Applied Network Science

RESEARCH Open Access

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

Abstract

Finding the influential vertexes 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 vertexes as N_s seeds which maximize the influence represented by the number of activated vertexes from the seeds in a social network, $N_s \geq 1$ is a constant number. The problem has many applications such as the viral marketing (Valente and Davis 1999; Guo et al. 2021b), brain activation (Morone et al. 2017), information dissemination (Lü et al. 2011) of community (Guo and Wu 2020), rumor blocking (Guo et al. 2021a) and halting global epidemic outbreaks in contact networks (Xu et al. 2020). 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 (Kempe et al. 2003). Note that a special case of IC with a constant infection probability on every links is the susceptible-infected-recovered (SIR) model (Pastor-Satorras et al. 2015) as mentioned later. However, the IMP is NP-hard (Karp 1972) for both IC and LT models (Pastor-Satorras et al. 2015). 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 (Borge-Holthoefer et al. 2012; Tanaka et al. 2012), k-core (Kitsak et al. 2010), local centrality (Chen et al. 2012), local structure centrality (Gao et al. 2014), and collective

influence (Teng et al. 2016). However, there are various drawbacks of 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) (Freeman 1977) and closeness centrality (CC) (Sabidussi 1966) can give better results (Dey et al. 2021) for finding multiple seeds, they are unsuitable for very large-scale social networks because of the high computational complexity (Guzman et al. 2014) 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 vertexes and edges, respectively. Although, various efforts have been made on the above research, the design of more effective method is still an open issue especially for finding multiple seeds.

On the other hands, for several NP-hard problems, there exist practically superior approximate algorithms in statistical physics approach. This gives our motivation for considering a new method to finding multiple seeds. In application point of view, the following problem setting to avoid the overlap of spreading from a fixed number of N_s seeds for the IMP:

- How to determine the the number N_s of seeds? $\Rightarrow$ We propose an applying of the extended minimum vertex cover (VC) on l-hop coverage.
- Is our method is better than the conventional selections of seeds as an approximate solution for the IMP? $\Rightarrow$ The spreading powers in our method and conventional methods are compared through numerical simulations.

Our innovative idea is a combination of l-hop coverage in computer science and the statistical physics approach (Weigt and Zhou 2006) for the minimum VC in addition through a faster simulation of information spreading. Note that these two research fields are quite different and not easily contacted. Here, the l-hop coverage means that seeds infect their l-hop neighbors. As the special case of $l = 1$, the set cover, dominating set, and the VC problems are corresponded to 1-hop coverage. Note that, minimum set cover problem can be reducted to the minimum VC problem (Karp 1972). However, the minimum VC problem is NP-hard (Karp 1972). In order to efficiently estimate the set of the minimum VC with global spreading power, we focus on collective computation by local interactions through message-passings based on statistical physics (Weigt and Zhou 2006). Moreover, to reduce the calculation time, we propose a faster simulation based on SIR model inspired from the collective influence (Teng et al. 2016) in physics community of network science. It does not need the average of behavior, therefore it is expected to be the number of samples times faster than the conventional SIR model. Because the SIR model (Pastor-Satorras et al. 2015) is usually 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.

The organization of this paper is as follows. In section "Our Combination Methods", we briefly review the statistical physics approach and propose our method. The conventional SIR model and our faster MP-SIR model are introduced in the subsection "Faster MP-SIR model". Through computer simulation, the spreading power of our method and

other heuristic methods are compared in the section "Simulation Results". Conclusion are given in the last section. In section "Appendix", we explain the conventional heuristic methods for finding seeds.

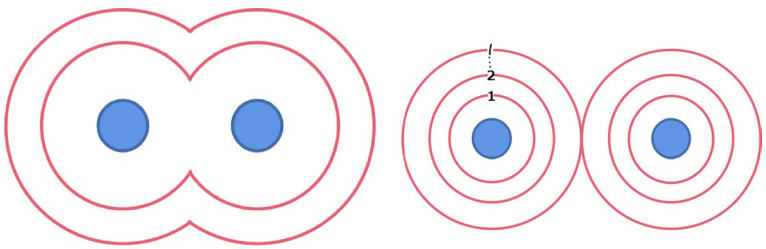
Our combination method

Although the conventional heuristic methods (Borge-Holthoefer et al. 2012; Tanaka et al. 2012; Kitsak et al. 2010; Chen et al. 2012; Gao et al. 2014; Teng et al. 2016) can be applied to find multiple seeds, they do not consider the overlap phenomena. As shown in Fig. 1a, when hubs are near to each other, High degree(HD) method (Borge-Holthoefer et al. 2012; Tanaka et al. 2012) 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.

