EPJ B
Condensed Matter and Complex Systems

EPJ.org
your physics journal

Eur. Phys. J. B (2016) 89: 188 DOI: 10.1140/epjb/e2016-60612-y

Phase transitions for information diffusion in random clustered networks

Sungsu Lim, Joongbo Shin, Namju Kwak and Kyomin Jung

edp sciences
Società Italiana di Fisica
Springer

Eur. Phys. J. B (2016) 89: 188
DOI: 10.1140/epjb/e2016-60612-y

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Phase transitions for information diffusion in random clustered networks

Sungsu Lim[1], Joongbo Shin[2], Namju Kwak[3], and Kyomin Jung[2,a]

[1] Graduate School of Knowledge Service Engineering, KAIST, 34141 Daejeon, Republic of Korea
[2] Department of Electrical and Computer Engineering, Seoul National University, 08826 Seoul, Republic of Korea
[3] Korea Development Bank, 07242 Seoul, Republic of Korea

Received 29 July 2015 / Received in final form 10 December 2015
Published online 5 September 2016 –

Abstract. We study the conditions for the phase transitions of information diffusion in complex networks. Using the random clustered network model, a generalisation of the Chung-Lu random network model incorporating clustering, we examine the effect of clustering under the Susceptible-Infected-Recovered (SIR) epidemic diffusion model with heterogeneous contact rates. For this purpose, we exploit the branching process to analyse information diffusion in random unclustered networks with arbitrary contact rates, and provide novel iterative algorithms for estimating the conditions and sizes of global cascades, respectively. Showing that a random clustered network can be mapped into a factor graph, which is a locally tree-like structure, we successfully extend our analysis to random clustered networks with heterogeneous contact rates. We then identify the conditions for phase transitions of information diffusion using our method. Interestingly, for various contact rates, we prove that random clustered networks with higher clustering coefficients have strictly lower phase transition points for any given degree sequence. Finally, we confirm our analytical results with numerical simulations of both synthetically-generated and real-world networks.

1 Introduction

Information diffusion plays an essential role in numerous human interactions, including the diffusion of innovations, ideas, rumours, and epidemics. The phase transitions of information diffusion have been observed in many complex systems, such as social, economic, and biological systems [1–3]. That is, there exists a *phase transition point* at which the fraction of the population who adopt the information becomes positive – this is called a *global cascade*. Identifying the conditions in which a global cascade occurs and analysing the behaviours of information diffusion are crucial problems because they are closely related to the eruption of epidemics in epidemiology, the initiation of trends in marketing, and so on.

The spread of information can be modeled using epidemic models [3–8]. Our paper deals with the Susceptible-Infected-Recovered (SIR) model, which has been a popular model in the standard literature [9–11]. In the SIR model, individuals are classified as susceptible, infected, or recovered. Initially, some nodes are infected, and all others are susceptible. Susceptibles can become infectious through meeting infecteds with given probability – called the *contact rate*, while infecteds will recover after some time and never be infectious again. To analyse the final outcome of information spreading, it is known that continuous recovery times for infecteds can be mapped without loss of generality to a single time step of the corresponding stochastic SIR model [12–16], while the basic SIR model has been studied as a continuous time process.

Although modeling information diffusion has been heavily studied, most previous work has considered overly simplified structures in two aspects; network topology and information diffusion pattern. One of the key features of modeling network topology is the effect of triadic closure (friends of friends are more likely to become friends), which occurs in many real-world networks [4,17] and is usually related to the *clustering coefficient*. Despite its importance, most previous studies on information diffusion have focused on limited types of networks including complete graphs [12,18] and locally tree-like networks, such as Erdős-Rényi random networks [19] and configuration models [20,21]. Also, the other key feature of modeling information diffusion patterns is considering the *heterogeneous contact rates*, i.e., contacts between individuals are not identical. Recent studies demonstrate that the impact of human activity patterns on information diffusion show significant heterogeneity [22,23]. However, many studies of spreading in complex networks have covered only constant contact rate [8,15].

Recently, Newman [24] and Miller [25] simultaneously proposed a model of random networks with clustering that displays a clustered structure by considering a given degree sequence and the number of triangles. Their model randomly generates a network based on $\mathbf{t} = (t_1, \ldots, t_n)$

[a] e-mail: kjung@snu.ac.kr

and $\mathbf{s} = (s_1, \ldots, s_n)$, which consist of the number of triangles and that of single edges at each node respectively. Note that this is a generalisation of the configuration model that generates a random network for a given *exact* degree sequence by incorporating clustering, while an alternative model called the Chung-Lu model [26,27] generates a random network with a given *expected* degree sequences. In this paper, we introduce a generalisation of the Chung-Lu model using the expected numbers of triangles and single edges at each node by incorporating clustering, as in the work of Newman and Miller. Here, we call this model the *random clustered network* for clarity. We also call the Chung-Lu model the *random unclustered network*, which is unlikely to have clustered structure. However, the previous studies on random graphs with clustering [24,25] do not cover the heterogeneous contact rates, a key feature of modeling information diffusion.

In this paper, we extensively analyse information diffusion in random clustered networks having *arbitrary* contact rates under the SIR model, and then identify the conditions for the phase transition of global cascades in random clustered networks. As a result, we also give conditions for phase transitions under random clustered networks and show that clustered networks promote the occurrence of a global cascade for various contact rates, which means that a clustered structure is an advantageous structure for information diffusion.

In more detail, we first use the branching process to analyse information diffusion in random unclustered networks with arbitrary contact rates, and provide novel iterative algorithms for estimating the occurrence probabilities and the expected sizes of global cascades respectively. The correctness of the algorithms is analytically proved. We extend our analysis to random clustered networks using a branching process at a macroscopic level by constructing a locally-tree like factor graph of a given network. To do so, we consider a factor graph with two types of nodes: a node and a factor node consisting of three nodes of each triangle from the original network. Furthermore, we performed numerical experiments on both synthetic and real-world networks to confirm the validity and accuracy of our results.

In order to further illustrate the importance of considering the heterogeneous contact rates, we introduce several classes of heterogeneous contact rate that can be used in modeling various contact patterns. For the contact rates of our overall studies, we assume that each infectious node i has a single chance to infect each of its susceptible neighbours j with probability $f(i,j)$ independently. The standard contact rates we used include, but are not limited to, the following:

(i) heterogeneous susceptibility, c/w_i,
(ii) heterogeneous infectivity, c/w_j,
(iii) heterogeneous infectivity and susceptibility simultaneously, $(c/2)/w_i + (c/2)/w_j$,

where c is the *infect/receive ability* and w_i, w_j are the degrees of a sender and a receiver, respectively. These scenarios are reasonable, because each user can only pay attention to a limited number of messages from their friends through Online Social Networks (OSNs), such as Twitter and Facebook [28,29].

The rest of this paper is organised as follows. In Section 2, we analyse information diffusion in random unclustered networks with arbitrary contact rates. In Section 3, we demonstrate our analysis can be applied to a random graph model of a clustered network, by using the macroscopic level branching process. In Section 4, our approach facilitates the analysis of the phase transition of a global cascade for various contact rates. The reliability of our algorithm is confirmed by conducting numerous simulations in Section 5. Finally, Section 6 concludes this paper.

2 Random unclustered networks

In this section, we analyse information diffusion in random networks having heterogeneous contact rates without considering any clustering, using the Chung-Lu model, which is a special case of the random clustered network model. By utilising the analytical tools we develop in this section, information diffusion analysis on random clustered networks will be given in the next section.

The *random unclustered network* $G(\mathbf{w})$ with a given expected degree sequence $\mathbf{w} = (w_1, \ldots, w_n)$ is defined as a probability space over the set of networks on the node set $V = \{1, \ldots, n\}$, where the edge set E is generated by:

$$Pr[\{i,j\} \in E] := \frac{w_i w_j}{\sum_k w_k}. \tag{1}$$

Intuitively, the edge set can be approximated by $\frac{w_i w_j}{\sum_k w_k}$ if the w_i's are not so large. This assumption is quite reasonable in many complex networks (e.g., Facebook has a maximum of 5000 friends limit [30], even if it has 1.55 billion monthly active users as of Sep. 2015[1].). For instance, we assume that almost all nodes are of degree $o(n^{1/\tau})$ for some $2 < \tau < 3$, e.g., a power-law exponent of the underlying degree distribution [31]. Therefore, the expected degree of i is w_i, since $\mathbb{E}[d_i] \approx \sum_j \frac{w_i w_j}{\sum_k w_k} = w_i$, where d_i is the actual degree of i in a network generated by $G(\mathbf{w})$.

