the spreading power per seed, $R(t_c)$ denotes the accumulated infections at the convergent time t_c.



(a)size of seed=3517 at $l = 1$

(b)size of seed=885 at $l = 2$

(c)size of seed=407 at $l = 4$

Figure 7 Accumulated infections at the convergent time t_c versus different infection parameter λ from 2 to 10 on the MP-SIR model for LastFM.

Conclusion

In summary, to efficiently find multiple seeds, we propose a new method in approximately solving the IMP problem. The key idea is the statistical physics approach for the minimum vertex cover and l-hop coverage, in order to avoid the overlap of spreading. We also propose the MP-SIR model which does not need many samples for averaging stochastic behavior, therefore it is approximately number of sample times faster than the conventional SIR model. We apply the faster MP-SIR model to simulate the spreading process quickly. As obtained results for the time evolution of accumulated infections, our method can outperform other conventional methods for social networks with different sizes. Thus, it is expected that our method is applied different scale networks efficiently.

In multi-hop coverage, how many hops are optimal for avoiding overlap is still an open problem. As future work, we will consider it and give an optimal number of hop that has the best effect for the IMP. Of course we may also integrate other statistical approach to find a better way and propose more effective methods.

ACKNOWLEDGEMENTS
Not applicable.

FUNDING
Not applicable.

CONSENT FOR PUBLICATION
Not applicable.

AVAILABILITY OF DATA AND MATERIAL
Please contact author for data requests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE
Not applicable.

ABBREVIATIONS

IMP: Influence maximization problem. IC: Independent Cascade model. LT: Linear Threshold model. SIR: susceptible-infected-recovered model. BC: betweenness centrality. CC: closeness centrality. HD: High degree method. LC: Local centrality . LSC: Local structure centrality. CI: Collective influence. FVS: Feedback vertex set. VC: Vertex cover problem. AVG-SIR: The averaging behavior in a stochastic SIR epidemic model. MP-SIR: Message passing for SIR model.

COMPUTING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

YH and LF designed research. LF performed research. YH and LF contributed to develop new methods, analyze data, and wrote the paper. All authors read and approved the final manuscript.

Author details

Division of Transdisciplinary Sciences, Japan Advanced Institute of Science and Technology, Nomi, Japan.

References

1. Thomas Valente and Rebecca Davis. Accelerating the diffusion of innovations using opinion leaders. *The Annals of the American Academy of Political and Social Science*, 566(1):55–67, 1999.
2. Flaviano Morone, Kevin Roth, Byungjoon Min, H. Eugene Stanley, and Hernán A. Makse. Model of brain activation predicts the neural collective influence map of the brain. *Proceedings of the National Academy of Sciences*, 114(15):3849–3854, 2017.
3. Linyuan Lü, Duan-Bing Chen, and Tao Zhou. The small world yields the most effective information spreading. *New Journal of Physics*, 13(12):123005, 2011.
4. Zhiou Xu, Xiaobin Rui, Jing He, Zhixiao Wang, and Tarik Hadzibeganovic. Superspreaders and superblockers based community evolution tracking in dynamic social networks. *Knowledge-Based Systems*, 192:105377, 2020.
5. David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.
6. Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, Aug 2015.
7. Richard Manning Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Springer, 1972.
8. Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. Locating privileged spreaders on an online social network. *Physical Review E*, 85(6):066123, 2012.
9. Gouhei Tanaka, Kai Morino, and Kazuyuki Aihara. Dynamical robustness in complex networks: the crucial role of low-degree nodes. *Scientific Reports*, 2(1):1–6, 2012.
10. Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.
11. Duanbing Chen, Linyuan Lü, Ming-Sheng Shang, Yi-Cheng Zhang, and Tao Zhou. Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications*, 391(4):1777–1787, 2012.
12. Shuai Gao, Jun Ma, Zhumin Chen, Guanghui Wang, and Changming Xing. Ranking the spreading ability of nodes in complex networks based on local structure. *Physica A: Statistical Mechanics and its Applications*, 403:130–147, 2014.
13. Xian Teng, Sen Pei, Flaviano Morone, and Hernán A Makse. Collective influence of multiple spreaders evaluated by tracing real information flow in large-scale social networks. *Scientific Reports*, 6(1):1–11, 2016.
14. Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.
15. G Sabidussi. The centrality index of a graph. *Psychometrika 31*, page 581–603, 1966.
16. Paramita Dey, Subhayan Bhattacharya, and Sarbani Roy. A survey on the role of centrality as seed nodes for information propagation in large scale network. *ACM/IMS Trans. Data Sci.*, 2(3), aug 2021.
17. Joshua D. Guzman, Richard F. Deckro, Matthew J. Robbins, James F. Morris, and Nicholas A. Ballester. An analytical comparison of social network measures. *IEEE Transactions on Computational Social Systems*, 1(1):35–45, 2014.
18. Martin Weigt and Haijun Zhou. Message passing for vertex covers. *Physical Review E*, 74(4):046110, 2006.
19. Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.
20. Nam Parshad Bhatia and Giorgio P Szegö. *Stability theory of dynamical systems*. Springer Science & Business Media, 2002.
21. Reuven Bar-Yehuda and Shimon Even. A local-ratio theorem for approximating the weighted vertex cover problem. In *Analysis and Design of Algorithms for Combinatorial Problems*, volume 109 of *North-Holland Mathematics Studies*, pages 27–45. North-Holland, 1985.
22. Hai-Jun Zhou. Spin glass approach to the feedback vertex set problem. *The European Physical Journal B*, 86(11):1–9, 2013.
23. Hao Chen and Jürgen Jost. Minimum vertex covers and the spectrum of the normalized laplacian on trees. *Linear Algebra and Its Applications*, 437(4):1089–1101, 2012.
24. Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Physical Review E*, 65:036104, Feb 2002.
25. Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 26(4):521–529, 2002.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ApplicationofNetworkscience.zip
- Supplement.pdf