The outline of our combination method is as follows. First, the number N_s of seeds is determined by the l-hop coverage, in which each distance between seeds is more than l hops. A fixed number N_s are corresponding to l=1, 2, 3, and 4 in order to compare the power of information spreading in the next section. Second, we explain the l-hop coverage. As shown in Fig. 2, vertex i is chosen as the first candidate of VC. After removing the vertex i and its l-hop neighbors, as the second candidate of VC, vertex j is chosen from the remaining network. Then, repeat the above steps until no vertexes exist in the network. The symbol table for our method is shown in Table 1.

Applying a survey propagation to minimum VC

We briefly review the approximate algorithm called survey propagation for the minimum VC problem. In the algorithm (Weigt and Zhou 2006), each vertex 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. As shown in Fig. 3, 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 take care that the following message-passing for estimating the minimum VC differs from information spreading on SIR model (Pastor-Satorras et al. 2015), although the message-passing and information spreading are similar words. For each vertex i, the message-passing equations (Weigt and Zhou 2006) are given by



(a)Close case (b)Distant case

Fig. 1 Two situations of **a** overlap and **b** avoiding the overlap

Table 1 Symbol table for our method

Notation	Description
$G(V, E)$	Graph with vertex set V and edge set E
$N = \mid V \mid, M = \mid E \mid$	Size of vertex set or edge set
n, m	$n = \mid V \mid, m = \mid E \mid$
k	Degree of a vertex (number of edges emanated from a vertex)
N_s	Size of seed set
l	Number of hops
i, j	Index of vertex
$\partial i \backslash j$	Set of the nearest neighbors of vertex i but not including j
e^{-y}	Penalty factor for minimizing the size of VC, y is an inverse temperature parameter
$i \rightarrow j$	Link from vertex i to j
$\hat{\pi}_i^{(0)}$	Probability variable of never covered state 0
$\hat{\pi}_i^{(1)}$	Probability variable of covered state 1
$\hat{\pi}_i^{(*)}$	Probability variable of sometimes covered and sometimes not joker state $*$
set of $\partial Ball(j, l-1)$	Set of the $l-1$ nearest neighbors of vertex j
N_s	Number of seeds
$\langle k \rangle$	Average degree: $2M/N$
β	Infection probability
$S(t)$, $I(t)$, $R(t)$	Cumulative probability of each state of S, I, or R at time t
$P_i^I(t)$, $P_i^R(t)$, and $P_i^S(t)$	Probability of state S, I, and R for a vertex i at time t
$\mid VC \mid$	Size of set of vertexes as vertex cover
$d_{i,j}$	Distance of i and j defined by the shortest path length between them
t_c	Convergent time until all infected vertexes are recovered
D	Diameter of network as the maximum distance of the shortest path between vertexes