The properties of $G(\mathbf{w})$ have been well studied by Chung and Lu [26,27]. For example, if w_i's are identical to a constant, $G(\mathbf{w})$ is equivalent to the Erdős-Rényi random network [19].

2.1 The occurrence of a global cascade

We now intend to compute the probability of the occurrence of a global cascade asymptotically on $G(\mathbf{w})$. For each node $i \in V$, let P_i be the probability that an initially infectious node i induces a global cascade, and let P be the probability that a randomly chosen initial infected induces a global cascade, hence $P = \frac{1}{n}\sum_{i=1}^{n} P_i$. For any $i \in V$, we can compute P_i by using the branching process. For each node i, the probability that an initial infected i

[1] http://newsroom.fb.com/company-info/

does not induce a global cascade ($=1-P_i$) can be represented by the product of the probabilities that every other node j is either not adjacent to i or is adjacent to i but does not induce a global cascade. It is possible because the cascades induced by child nodes of i are independent, since a random unclustered network has a locally tree-like structure. Thus, we find that

$$1-P_i=\prod_{j:j\neq i}\left(1-k(i,j)P_j\right), \tag{2}$$

where $k(i,j)=\frac{w_iw_j}{Vol(G)}f(i,j)$ and $Vol(G)=\sum_{k=1}^{n}w_k$. Note that the probability that an infectious node i infects a susceptible node j in $G(\mathbf{w})$ is equal to $k(i,j)$, because the probability that j is adjacent to i is $\frac{w_iw_j}{Vol(G)}$ and the probability that i infects its neighbour j is $f(i,j)$.

By taking the logarithm of both sides of (2) and applying the first-order Taylor series approximation for the right-hand side, we obtain that for each $i=1,\ldots,n$,

$$\log(1-P_i)=-\sum_{j:j\neq i}k(i,j)P_j+\tilde{\epsilon}_i$$
$$\Rightarrow P_i=1-\exp\left\{-\sum_j k(i,j)P_j\right\}+\epsilon_i, \tag{3}$$

where $\tilde{\epsilon}_i$ and ϵ_i are negligible errors caused by the Taylor series approximation (see Appendix A).

From equation (3), we can also derive a formula to compute the probability that a randomly chosen initial infected induces a global cascade on the network asymptotically, where $\bar{\epsilon}=\frac{1}{n}\sum_i\epsilon_i$:

$$P=1-\frac{1}{n}\sum_i\exp\left\{-\sum_j k(i,j)P_j\right\}+\bar{\epsilon}. \tag{4}$$

It can be shown that the solution of P_i by equation (2) and the solution of $\hat{P}_i$ by equation (3), ignoring the error terms, are the same asymptotically (i.e., the difference between them tends to 0 as $n\to\infty$). Using this result, we can design an algorithm for finding $\hat{P}_i$ as the estimate of P_i. Let $g^{(0)}$ and $g^{(1)}$ be functions from $[0,1]^n$ to $[0,1]^n$ for all $\mathbf{x}=(x_1,\ldots,x_n)\in[0,1]^n$,

$$g^{(0)}(\mathbf{x})=\left(1-\prod_{j:j\neq i}\{1-k(i,j)x_j\}\right)_{i=1,\ldots,n}$$

and

$$g^{(1)}(\mathbf{x})=\left(1-\exp\left\{-\sum_j k(i,j)x_j\right\}\right)_{i=1,\ldots,n}. \tag{5}$$

$\mathbf{P}=(P_1,\ldots,P_n)$ and $\hat{\mathbf{P}}=(\hat{P}_1,\ldots,\hat{P}_n)$ then satisfy that $\mathbf{P}=g^{(0)}(\mathbf{P})$ and $\hat{\mathbf{P}}=g^{(1)}(\hat{\mathbf{P}})$ by equation (2) and equation (3). It means that $\mathbf{P}$ and $\hat{\mathbf{P}}$ are *fixed points* of $g^{(0)}$ and $g^{(1)}$ respectively.

In a discrete dynamical system, we can analyse the behaviour of the system by considering the *stability* of a fixed point. A fixed point $\mathbf{x}$ of a system g is called *unstable* if whose nearby solutions diverge away from the fixed point, which means that the system g does not converge to $\mathbf{x}$ if an initial input is not $\mathbf{x}$. Likewise, a fixed point $\mathbf{x}$ of the system g is *stable* when an initial input close to $\mathbf{x}$ implies that the system converges to $\mathbf{x}$. For a continuously differentiable map g, it is known that a fixed point $\mathbf{x}$ is stable if $\rho=\rho(\nabla g(\mathbf{x}))<1$, where ρ is the largest (in magnitude) eigenvalue of the Jacobian of g at $\mathbf{x}$, $\nabla g(\mathbf{x})=(\frac{\partial g_i}{\partial x_j})_{i,j=1,\ldots,n}$, with the i-th component g_i of the vector-valued function g [32].

It can be easily checked that 0 is a fixed point of both $g^{(0)}$ and $g^{(1)}$. Consider the case in which 0 is a *stable* fixed point of a discrete dynamical system g. If we assume that a (few) initial nodes are infected at the beginning, then since 0 is a stable fixed point of g, a global cascade does not take place. We then consider the case where 0 is unstable, so that there is a non-zero fixed point $\mathbf{x}$ of g. Using a similar argument, we identify the conditions for phase transitions in Section 4.

Algorithm 1 Computing the estimates of P_is.

Require: $\hat{\mathbf{P}}^{(0)}=(\hat{P}_1{}^{(0)},\ldots,\hat{P}_n{}^{(0)})\in[0,1]^n$, $\epsilon_{tol}>0$
Ensure: $\hat{\mathbf{P}}=(\hat{P}_1,\ldots,\hat{P}_n)$ and $\hat{P}$

```
repeat
    P̂^(t+1) ← g^(1)(P̂^(t))
    t ← t + 1
until ||g^(1)(P̂^(t)) − P̂^(t)||_1 < ε_tol
P̂ ← g^(1)(P̂^(t))
P̂ ← (1/n) Σ_i P̂_i
```

We now propose a novel iterative algorithm that computes the estimates of the P_is by finding a non-zero fixed point of $g^{(1)}$. As described above, this algorithm guarantees that the errors of the estimates are negligible. The pseudocode of the algorithm is given in Algorithm 1. In this algorithm, the pre-specified tolerance ϵ_{tol} determines the stopping criteria. By using the triangle inequality, we prove that $\frac{1}{n}\|\hat{\mathbf{P}}-\mathbf{P}\|_1=\frac{1}{n}\sum_i|\hat{P}_i-P_i|$ is $o(1)$ (details are in Appendix B), and it means our proposed algorithm provides estimates of P_i and P that are asymptotically equivalent to the actual values.

Moreover, one can extend our iterative algorithm to various situations. For instance, let x_i be the probability that a node i is chosen to be an initial infected for all $i=1,\ldots,n$ independently so that $\sum_i x_i=1$. One then obtains P by solving $P=\sum_{i=1}^{n}x_iP_i$ with the same error bound $o(1)$ even if an initial infected is not chosen uniformly at random, because P is the weighted average of P_is with the same error bound $o(1)$.

2.2 The expected size of a global cascade

We turn to the calculation of the expected size of a global cascade once it is triggered. Let S_j be the probability that a node j is contained in a global cascade and S be the probability that a randomly chosen node is contained in a global cascade. Therefore, $S = \frac{1}{n}\sum_{j=1}^{n} S_j$ is the expected size of a global cascade once it is triggered. For any $j \in V$, we compute S_j by using the backward branching process approach. Any initially non-infectious node j is contained in a global cascade if there is another infectious node i that belongs to a global cascade and j is infected from i. The probability that a susceptible node j is not contained in a global cascade ($=1-S_j$) can be represented by the product of the probabilities that every other node i is either not adjacent to j or is adjacent to j but is not contained in a global cascade. Therefore,

$$1 - S_j = \prod_{i:i\neq j}\Big(1 - k(i,j)S_i\Big), \tag{6}$$

where $k(i,j) = \frac{w_i w_j}{Vol(G)} f(i,j)$, which is the probability that an infectious node i infects a susceptible node j in a random unclustered network, as in Section 2.1.

We take the logarithm of both sides, and apply the first-order Taylor series approximation for the right-hand side, then for each $j = 1, \ldots, n$,

$$\log(1 - S_j) = -\sum_{i;i\neq j} k(i,j)S_i + \tilde{\epsilon}'_j$$
$$\Rightarrow S_j = 1 - \exp\left\{-\sum_i k(i,j)S_i\right\} + \epsilon'_j, \tag{7}$$

and thus,

$$S = 1 - \frac{1}{n}\sum_j \exp\left\{-\sum_i k(i,j)S_i\right\} + \tilde{\epsilon}', \tag{8}$$

where the errors are caused by the Taylor series approximation, as in Section 2.1. By the same argument, the error term of (8) is negligible in many diffusion processes for sufficiently large n. Let $g^{(2)} : [0,1]^n \to [0,1]^n$ be a function that satisfies

$$g^{(2)}(\mathbf{S}) = \left(1 - \exp\left\{-\sum_i k(i,j)S_i\right\}\right)_{j=1,\ldots,n}.$$

We then have a formula to compute the estimates of S_j and S numerically using equations (7) and (8). The iterative algorithm is given in the Algorithm 2.

Note that the result of this section is an extension of the model to allow for heterogeneous contact rates. In the case of Erdős-Rényi random networks $G(n,p)$ with $f(i,j) \equiv 1$, the probability and the size of the giant component S are the solutions of $S = 1 - e^{-npS}$ by equations (4) and (8). This is identical to the well-known result in references [19,33], and it is equivalent to the bond percolation problem if $f(i,j) \equiv 1$ [34–36]. We note that if $f(i,j) \equiv T$ for a constant T and a given network is fully connected, information diffusion in the network can be mapped exactly onto a bond percolation with the edge occupation probability T [13,15,25]. It is also worth noting that the algorithms for estimating P and S are from the same approach even with heterogeneity, in agreement with previous work [12,13,37].

Algorithm 2 Computing the estimates of S_js.

Require: $\hat{\mathbf{S}}^{(0)} = (\hat{S_1}^{(0)}, \ldots, \hat{S_n}^{(0)}) \in [0,1]^n$, $\epsilon_{tol} > 0$
Ensure: $\hat{\mathbf{S}} = (\hat{S}_1, \ldots, \hat{S}_n)$ and $\hat{S}$

```
repeat
    S^(t+1) ← g^(2)(Ŝ^(t))
    t ← t + 1
until ‖g^(2)(Ŝ^(t)) − Ŝ^(t)‖_1 < ε_tol
Ŝ ← g^(2)(Ŝ^(t))
Ŝ ← (1/n) Σ_j Ŝ_j
```

3 Random clustered networks

In this section, we describe the details of the random clustered network [24], and extend our algorithms for computing the estimates of global cascade on random unclustered networks to the case of random clustered networks. It is a remarkable fact that our algorithm captures the heterogeneity of contact rates on random clustered networks.

Suppose that we have two sequences; a *single-edge* sequence $\mathbf{s} = (s_1, \ldots, s_n)$ and a *triangle-edge* sequence $\mathbf{t} = (t_1, \ldots, t_n)$, where s_i is the number of edges of a node i that are not included in any triangle and t_i is the number of triangles that contain a node i. The random clustered network $G(\mathbf{s}, \mathbf{t})$ with two given sequences $\mathbf{s}$ and $\mathbf{t}$ is a probability space over the set of networks on the node set $V = \{1, \ldots, n\}$. The procedure for generating a random clustered network is as follows: let us consider a random unclustered network $\mathcal{G}_{\mathbf{s}}$ with a given expected degree sequence $\mathbf{s}$. Let us also consider a random network $\mathcal{G}_{(i,j,k)}$ for each $1 \le i < j < k \le n$ so that it has only three edges $\{i,j\}$, $\{i,k\}$, and $\{j,k\}$ that form a triangle with probability $\frac{t_i t_j t_k}{\sum_{x<y} t_x t_y}$, and that has no edges otherwise. We get a random clustered network by taking the union of two networks generated by $\mathcal{G}_{\mathbf{s}}$ and $\mathcal{G}_{(i,j,k)}$ for all $1 \le i < j < k \le n$. An example of the procedure for $n = 4$ is shown in Figure 1. The generating procedure does not make the same edge twice with high probability, which means that each operation that decides whether there is a triangle for each three nodes i, j, k or not does not affect almost all of the others (details are in Appendix C).

3.1 The extensibility of random clustered networks

Note that the random clustered network is a generalised version of the random unclustered network, since $G(\mathbf{s}) = G(\mathbf{s}, \mathbf{0})$. Indeed, the expected degree of each node i in $G(\mathbf{s}, \mathbf{t})$ is $\mathbb{E}[d_i] \approx \sum_j \frac{s_i s_j}{\sum_x s_x} + 2\sum_{j<k} \frac{t_i t_j t_k}{\sum_{x<y} t_x t_y} = s_i + 2t_i$.

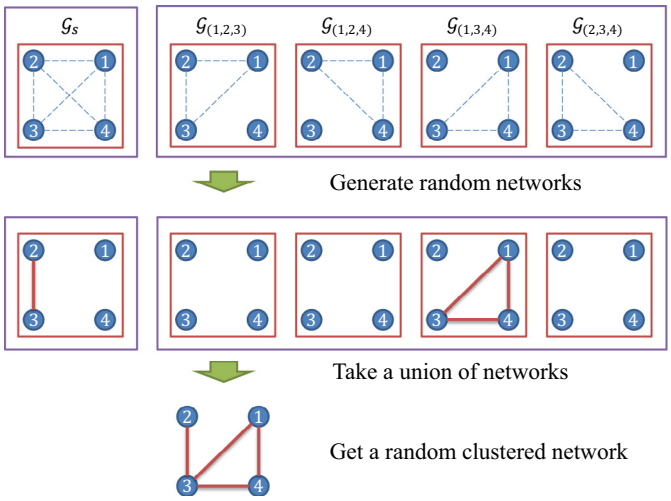


Fig. 1. An example of an overall procedure for generating random clustered network with $n = 4$.

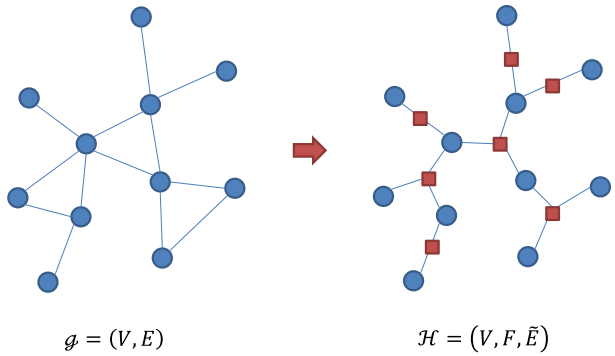


Fig. 2. A network $\mathcal{G} = (V, E)$ and its corresponding factor graph $\mathcal{H} = (V, F, \tilde{E})$. ● represents a node, ■ represents a factor node, and there are edges of $\mathcal{H}$ joining them corresponding to the structure of $\mathcal{G}$.

If there is no edge that is contained in more than one triangle, then the expected degree is equal to $s_i + 2t_i$ for each $i = 1, \ldots, n$.

In a general random clustered network, a branching process is not applicable. The following *factor graph* that we suggest allows us to apply the branching process for random clustered networks. For a given network $\mathcal{G} = (V, E)$, we consider a factor graph $\mathcal{H} = (V, F, \tilde{E})$ that corresponds to $\mathcal{G}$ as follows: we first define

$$\tilde{F} = \{\{i, j, k\} | i, j, k \text{ form a triangle}\} \cup \{\{i, j\} | \{i, j\} \in E \text{ and it is a single-edge}\}.$$

We then define a set $F = \{v_c | c \in \tilde{F}\}$, where each element in F is called a *factor node*. The set of edges of $\mathcal{H}$ is defined by $\tilde{E} = \{\{i, v_c\} | i \in c, i \in V, v_c \in F\}$. Figure 2 shows an example of a network and its corresponding factor graph: there are 4 single-edges and 3 triangle-edges in $\mathcal{G}$, so that the corresponding factor graph has 7 (=4+3) factor nodes and 17 ($= 4 \cdot 2 + 3 \cdot 3$) edges in $\mathcal{H}$. If there are only a small number of triangles, then $\mathcal{H}$ has a locally tree-like structure. Thus, we can use a branching process method at the macroscopic level of the structure of the network, as in the random unclustered network.

3.2 The estimates of global cascades on random clustered networks

We now aim at the calculation of the occurrence probabilities and the expected sizes of global cascades on random clustered networks. We define P_i and P, as in Section 2. To compute P_i, we consider both single-edges and triangle-edges that are incident with i on the random clustered network. For each node $i = 1, \ldots, n$, the probability that i does not induce a global cascade ($= 1 - P_i$) can be represented by the product of the probabilities $P_i^{(1)}$ and $P_i^{(2)}$, which are the probabilities that there is no global cascade that is triggered by a node infected by an initially infectious node i through a single-edge and a triangle-edge respectively. Note that the events that i infects through single-edges and triangle-edges are independent, since we assume that $G(\mathbf{s}, \mathbf{t})$ does not create the same edge more than once.