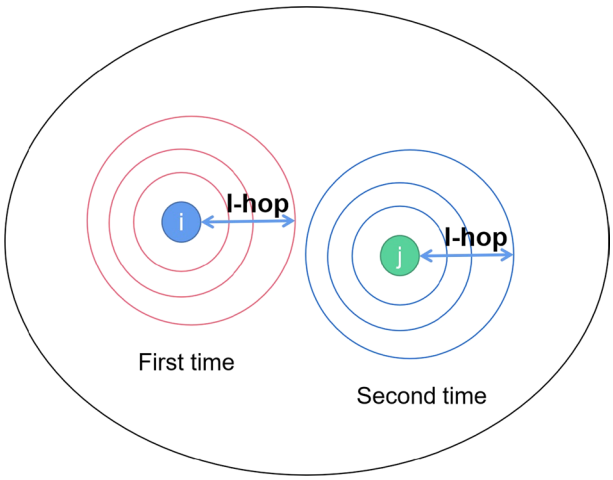


Fig. 2 Illustration of l-hop coverage for determining N_s

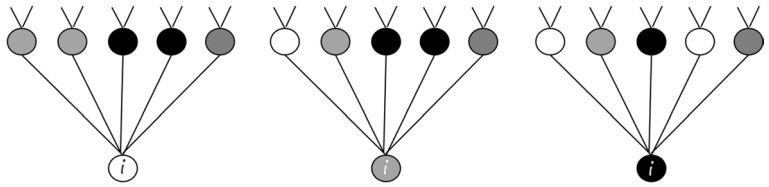


Fig. 3 How to probabilistically determine three different states in the survey propagation. Values 0, ∗, and 1 are represented by a white, gray, and black circles. The state of vertex *i* at bottom is determined by the states of its neighbors ∂i at top. **a** There are no white circle in ∂i, the bottom circle *i* is not necessary to be covered and gets color white (state 0). **b** If there is exactly one white circle in ∂i, the bottom circle *i* becomes gray (joker state). **c** If there are two or more white circles in ∂i, the bottom circle *i* is black as an always covered (state 1)

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

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

$$\hat{\pi}_i^{(1)} = C_i^{-1} e^{-y} \left[1 - \prod_{j \in \partial i} (1 - \hat{\pi}_{j \to i}^{(0)}) - \sum_{j \in \partial i} \hat{\pi}_{j \to i}^{(0)} \prod_{j' \in \partial i \backslash j} (1 - \hat{\pi}_{j' \to i}^{(0)}) \right], \tag{3}$$

where $\partial i \backslash j$ is the set of the nearest neighbors of vertex *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

$$C_i = e^{-y} \left[1 - (1 - e^{y}) \prod_{j \in \partial i} (1 - \hat{\pi}_{j \to i}^{(0)}) \right]. \tag{4}$$

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

$$\hat{\pi}_{i \to k}^{(0)} = C_{i \to k}^{-1} \prod_{j \in \partial i \backslash k} (1 - \hat{\pi}_{j \to i}^{(0)}), \tag{5}$$

$$C_{i \to l} = e^{-y} \left[1 - (1 - e^{y}) \prod_{j \in \partial i \backslash k} (1 - \hat{\pi}_{j \to i}^{(0)}) \right]. \tag{6}$$

Equations (3)–(6) are calculated through *T* round iterations. After the convergence, the vertex *i* with the largest $\hat{\pi}_i^{(1)}$ is selected as the VC. Then it is removed and recalculate the $\hat{\pi}_i$ until all vertexes are covered in the following decimation process on the *l*-hop coverage. The detail of the extended minimum VC on *l*-hop coverage is described as follows

Step 1 By using Eqs. (3)–(6), the probability $\hat{\pi}_i^{(1)}$ of vertex *i* is calculated for estimating the minimum VC.

Step 2 As the decimation process, the vertex *j* with the highest $\hat{\pi}_j^{(1)}$ is selected as a

seed, the chosen vertex 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 vertexes have been removed in the network. Finally, the size of multiple seeds is obtained as N_s.

Faster MP-SIR model

Let us consider the averaging behavior in a stochastic SIR epidemic model (we call it AVG-SIR) (Pastor-Satorras et al. 2015) with three states S: susceptible (inactive) vertexes represents the individuals susceptible to the disease, I: infected (active) vertexes 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 (Pastor-Satorras et al. 2015). At each time step, in the spreading process, an infected vertex changes the states of its neighbors from S to I with probability $\beta = \lambda \frac{\langle k \rangle}{\langle k^2 \rangle}$ (Pastor-Satorras and Vespignani 2002), 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 vertexes as multiple seeds, however many trials of spreading is necessary for the averaging of stochastic behavior with probability β in samplings. 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 (Teng et al. 2016).

$$P_i^I(t+1) = P_i^S(t)\left[1 - \prod_{j \in \partial i}(1 - \beta P_j^I(t))\right], \tag{7}$$

$$P_i^R(t+1) = P_i^R(t) + P_i^I(t), \tag{8}$$

$$P_i^S(t+1) = 1 - P_i^I(t+1) - P_i^R(t+1), \tag{9}$$

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 vertex i at time $t+1$, respectively. Note that already averaged probability values $P_i^I(t)$, $P_i^R(t)$, and $P_i^s(t)$ are updated by time step. We call it MP-SIR model. These message-passing Eqs. (7), (8), and (9) are also physics approach. As the remarkable difference, MP-SIR model is based on already averaged probability variables, therefore it does not need many samples for averaging stochastic behavior.