As in Section 2, $P_i^{(1)} = \prod_{1 \le j \le n, j \ne i} (1 - \frac{s_i s_j}{\sum_x s_x} f(i, j) P_j)$. For the $P_i^{(2)}$, Figure 3 illustrates all cases that an infectious node i and a triangle with nodes i, j, and k do not induce a global cascade. For each $1 \le i < j < k \le n$, the probability that i, j, k do not form a triangle is $1 - \frac{t_i t_j t_k}{\sum_{x<y} t_x t_y}$. The probability that j and k do not induce a global cascade is the sum of the probabilities of (Case I)-(Case IV); (Case I) i infects j and k, but they do not induce a global cascade, (Case II)-(Case III) i infects only one out of j and k, but there is no global cascade induced by either j or k, and (Case IV) i does not infect either j or k. Therefore, the probability that i does not induces a global cascade is

$$\begin{aligned} 1 - P_i = & \prod_{1 \le j \le n, j \ne i} \left(1 - \frac{s_i s_j}{\sum_x s_x} f(i, j) P_j\right) \\ & \prod_{1 \le j < k \le n, j, k \ne i} \Bigg(1 - \frac{t_i t_j t_k}{\sum_{x<y} t_x t_y} \\ & + \frac{t_i t_j t_k}{\sum_{x<y} t_x t_y} \Big(f(i, j)(1 - P_j) f(i, k)(1 - P_k) \\ & + f(i, j)(1 - P_j)(1 - f(i, k))(1 - f(j, k) P_k) \\ & + (1 - f(i, j))(1 - f(k, j) P_j) f(i, k)(1 - P_k) \\ & + (1 - f(i, j))(1 - f(i, k)) \Big) \Bigg). \end{aligned} \tag{9}$$

We can derive the following equation by taking the logarithm of both sides and applying the first-order Taylor

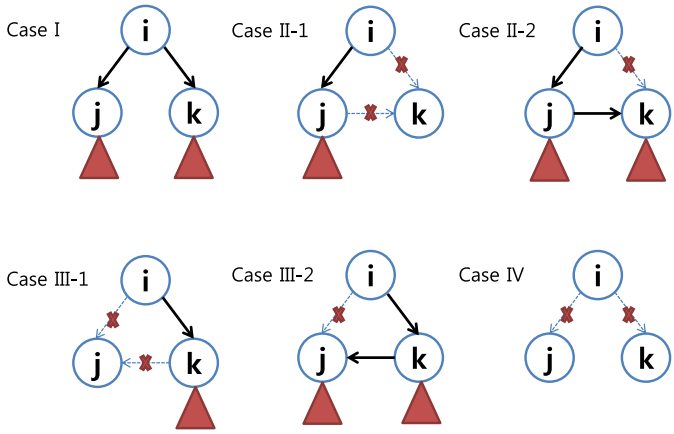


Fig. 3. All possible cases when an infection does not induce a global cascade in a triangle. The solid and dashed lines correspond to the occurrence and non-occurrence of information diffusion. Each ▲ indicates that the corresponding node does not induce a global cascade.

series approximation for the right-hand side.

$$\begin{aligned}\log(1-P_i) = &\sum_j \left(-\frac{s_i s_j}{\sum_x s_x} f(i,j) P_j \right) \\ &+ \sum_{j<k} \Bigg(-\frac{t_i t_j t_k}{\sum_{x<y} t_x t_y} \\ &+ \frac{t_i t_j t_k}{\sum_{x<y} t_x t_y} \Big(f(i,j)(1-P_j) f(i,k)(1-P_k) \\ &+ f(i,j)(1-P_j)(1-f(i,k))(1-f(j,k)P_k) \\ &+ (1-f(i,j))(1-f(k,j)P_j) f(i,k)(1-P_k) \\ &+ (1-f(i,j))(1-f(i,k)) \Big) \Bigg) + \tilde{\epsilon}_i, \end{aligned} \tag{10}$$

where $\tilde{\epsilon}_i$ is the error caused by the Taylor series approximation. For a function $F(j,k)$ of j and k, and for sufficiently large n, $\sum_{j<k}(F(j,k)+F(k,j)) \approx \sum_j \sum_k F(j,k)$ holds. Then, from equation (10), we obtain that

$$\log(1-P_i) \approx -\sum_j k(i,j) P_j + \tilde{\epsilon}_i, \tag{11}$$

where

$$\begin{aligned} k(i,j) = &\frac{s_i s_j}{\sum_x s_x} f(i,j) + \frac{t_i t_j}{\sum_{x<y} t_x t_y} \Bigg(f(i,j) \sum_k t_k \\ &- \frac{1}{2} f(i,j) \sum_k t_k f(i,k) P_k + (1-f(i,j)) \\ &\times \sum_k t_k f(i,k)(1-P_k) f(k,j) \Bigg). \end{aligned} \tag{12}$$

The error analysis of equation (11) is straightforward from Section 2. Moreover, computing the sizes of a global cascade can be expressed with the same $k(i,j)$ using the backward branching process method. To compute the estimates of P_i and S_j, we use iterative algorithms by solving the fixed point problem ignoring the error term, as in Section 2. Hence, we can solve P_i and S_j by solving the following equations iteratively.

$$P_i \approx 1 - \exp\left\{ -\sum_j k(i,j) P_j \right\}, \tag{13}$$

$$S_j \approx 1 - \exp\left\{ -\sum_i k(i,j) S_i \right\}. \tag{14}$$

Then, P and S satisfy that

$$P \approx 1 - \frac{1}{n} \sum_i \exp\left\{ -\sum_j k(i,j) P_j \right\}, \tag{15}$$

$$S \approx 1 - \frac{1}{n} \sum_j \exp\left\{ -\sum_i k(i,j) S_i \right\}. \tag{16}$$

Note that if one ignores all the triangle-edges and sets $\mathbf{s} = \mathbf{w}$, then this is identical to the results in the unclustered random network case. In addition, our argument can be applied to any specific substructure of a network of finite size, including cliques of size larger than 3, cycles, and motifs, rather than triangles, which are used in the above argument. This framework can explain specified small-scale structures of a network. More specifically, when there is a modular structure of a network that is locally tree-like, one can determine the values of $k(i,j)$ and analyse the occurrence of a global cascade using our formula.

3.3 When the network topology is given

Until now, we have provided a formula for estimating the occurrence probabilities and the expected sizes of global cascades on random networks. For analysing real-world networks, whose network topology is given and not random, having a formula for estimating the probability and the size of a global cascade would be useful.

Suppose that the actual network topology is given, and that the outbreak of the propagation begins with a single infectious node. When information spreads through the given network, the dynamics of information diffusion can be represented by a locally tree-like structure that is a subgraph of the given network [6,38]. In this respect, we obtain a simple formula for estimates of a global cascade. For each $i = 1, \ldots, n$, let $n(i)$ be the set of neighbours of i. Then,

$$1 - P_i = \prod_{j \in n(i)} \big(1 - f(i,j) P_j\big), \tag{17}$$

$$1 - S_j = \prod_{i \in n(j)} \big(1 - f(j,i) S_i\big), \tag{18}$$

for each $i = 1, \ldots, n$, and $j = 1, \ldots, n$.

In equation (17), the probability that an infectious node i induces no global cascade can be represented as the product of the probabilities that each node j, that is infected by its neighbour i, induces no global cascade. Similarly, in equation (18), any non-initial infectious node j is contained in a global cascade if there is a node i that belongs to a global cascade and i infects j.

An important application of this work is that S_j can be used to measure the influence. When one restricts some S_js as specified values, the solutions of equation (18) indicate the influences for each corresponding individuals under the given condition. For example, suppose that one chooses a set of individuals A and takes $S_{j^*} = 1$ for each $j^* \in A$. The solution of equation (18) with this restriction gives a measure of the influence from the users who are selected first. When S_j is large for some node j, it indicates that j is influenced a lot by A, the set of selected nodes. Furthermore, our result can be applied to the influence maximisation problem [39]. By using S_j, one can determine the set of candidates that are highly influential under some constraints. This plays a role in a preprocessing procedure to solve the influence maximisation problem more efficiently.

4 Conditions for phase transitions for various contact rates

In this section, we identify the conditions for phase transitions under random unclustered and clustered networks, when the contact rates are heterogeneous. We consider three different types of basic contact rates (i) c/w_i, (ii) c/w_j, and (iii) $(c/2)/w_i + (c/2)/w_j$, where c is the *infect/receive ability* of information diffusion and w_i is the number of neighbours for each node i. Under model (i), the average infection for each node is equivalent to c, but the susceptibility differs in general – heterogeneous susceptibility. Under model (ii), all nodes are likely to be influenced by c neighbours on average, and this model shows heterogeneous infectivity. And under model (iii), we combine them so that it describes the multiple contact scenario. It is heterogeneous infectivity and susceptibility simultaneously. Because each node can affect or be affected by a limited number of its neighbours in reality, these scenarios are reasonable. We parameterise this limit as c in the above three models. These three contact rates are meaningful and applicable models to analyse the spread of information. Let us now identify a condition on c for phase transitions under random unclustered and clustered networks for various contact rates.

4.1 The basic reproduction number

The basic reproduction number of information diffusion, R_0, is the expected number of secondary infections generated by a single infectious node. This has been used to estimate the phase transition point (or the *epidemic threshold*, equivalently) of information diffusion [34,40–43]. Recently, Prakash et al. [42] provided a formula to find the phase transition points for several virus propagation models on arbitrary networks using R_0. They showed that the effect of an undirected underlying topology is determined only by the largest eigenvalue of the adjacency matrix. However, their analysis is based on a strong assumption that the likelihood that a node is infected is independent of the status of its neighbours.

Note that by the theory of difference equations, a dynamical system $g : [0,1]^n \to [0,1]^n$ given by $\mathbf{P}^{(t+1)} = g(\mathbf{P}^{(t)})$ is asymptotically stable at an equilibrium point $\mathbf{x} \in [0,1]^n$ if the largest eigenvalues of the Jacobian $\partial = \nabla g(\mathbf{x})$ of the vector-valued function g at $\mathbf{x}$ is less than 1, where $\partial_{ij} = [\nabla g(\mathbf{x})]_{i,j} = \frac{\partial}{\partial \mathbf{x}_j} g(\mathbf{x}_i)$. Thus, $\tilde{\mathbf{P}} = 0$ is asymptotically stable where $\rho(\partial)$, the largest eigenvalue of ∂, is less than 1. Therefore, R_0 is given by $\rho(\partial)$ and $R_0 = 1$ is the phase transition point of the corresponding discrete-time SIR process.

Unlike in the previous studies on the basic reproduction number, we provide an estimate of the basic reproduction number using arbitrary contact rates $f(i,j)$ between an infectious node i and a susceptible node j. Then, we derive the partial derivatives $\partial_{ij} = k(i,j)$ for random clustered networks using equations (13) and (14). In particular, $\partial_{ij} = \frac{w_i w_j}{Vol(G)} f(i,j)$ for random unclustered networks from equations (3) and (7). As a result, the phase transition point of information diffusion in random clustered networks with arbitrary contact rates is $R_0 = \rho(\partial)$, where $\partial = [k(i,j)]_{i,j=1,\ldots,n}$ and $k(i,j)$ is defined in equation (12).

One can easily check that it is an extended version of the analysis on phase transitions in random unclustered networks [34], since the phase transition point R_0 is asymptotically equilvalent to the previous work when the $f(i,j)$ is constant and the triangle sequence $\mathbf{t} \equiv 0$ in our setting. Moreover, by using the fact $\rho(\partial) \leq \min\{\sum_i \partial_{ij}, \sum_j \partial_{ij}\}$, we conclude that there is no global cascade if $c < 1$ for model (i) and (ii) since $\sum_i \frac{cw_i}{Vol(G)} = \sum_j \frac{cw_j}{Vol(G)} = c$. Similarly, there is no global cascade if $c < 1$ for model (iii). We analyse the phase transitions for models (i), (ii) and (iii) more rigorously in the rest of this section.

4.2 Random unclustered networks

When c is a constant: by using equations (4) and (8), we can easily compute the occurrence probabilities and the expected sizes of global cascades on the random unclustered networks asymptotically for model (i) and (ii):

$$\text{(i) } P \approx 1 - e^{-cP},\ S \approx 1 - \frac{1}{n}\sum_j e^{-\left(\frac{cw_j}{Vol(G)}\right)nS}. \quad (19)$$

$$\text{(ii) } P \approx 1 - \frac{1}{n}\sum_i e^{-\left(\frac{cw_i}{Vol(G)}\right)nP},\ S \approx 1 - e^{-cS}. \quad (20)$$

A previous study on the configuration model [14] has shown that the probability of a global cascade is independent of heterogeneity in susceptibility, and the expected

size of a global cascade is independent of heterogeneity in infectivity for any Erdős-Rényi random network. We prove the same thing for model (i) and (ii) for the Chung-Lu model, which allows arbitrary degree distributions. For model (i), the P_is are the same and P is independent of the network structure, since it is not just a heterogeneous susceptibility case but a homogeneous infectivity case. Similarly for model (ii), the S_js are the same and S is independent of the network structure, since it is not just a heterogeneous infectivity case but a homogeneous susceptibility case. Thus, P for model (i) and S for model (ii) can be computed using the same formula with Erdős-Rényi random networks; Erdős and Rényi showed that $P = 1 - e^{-cP}$ and $S = 1 - e^{-cS}$ for Erdős-Rényi random networks and showed that the critical point of both P and S is $c = 1$ [19]. It implies that the phase transition point of both P for model (i) and S for model (ii) is $c = 1$.

We now show that the phase transition point of P for model (ii) is also $c = 1$. In the Chung-Lu network, the P_is for model (ii) are not homogeneous, unlike in the Erdős-Rényi random network. In order to identify the conditions for phase transition, we use a fixed-point analysis.

Let d be a metric induced by the L_1-norm. That is, $d(x_1, x_2) := ||x_1 - x_2||_1 = |x_{11} - x_{21}| + \ldots + |x_{1n} - x_{2n}|$. Let $\Phi : [0,1]^n \to [0,1]^n$ be a function that satisfies

$$\Phi(\mathbf{P}) = \left(1 - \exp\left\{-\sum_j \left(\frac{w_i w_j}{Vol(G)}\right) f(i,j) P_j\right\}\right)_{i=1,\ldots,n}$$

for all $\mathbf{P} = (P_1, \ldots, P_n) \in [0,1]^n$. If we suppose that $f(i,j) = c/w_j$, then

$$\begin{aligned} d(\Phi(\mathbf{P}), \Phi(\mathbf{P}')) &\leq \sum_{i=1}^{n} \left| \frac{cw_i}{Vol(G)} \sum_j P_j - \frac{cw_i}{Vol(G)} \sum_j P'_j \right| \\ &\leq \sum_i \frac{cw_i}{Vol(G)} \sum_j |P_j - P'_j| \\ &= c \sum_j |P_j - P'_j| \\ &= cd(\mathbf{P}, \mathbf{P}'), \end{aligned} \tag{21}$$

because $|e^{-x_1} - e^{-x_2}| \leq |x_1 - x_2|$ where $x_1, x_2 > 0$. Thus, Φ is a contraction map if $c < 1$. Hence, for any initial value $\mathbf{P} = (P_1, \ldots, P_n) \in [0,1]^n$, the iteration must converge to a unique fixed point within $[0,1]^n$ by Banach's fixed point theorem. Since $\Phi(0) = 0$, we conclude that 0 is a stable fixed point and a global cascade does not occur if $c < 1$. Note that we already have the same result in Section 4.1 to derive an upper bound of the largest eigenvalue of the Jacobian matrix.

To obtain the sufficient and necessary conditions for phase transitions, we also identify the conditions for inducing global cascades. Let $\Psi : [0,1]^n \to [0,1]$ be a function that satisfies

$$\Psi(\mathbf{P}) = 1 - \frac{1}{n}\sum_i \exp\left\{-\sum_j \frac{w_i w_j}{Vol(G)} f(i,j) P_j\right\} - \frac{1}{n}\sum_i P_i$$

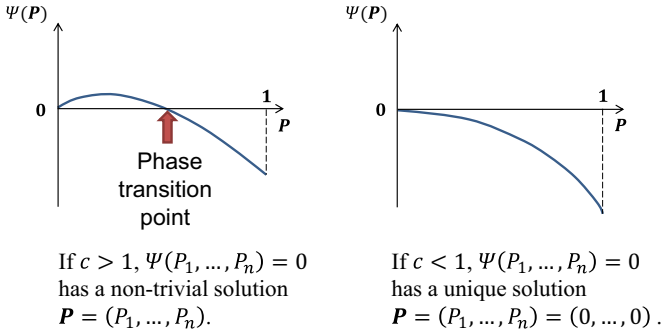


Fig. 4. For model (i) and (ii), there is a non-trivial solution of $\Psi(P) = 0$ if $c > 1$. Here, $c = 1$ is the phase transition point for both model (i) and (ii). Note that $\Psi : [0,1]^n \to [0,1]$ and the dimension of the x-axis is n in the above figures.

for all $\mathbf{P} = (P_1, \ldots, P_n) \in [0,1]^n$. For each $j = 1, \ldots, n$

$$\begin{aligned} \frac{\partial \Psi(\mathbf{P})}{\partial P_j}\Big|_{\mathbf{P}=0} &= \frac{1}{n}\sum_i \frac{w_i w_j}{Vol(G)} f(i,j) - \frac{1}{n} \\ &= \frac{1}{n}(c-1), \end{aligned} \tag{22}$$

where $f(i,j) = c/w_j$. Since $\Psi(0) = 0$ and $\Psi(\mathbf{1}) < 0$, there exists a non-zero solution $\mathbf{P} \in [0,1]^n$ so that $\Psi(\mathbf{P}) = 0$ if $c > 1$. Figure 4 shows this result and it implies that there is a non-trivial solution P of equation (19) if $c > 1$. Therefore, we conclude that the phase transition point of P for model (ii) is $c = 1$, which is the same result as for model (i). Moreover, the phase transition points of S for models (i) and (ii) are also the same, because the equations (19) and (20) are symmetric.