In summary, there are two contributions to algorithm design as follows:

- By combining survey propagation for finding the minimum VC and l-hop coverage, we propose a new method for finding multiple seeds.
- We propose the faster MP-SIR model by message-passing.

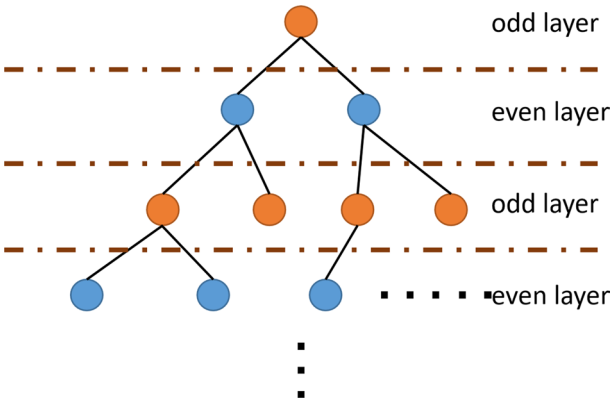


Fig. 4 Well-known assignment method (Chen and Jost 2012) for finding VC in a tree

Table 2 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

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			

The bold numbers are the best results with the minimum VC

Simulation results

The minimum VC by survey propagation versus 2-approximation method

We compare the survey propagation and 2-approximation method through numerical simulations in a realistic network named LastFM to show that the survey propagation can efficiently estimate the minimum VC. As shown in Table 2, the solution by the survey propagation seems to be nearly optimal, while that by 2-approximation method (Bar-Yehuda and Even 1985) 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 (Zhou 2013) 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. 4, we apply the well known method (Chen and Jost 2012) 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 2 shows that $|\ VC\ |$ estimated by the survey propagation is slightly better than $|\ VC\ |$ estimated 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 3 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.

Table 3 VC by the survey propagation under different round *T* for the social network lastFM with N=7624

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

Bold indicates mean the top performers

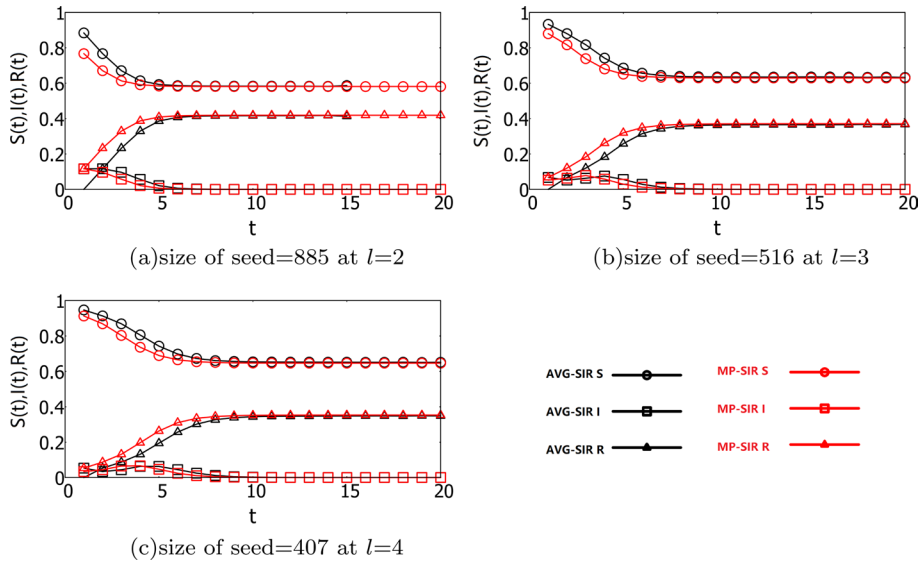


Fig. 5 Time evolution of the fractions of three states S, I, and R, in the conventional AVG-SIR and faster MP-SIR models