For model (iii), it is generally supposed that the contact rate of information diffusion is $f(i,j) = c_1/w_i + c_2/w_j$. We then find by the same argument that a phase transition occurs when $c_1 + c_2 > 1$, because there is no global cascade if $\rho(\partial) \leq \min\{\sum_i \partial_{ij}, \sum_j \partial_{ij}\} = c_1 + c_2 < 1$ and $\frac{\partial \Psi(\mathbf{P})}{\partial P_j}\Big|_{\mathbf{P}=0} = n(c_1 + c_2 - 1) > 0$ if $c_1 + c_2 > 1$. For simplicity, if we assume that $c_1 = c_2 = c/2$ for some constant c as in model (iii), then $c = 1$ is the phase transition point of this model.

When c follows a power-law distribution: however, the infect/receive abilities of information diffusion are not identical for real-world applications. We observe that in many cases these abilities follow the power-law distribution with heavy-tails [22]. Our algorithm is still valid even if both degree and infect/receive ability c follow power-law distributions with the assumption that the power-law exponent of c is larger than that of the degree distribution. Details are in Appendix D.

4.3 Random clustered networks

We extend the analysis on phase transitions of information diffusion in random unclustered networks to the random

clustered network case. We show that the conditions for phase transitions are affected by the clustering coefficient of a network. The clustering coefficient is a measure of the degree to which nodes in a network tend to cluster together, which is defined as below:

$$\mathcal{C} = \frac{3 \times \text{the number of triangles}}{\text{the number of connected triples}}.$$

We first assume that there is a constant γ so that $w_i = s_i + 2t_i$ and $t_i = \gamma w_i$ for all $i = 1, \ldots, n$. Consider a random clustered network with two given sequences $(s_1, \ldots, s_n) = ((1-2\gamma)w_1, \ldots, (1-2\gamma)w_n)$ and $(t_1, \ldots, t_n) = (\gamma w_1, \ldots, \gamma w_n)$. The expected clustering coefficient is then

$$\mathcal{C} = \frac{\sum_i t_i}{\sum_i w_i(w_i-1)/2} = \frac{2\gamma \sum_i w_i}{\sum_i w_i(w_i-1)} \propto \gamma.$$

Hence, the expected clustering coefficient is increasing linearly with the value of γ, and we can regard γ as the degree of clustering. Let $\Psi : [0,1]^n \to [0,1]$ be a function that satisfies

$$\Psi(\mathbf{P}) = 1 - \frac{1}{n}\sum_i \exp\left\{-\sum_j k(i,j)P_j\right\} - \frac{1}{n}\sum_i P_i$$

for all $\mathbf{P} = (P_1, \ldots, P_n) \in [0,1]^n$, where $k(i,j)$ is from equation (11). For each $j = 1, \ldots, n$, we then approximate a value for $\frac{\partial \Psi(\mathbf{P})}{\partial P_j}$ at $\mathbf{P} = 0$:

$$\begin{aligned}\frac{\partial \Psi(\mathbf{P})}{\partial P_j}\Big|_{\mathbf{P}=0} &\approx \frac{\partial}{\partial P_j}\left(-\frac{1}{n}\sum_i\left(-\sum_j k(i,j)P_j\right)\right)\Big|_{\mathbf{P}=0} \\ &= \sum_i \frac{1}{n}\Bigg((1-2\gamma)\frac{w_i w_j}{Vol(G)}f(i,j) \\ &\quad + \frac{w_i w_j}{\sum_{x<y} w_x w_y}\Bigg(\gamma f(i,j)Vol(G) \\ &\quad + \gamma(1-f(i,j))\sum_k w_k f(i,k)f(k,j)\Bigg)\Bigg) - \frac{1}{n}.\end{aligned} \tag{23}$$

For model (i), we substitute $f(i,j)$ with c/w_i:

$$\begin{aligned}\frac{\partial \Psi(\mathbf{P})}{\partial P_j}\Big|_{\mathbf{P}=0} &\approx \frac{1-2\gamma}{n}\sum_i \frac{cw_j}{Vol(G)} + \frac{\gamma}{n}\sum_i\Bigg\{\frac{cw_j Vol(G)}{\sum_{x<y} w_x w_y} \\ &\quad + \frac{c^2 w_j}{\sum_{x<y} w_x w_y}\left(1-\frac{c}{w_i}\right)\Bigg\} - \frac{1}{n} \\ &\approx \frac{cw_j}{Vol(G)} + \gamma\frac{nc^2 w_j}{\sum_{x<y} w_x w_y} - \frac{1}{n}.\end{aligned} \tag{24}$$

By applying the chain rule, we derive a value for $\frac{\partial \Psi(\mathbf{P})}{\partial P}$ at $\mathbf{P} = 0$:

$$\begin{aligned}\frac{\partial \Psi(\mathbf{P})}{\partial P}\Big|_{\mathbf{P}=0} &= \sum_j \frac{\partial \Psi(\mathbf{P})}{\partial P_j}\frac{\partial P_j}{\partial P} \\ &\approx \sum_j\left(\frac{cw_j}{Vol(G)} + \gamma\frac{nc^2 w_j}{\sum_{x<y} w_x w_y} - \frac{1}{n}\right)n \\ &= n\left(c + \frac{2\gamma nc^2}{Vol(G)} - 1\right).\end{aligned} \tag{25}$$

$\Psi(0) = 0$ and $\Psi(\mathbf{1}) < 0$ imply that there exists a non-zero solution $\mathbf{P} \in [0,1]^n$ so that $\Psi(\mathbf{P}) = 0$ if $c + \frac{2\gamma nc^2}{Vol(G)} > 1$, and it follows that there is a global cascade if $c > 1 - \mathcal{O}(\frac{\gamma n}{Vol(G)})$. Therefore, the phase transition point is less than 1, but it is close to 1 if the average degree $Vol(G)/n$ is sufficiently large. For example, if $Vol(G)/n = 60$ and $\gamma = 0.3$, then $c + \frac{2\gamma nc^2}{Vol(G)} > 1$ is equivalent to $c > 0.99$.

For model (ii), we substitute $f(i,j)$ with c/w_j:

$$\begin{aligned}\frac{\partial \Psi(\mathbf{P})}{\partial P_j}\Big|_{\mathbf{P}=0} &\approx \frac{1-2\gamma}{n}\sum_i \frac{cw_i}{Vol(G)} + \frac{\gamma}{n}\sum_i\Bigg\{\frac{cw_i Vol(G)}{\sum_{x<y} w_x w_y} \\ &\quad + \frac{nc^2 w_i}{\sum_{x<y} w_x w_y}\left(1-\frac{c}{w_j}\right)\Bigg\} - \frac{1}{n} \\ &\approx \frac{1}{n}\left(c + \frac{2\gamma nc^2}{Vol(G)} - 1\right),\end{aligned} \tag{26}$$

and $\frac{\partial \Psi(\mathbf{P})}{\partial P}\Big|_{\mathbf{P}=0}$ has the same value by applying the chain rule. Therefore, the phase transition point for model (ii) is also $c = 1 - \mathcal{O}(\frac{\gamma n}{Vol(G)})$, as in model (i). When the contact rate is $c_1/w_i + c_2/w_j$, we obtain by the same argument that

$$\frac{\partial \Psi(\mathbf{P})}{\partial P}\Big|_{\mathbf{P}=0} \approx n\left(c_1 + c_2 + \frac{2\gamma nc^2}{Vol(G)} - 1\right). \tag{27}$$

Hence, we obtain the phase transition point that $c_1 + c_2 > 1 - \frac{2\gamma nc^2}{Vol(G)}$, or $c > 1 - \mathcal{O}(\frac{\gamma n}{Vol(G)})$ if $c_1 = c_2 = c/2$.