Faster MP-SIR versus AVG-SIR

In this subsection, we show that Faster MP-SIR and AVG-SIR models have similar spreading behaviors. In Fig. 5, we investigate 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 $\beta = 0.12$ at percolation threshold $\lambda \frac{\langle k \rangle}{\langle k^2 \rangle}$, $\lambda = 1$. These values of l give different size N_s of seeds. The rate of seeds are $N_s/N = 0.12$ ($l = 2$), 0.07 ($l = 3$), and 0.05 ($l = 4$), respectively. Note that $N_s/N \leq 20\%$ is realistic (Kitsak et al. 2010). Here, we define $S(t) = \sum_{i=1}^{N} P_i^S(t)/N$, $I(t) = \sum_{i=1}^{N} P_i^I(t)/N$, $R(t) = 1 - S(t) - I(t)$. In Fig. 5a–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 vertexes 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

Table 4 Gap of the lowest and the highest accumulated infection $R(t)$ in samples for spreading time from 2 to 7 on AVG-SIR model.

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 5 The difference and Calculation time until convergence in AVG and MP SIR model (CPU:i7-11800H, Memory:16GB)

	AVG	MP
Difference	Need the averaging of stochastic behavior	Equations for already averaged variables
Calculation time (Sec)	602.28630	20.06478

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 $R(t)$ and the lowest $R(t)$ in samples on AVG-SIR model as shown in Table 4. Note that, the gap corresponds to the difference between the red line and the black line from $t = 2$ to $t = 7$ in Fig. 5. 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 vertexes is large enough because of the network size $N = 7624$. Moreover, Table 5 shows our faster MP-SIR model is 30 time faster than the AVG-SIR, it means 1000 samples $\approx 30\times$T-rounds ($T = 50$). The rate of speed up (calculation time of AVG-SIR / MP-SIR) are from 20 to 30. The detail is shown in the Additional file 1.

Our method versus conventional methods for finding seeds

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. Note that our method and conventional methods have the same seed size. Since the CI3 outperform the other conventional methods, we consider it result as the base line. Similar results are obtained for the remaining 7 real networks in the Additional file 1.

Figure 6 shows the time evolution of accumulated infections $R(t)$. As shown in Fig. 6a, 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 later in Fig. 7, 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. 6b, 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

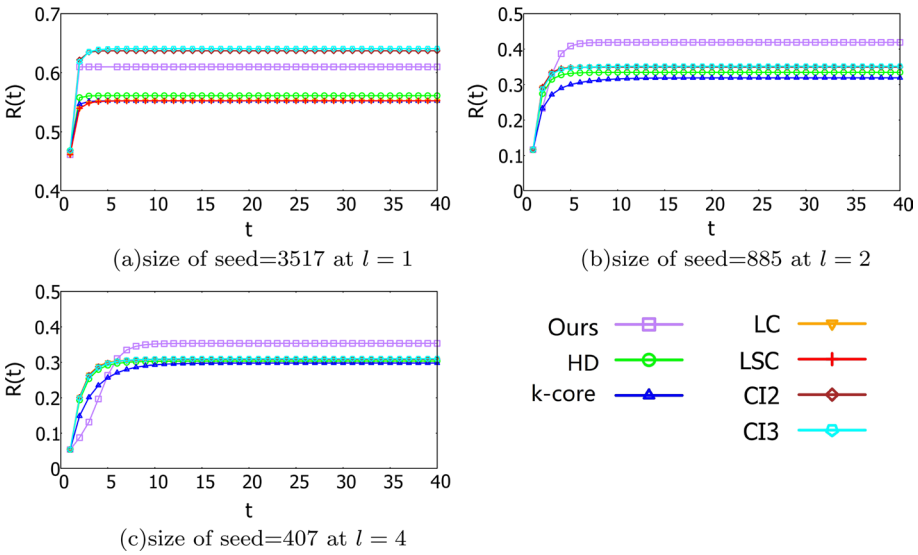


Fig. 6 Time evolution of accumulated infections $R(t)$ on the MP-SIR model for LastFM with $\lambda = 2$ and $l = 1, 2, 4$

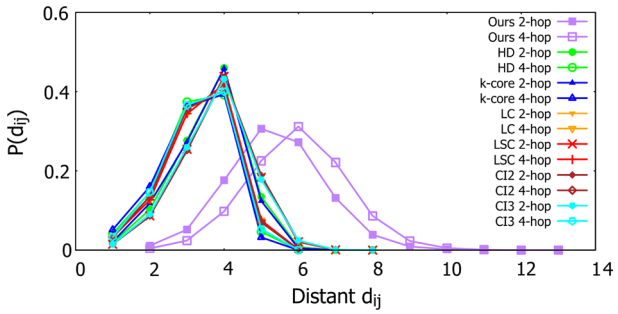


Fig. 7 Distribution of distance $d_{i,j}$ of each pair of seeds i,j on 2- or 4-hop coverage ($l = 2, 4$)

(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 later in Fig. 7, it is considered as that seeds chosen by our method are located away from each other as illustrated in Fig. 1b. Moreover, after the convergent time t_c, the gap between purple line with square mark and other lines in Fig. 6b is larger than ones in Fig. 6c. 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.

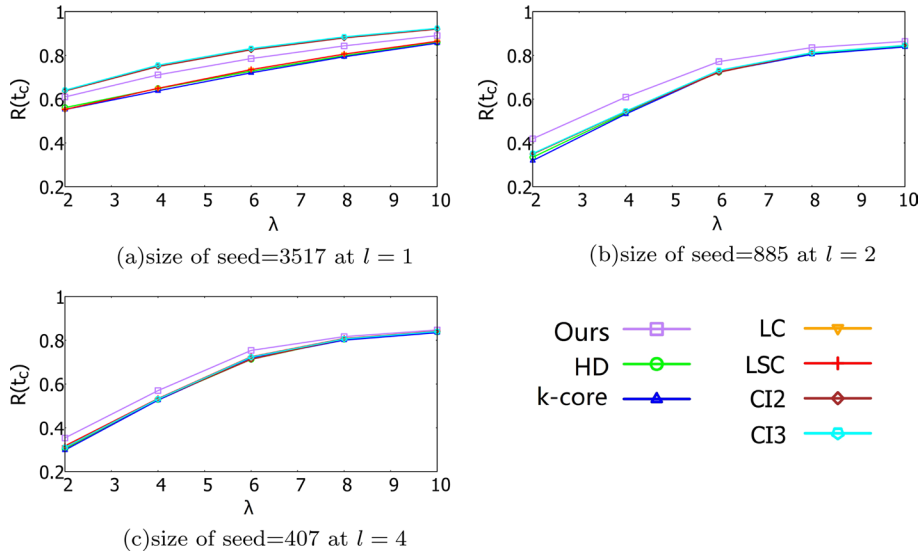


Fig. 8 Accumulated infections at the convergent time t_c versus different infection parameter λ from 2 to 10 on the MP-SIR model for LastFM

Table 6 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

	$N \times R(t_c)/N_s$ **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

Bold indicates mean the top performers

Figure 7 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 6.

With different spreading rates $\beta = \lambda \frac{\langle k \rangle}{\langle k^2 \rangle}$, λ=2, 4, 6, and 8, we investigate the performance of our method for finding multiple seeds. Note that a higher spreading rate than the

percolation threshold $\beta = \lambda \frac{\langle k \rangle}{\langle k^2 \rangle}$ (Pastor-Satorras and Vespignani 2002) is realistic (Moreno et al. 2002). As shown in Fig. 8, 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. 6. The vertical axis $R(t_c)$ is the accumulated infections at the convergent time t_c. As shown in Fig. 8a, 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 vertexes.

Furthermore, from Table 6, 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. Here, D is the diameter of network defined as the maximum distance of the shortest path between vertexes. Because when l is larger than $D - 1$, all vertexes are removed after the first seed are chosen. Remember that, the N_s is determined by the number of VC.

Conclusion

In summary, to efficiently find multiple seeds, we propose a new method in approximately solving the IMP problem. The key idea is a combination of the statistical physics approach for the minimum VC 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 samples / T-rounds 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.

However, 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 gives the best effect for the IMP.

Moreover, there are two algorithms (Han et al. 2020; Guo et al. 2020) based on IC model. They are quite different from our method which is based on SIR model as a special case of IC model. If we consider IC model extendedly, some relations maybe exist between these two algorithms (Han et al. 2020; Guo et al. 2020) and our method for spread overlap issue. We may find some new way to solve the IMP.

Appendix

The symbol table for the conventional methods is shown in Table 7.

Table 7 Symbol table for the conventional methods.