As a result, we have shown that the phase transition point of information diffusion in a random clustered network is strictly lower than that of the corresponding random unclustered network. Therefore, the condition required to induce a phase transition is strictly easier for networks with higher clustering coefficients for contact rates (i)–(iii). However, if either the average degree $Vol(G)/n$ is large or the clustering coefficient is small, the effect of clustering is not significant.

5 Experimental results

5.1 Methods

In order to demonstrate the performance of our proposed algorithm, we investigated the results on real-world social network topologies. By crawling the web, we obtained

two network topologies, Facebook [44] and MySpace [45] friendship networks; the Facebook friendship network was obtained from the New Orleans regional network in Facebook during December 29th, 2008 and January 3rd, 2009, and it consists of 63,392 users and 816 886 friendship links. The MySpace friendship network was obtained by crawling the MySpace online web site from September to October, 2006, and it consists of 100 000 individuals and 6 854 231 social relationships.

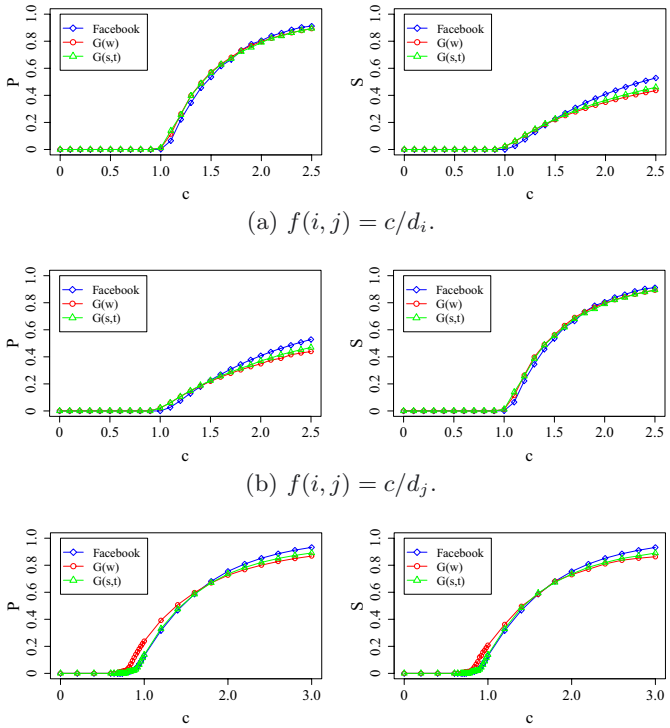
We applied our algorithm to the above two network topologies and the corresponding random unclustered and clustered networks with the same number of nodes and expected substructures. Let $\mathbf{w} = (w_1, \ldots, w_n)$ be the degree sequence of a friendship network. We could then generate the random unclustered network with a given expected degree sequence $\mathbf{w}$. In addition, the random clustered network can be generated with two sequences, $\mathbf{s} = (s_1, \ldots, s_n)$ and $\mathbf{t} = (t_1, \ldots, t_n)$. We estimated $\mathbf{s}$ and $\mathbf{t}$ efficiently as follows: for a given network topology, we found and deleted triangles one by one from the network. We then defined t_i as half the number of deleted edges that were incident with i and $s_i = w_i - 2t_i$ for each node i. Note that s_i and t_i are accurate when the triangles in the network do not have pairwise overlap. Once we estimated the degree sequences for the two real networks, s_i and t_i as well as w_i follow the power-law distribution.

In the analysis of the occurrence of global cascades, we propose a novel iterative algorithm to compute the probability that an originally infected node i induces a global cascade ($=P_i$) and the probability that a primarily uninfected node j is contained in a global cascade ($=S_j$). For the simulation, we used three contact rates $f(i,j)$ between nodes i and j, as discussed earlier: (i) c/w_i; (ii) c/w_j; and (iii) $(c/2)/w_i + (c/2)/w_j$, where c is the infect/receive ability.

In the experiments, we performed 10 000 simulation trials for each value of c, while this value varied from 0 to 3. We compared the simulation results of the original network and two synthetic networks; random unclustered and clustered networks. We also conducted Monte-Carlo experiments on information diffusion in real-world social network topologies, and these results are compared with the computed values of the estimations of global cascades based on our proposed algorithm. By varying the clustering coefficient values, we found that networks with higher clustering coefficients have lower phase transition points for global cascades. Finally, in order to check the effect of the clustering coefficient on the degree correlation in random clustered networks, we used several local structures; the actual degree of the node and that of its neighbours, and the degree correlation function [46].

5.2 Simulation results

Here, we show that the random clustered network is a good underlying network structure for estimating the information diffusion in networks (Fig. 5) and its accuracy is good for different initial infectious nodes (Fig. 6). We also studied the effect of clustering on phase transition



(a) $f(i,j) = c/d_i$.

(b) $f(i,j) = c/d_j$.

(c) $f(i,j) = (c/2)/d_i + (c/2)/d_j$.

Fig. 5. P and S for various contact rates where c ranges from 0 to 3 on Facebook friendship network. Simulation trials are performed on Facebook friendship network (◊). The other two lines represent the predictions by the algorithm based on the random unclustered network $G(\mathbf{w})$ (○) and the random clustered network $G(\mathbf{s}, \mathbf{t})$ (△), respectively.

and checked that the phase transition points are lower for higher clustering (Fig. 7). We then analysed why this phenomenon happens to examine the degree correlation of random clustered networks (Fig. 8).

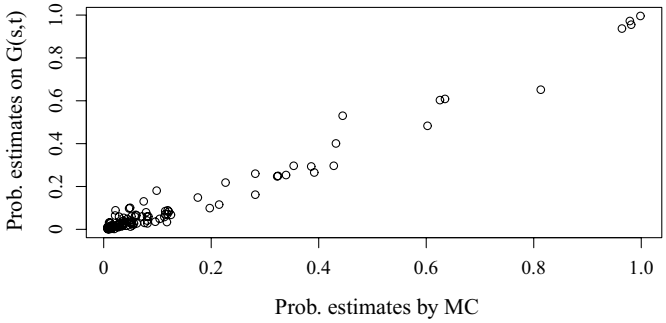
First, we check the final cascade sizes of information diffusion for each value of c were concentrated around their mean once a global cascade was triggered. Table 1 shows that S on Facebook and MySpace friendship networks are tightly concentrated when they occur at various contact rates. Since the variations are very small, it is meaningful to analyse S of information in real-world networks.

Through equations (15) and (16), we estimated the P and S of information diffusion in random clustered networks. Figure 5 represents these estimates for Facebook friendship networks when the contact rate is c/w_i (1st row), c/w_j (2nd row), and $(c/2)/w_i + (c/2)/w_j$ (3rd row). It shows a comparison of experiments and theories for the values of c. The prediction of the cascading behaviour using the random clustered network is more accurate than that obtained with the random unclustered network. Hence, it is worthwhile to consider the random clustered network as a contact network to model the spread of information in networks.

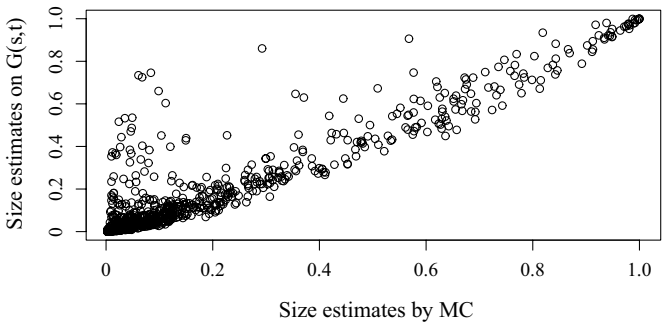
In Figure 6, we plotted the estimates using Algorithms 1 and 2 against the estimates obtained by the Monte-Carlo experiments (denoted as MC). The P_i and S_j

Table 1. The experimental results for model (i), (ii), and (iii) with various infect/receive abilities. The expected sizes of global cascades on Facebook and MySpace friendship networks are tightly concentrated around their mean for all cases.

Contact rate	Facebook		MySpace	
	Exp. size	Std. dev.	Exp. size	Std. dev.
$1.0/w_i$	0.012	0.002	0.012	0.001
$1.5/w_i$	0.227	0.005	0.081	0.002
$2.0/w_i$	0.409	0.004	0.148	0.002
$2.5/w_i$	0.529	0.003	0.203	0.002
$1.0/w_j$	0.018	0.009	0.002	0.001
$1.5/w_j$	0.542	0.008	0.582	0.002
$2.0/w_j$	0.809	0.003	0.819	0.002
$2.5/w_j$	0.914	0.002	0.912	0.002
$(1.0/2)/w_i + (1.0/2)/w_j$	0.129	0.002	0.060	0.002
$(2.0/2)/w_i + (2.0/2)/w_j$	0.761	0.002	0.274	0.001
$(3.0/2)/w_i + (3.0/2)/w_j$	0.931	0.001	0.433	0.001



(a) Estimates of P_is using Algorithm 1 and MC.