Notation	Description
k_s	Level of shell
u, v, and w	Index of vertexes
α	Tunable balance parameter
$C_L(v)$	Local centrality of vertex v
$C_{LS}(v)$	Local structure centrality of vertex v
∂u	Set of the nearest neighbors of vertex u
$\partial Ball(u,2)$	Set of the next nearest neighbors of vertex u
k_i, k_j	Degree of vertex i or j
$\mathcal{R}$	Radius of the $Ball(u,\mathcal{R})$ from vertex u
$CI_{\mathcal{R}}(i)$	Collective influence of vertex i with radius $\mathcal{R}$
$i \rightarrow j$	Link from i to j

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

High degree

The High degree (HD) method selects k vertexes in decreasing order of degrees as the influential seeds (Borge-Holthoefer et al. 2012; Tanaka et al. 2012). 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 (Kitsak et al. 2010), seeds are ranked according to their k_s values, which are calculated through the k-shell decomposition. In the k-shell decomposition, vertexes 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 vertexes. Similarly, the next k-shells with index $k_s \leftarrow k_s + 1$ are extracted, the vertexes located within the core have the highest k_s values. Actually, in the k-shell decomposition, all vertexes are divided into shells. In comparison with the peripheral vertexes, the core vertexes tend to involve larger spreading from them. Therefore, the vertex 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 (Chen et al. 2012). The LC is defined as

$$Q_{LC}(u) = \sum_{w \in \partial u} \sum_{k=0}^{2} \mid \partial Ball(w,k),$$

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

where ∂u denotes the set of the nearest neighbors of vertex u, $\partial Ball(w,k)$ denotes a set of vertexes within k hops from vertex w as shown in Fig. 9. denotes its size. As a seed, v is selected in decreasing order of $C_{LC}(v)$. Note that LC gives similar spreading power as good as the closeness centrality (Chen et al. 2012).

In addition, Local structure centrality (LSC) is an extension of LC (Gao et al. 2014). The LSC is defined by the linear interpolation of local clustering coefficient C_w (Watts and Strogatz 1998) and LC with a tunable balance parameter $0 \leq \alpha \leq 1$.

$$Q_{LSC}(u) = \alpha \sum_{k=0}^{2} \mid \partial Ball(u,k) \mid + (1-\alpha) \sum_{w' \in \partial Ball(u,2)} C_{w'},$$

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

where $\partial Ball(u,2)$ is a set of the next nearest neighbors of vertex u. As mentioned in Gao et al. (2014), we set $\alpha = 0.7$

Collective influence

Collective influence (CI) aims to find the minimum set of vertexes for the IMP as follows (Teng et al. 2016). At the origin $\{v_{i \to j}\} = \{0\}$, the stability of nonlinear message-passing equation

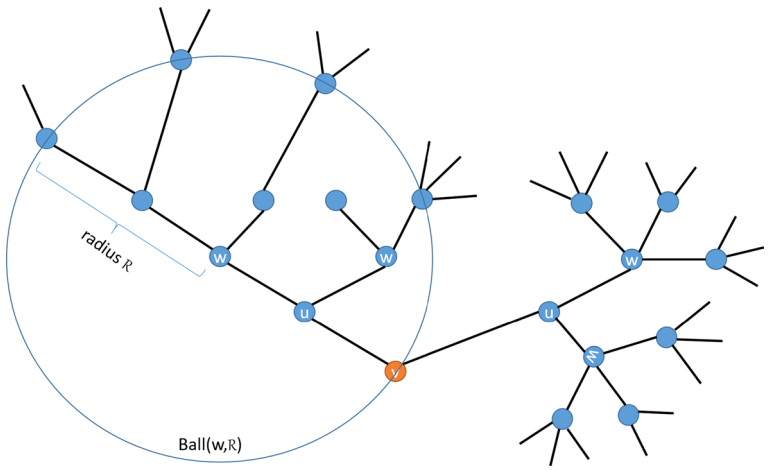


Fig. 9 A set of vertexes from 2-hops from vertex w in the process of LC

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

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 vertexes $\{i : n_i = 0\}$ as influences. Thus, by using a greedy algorithm to minimize the eigenvalue, CI is derived (Bhatia and Szegö 2002) through a power method for each vertex 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 vertex i, $CI_{\mathcal{R}}(i')$ is recalculated for the remaining vertexes $i' \in V$ in the network. It needs only local topological structure within the ball of the radius $\mathcal{R}$ instead of the whole network.

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 modelSupplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s41109-022-00491-x.

Additional file 1. Supplementary tables and figures.

Author 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.

Availability of data materials

Please contact author for data requests.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 31 March 2022 Accepted: 12 July 2022
Published online: 25 July 2022

References

Bhatia NP, Szegö GP (2002) Stability theory of dynamical systems. Springer, Berlin
Borge-Holthoefer J, Rivero A, Moreno Y (2012) Locating privileged spreaders on an online social network. Phys Rev E 85(6):066123

Chen H, Jost J (2012) Minimum vertex covers and the spectrum of the normalized Laplacian on trees. Linear Algebra Appl 437(4):1089–1101
Chen D, Lü L, Shang M-S, Zhang Y-C, Zhou T (2012) Identifying influential nodes in complex networks. Phys A Stat Mech Appl 391(4):1777–1787
Dey P, Bhattacharya S, Roy S (2021) A survey on the role of centrality as seed nodes for information propagation in large scale network. ACM/IMS Trans Data Sci 2(3):1–25
Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35–41
Gao S, Ma J, Chen Z, Wang G, Xing C (2014) Ranking the spreading ability of nodes in complex networks based on local structure. Phys A Stat Mech Appl 403:130–147
Guo J, Wu W (2020) Influence maximization: seeding based on community structure. ACM Trans Knowl Discov Data 14(6):1–22
Guo J, Chen T, Wu W (2021) A multi-feature diffusion model: rumor blocking in social networks. IEEE/ACM Trans Netw 29(1):386–397
Guzman JD, Deckro RF, Robbins MJ, Morris JF, Ballester NA (2014) An analytical comparison of social network measures. IEEE Trans Comput Soc Syst 1(1):35–45
Han K, Xiao X, Chen W, Sun A, Tang X, Lim A, Huang K, Tang J (2020) Efficient approximation algorithms for adaptive influence maximization. Int J Very Large Data Bases 29:1385–1406
Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888–893
Lü L, Chen D-B, Zhou T (2011) The small world yields the most effective information spreading. New J Phys 13(12):123005
Moreno Y, Pastor-Satorras R, Vespignani A (2002) Epidemic outbreaks in complex heterogeneous networks. Eur Phys J B Condens Matter Complex Syst 26(4):521–529
Morone F, Roth K, Min B, Stanley HE, Makse HA (2017) Model of brain activation predicts the neural collective influence map of the brain. Proc Natl Acad Sci 114(15):3849–3854
Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. Phys Rev E 65:036104
Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. Rev Mod Phys 87(3):925–979
Sabidussi G (1966) The centrality index of a graph. Psychometrika 31:581–603
Tanaka G, Morino K, Aihara K (2012) Dynamical robustness in complex networks: the crucial role of low-degree nodes. Sci Rep 2(1):1–6
Teng X, Pei S, Morone F, Makse HA (2016) Collective influence of multiple spreaders evaluated by tracing real information flow in large-scale social networks. Sci Rep 6(1):1–11
Valente T, Davis R (1999) Accelerating the diffusion of innovations using opinion leaders. Ann Am Acad Polit Soc Sci 566(1):55–67
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
Weigt M, Zhou H (2006) Message passing for vertex covers. Phys Rev E 74(4):046110
Xu Z, Rui X, He J, Wang Z, Hadzibeganovic T (2020) Superspreaders and superblockers based community evolution tracking in dynamic social networks. Knowl Based Syst 192:105377
Zhou H-J (2013) Spin glass approach to the feedback vertex set problem. Eur Phys J B 86(11):1–9
Bar-Yehuda R, Even S (1985) 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. North-Holland, pp 27–45
Guo Q, Wang S, Wei Z, Chen M (2020) Influence maximization revisited: efficient reverse reachable set generation with bound tightened. In: Proceedings of the 2020 ACM SIGMOD international conference on management of data, pp 2167–2181
Guo J, Zhang Y, Wu W (2021) An overall evaluation on benefits of competitive influence diffusion. IEEE Trans Big Data 1
Karp RM (1972) Reducibility among combinatorial problems. In: Complexity of computer computations. Springer, pp 85–103
Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, pp 137–146

Publisher's Note