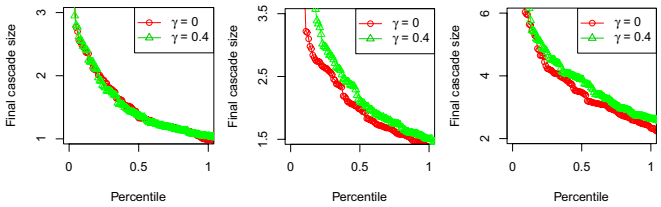


(b) Estimates of S_js using Algorithm 2 and MC.

Fig. 6. Comparisons of the estimated values of P_i and S_j based on our algorithms and that of Monte-Carlo simulations.

were estimated by our algorithms in equations (13) and (14) with the degree sequence of Facebook friendship network. Equations (17) and (18) were used for the Monte-Carlo simulations on the real network topology. We checked that there is a strong linear relationship, which means that the estimation accuracies are high for randomly selected initial infectious nodes even if an initial infectious node was given.

In addition, we conducted a set of simulations to quantify how the clustering coefficient affects the conditions for the phase transitions. We considered a power-law degree distribution $(w_1, \ldots, w_n)$ with power-law exponent 3, average degree 10, maximum degree 50 and $n = 10\,000$. As

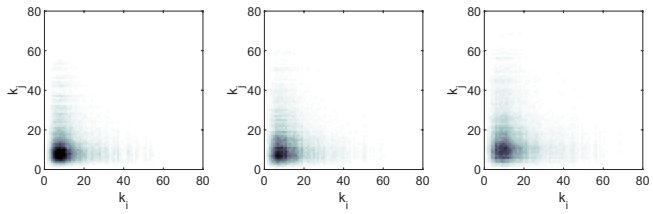


(a) $c = 0.96$. (b) $c = 1$. (c) $c = 1.04$.

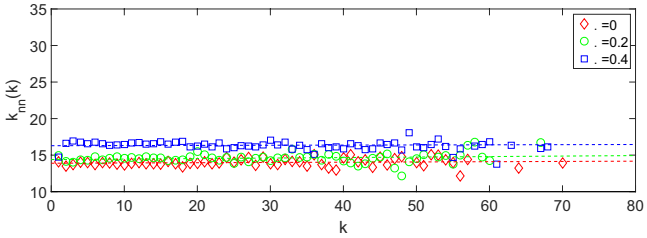
Fig. 7. P on random clustered networks for model (ii) when c varies from 0.96 to 1.04 and γ is 0 (○) and 0.4 (△).

shown in Section 4.3, the clustering coefficient of the random clustered network using two sequences $(s_1, \ldots, s_n) = ((1-2\gamma)w_1, \ldots, (1-2\gamma)w_n)$ and $(t_1, \ldots, t_n) = (\gamma w_1, \ldots, \gamma w_n)$ is proportional to $\gamma \in [0, 0.5]$. In order to conduct a set of experiments on the effect of clustering, we set the value γ to 0 (random unclustered network) and 0.4 (random clustered network). Figure 7 shows the probabilities of the occurrences of global cascades for model (ii) on random clustered networks. From the results in Section 4.2, we estimate by the proposed algorithm that a phase transition occurs when $c \approx 1 - \frac{2\gamma n}{Vol(G)} = 1 - \frac{2\gamma}{20} = 1$ and 0.96 for $\gamma = 0$ and 0.4, respectively. In Figure 7a, a phase transition does not occur in either case. In Figure 7b, the cascade sizes of the random clustered network are strictly larger than that of the random unclustered network. In Figure 7c, the cascade sizes seem to reach a positive fraction of the population, which means that phase transitions occur. Here we checked that a higher clustering can induce a larger cascade size for a random clustered network at the beginning of the phase transition, i.e., the value of c is around 1.

In Figure 8, we generated random clustered networks with the same conditions used in Figure 7. In order to check the clustering effect on the degree correlations on random clustered networks, we quantified the actual degree pair (k_i, k_j) in random networks with different clustering coefficients, where k_i is the actual degree of the node. The quantities are marked by colour in Figures 8a–8c. In the same networks, we also quantified



(a) $\gamma = 0$. (b) $\gamma = 0.2$ (c) $\gamma = 0.4$.



(d) Degree correlations.

Fig. 8. The actual degree of a node and that of its neighbours (a)–(c) and degree correlations (d) on a random clustered network with several γ values varied from 0 to 0.4.

$k_{nn}(k)$ (named degree correlation function in Ref. [46]), which is the average degree of the neighbours of all nodes with degree k, and these results are in Figure 8d. As a result, it was found that modifying the clustering coefficient does not affect the degree correlation in the Chung-Lu network. Interestingly, the higher the clustering coefficient, the more distributed connections were shown in Figures 8a–8c. It would be one of the reasons why the effect of clustering on the percolation threshold in a random clustered networks (i.e., a generalised version of the Chung-Lu model) is contrary to that in a random graph with clustering (i.e., a generalised version of the configuration model), which is known as the epidemic threshold analysis based on the configuration model showing that clustering raises the threshold [47,48].

One of the main characteristic of the Chung-Lu model is that the probability of having an edge between two vertices is independent of the other edges, unlike the configuration model. This is because the probability depends only on the expected degree of the two vertices and the sum of expected degrees in the network, while the configuration model does not. If similar independence exists in the random clustered network model we used, then we can expect some interesting results different from the previous work. More numerical analysis in Appendix E consolidates our results.

6 Conclusion

In this paper, we have proposed a novel iterative algorithm for predicting the probability and size of a global cascade of information diffusion in a random clustered network, when an initial infectious node is given or selected at random. To account for the heterogeneity, the contact rates between two individuals are assumed to be functions that depend on their local information, such as heterogeneous infectivity, heterogeneous susceptibility, and so on. We have found the conditions required to induce a phase transition with general contact rates. Interestingly, we have shown that random clustered networks with higher clustering coefficients have lower phase transition points than unclustered networks. The experimental results of real-world network topologies and synthetic networks confirmed that our algorithm produces accurate estimates.

K.J. is with the Department of Electrical and Computer Engineering, ASRI, Seoul National University, Seoul, Korea. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (2016R1A2B2009759), and supported by the Brain Korea 21 Plus Project in 2016.

Appendix A: Taylor series approximation error

Error terms $\tilde{\epsilon}_i$ and ϵ_i in equation (3) satisfy

$$\tilde{\epsilon}_i = \mathcal{O}\left(\sum_j (k(i,j))^2\right)$$

and

$$\epsilon_i = \mathcal{O}\left(\exp\left\{-\sum_j k(i,j)P_j\right\}(\exp(\tilde{\epsilon}_i) - 1)\right). \quad \text{(A.1)}$$

In the error term ϵ_i in equation (A.1),

$$\begin{aligned}\exp\left\{-\sum_j k(i,j)P_j\right\} &\geq \exp\left\{-\sum_j k(i,j)\right\} \\ &\geq \exp\left\{-\sum_j \frac{w_i w_j}{Vol(G)}\right\} \\ &= \exp(-w_i) = o(1)\end{aligned}$$

if $w_i = \omega(\log n)$. Thus, the error term converges to zero for a slightly dense graph, where n is sufficiently large. In fact, ϵ_i is negligible even if there are many nodes with small degrees (e.g., a degree sequence follows a power-law distribution). In the error term $\tilde{\epsilon}_i$ in equation (A.1), $\sum_j (k(i,j))^2 \leq \sum_j (\frac{w_i w_j}{Vol(G)})^2 \leq \max_i(w_i)^2 \sum_j (\frac{w_j}{Vol(G)})^2$. If we approximate that a given degree sequence is a sequence of observations of a continuous random variable X with the probability density function $f_X(x)$, then

$$\sum_{j:j\neq i}\left(\frac{w_j}{Vol(G)}\right)^2 \approx \frac{n\int_{-\infty}^{\infty} x^2 f_X(x)dx}{(n\int_{-\infty}^{\infty} x f_X(x)dx)^2} \propto \frac{1}{n}, \quad \text{(A.2)}$$

where the mean and the variance of X is finite. Therefore, $\tilde{\epsilon}_i$ is only $\mathcal{O}(\max_i\{w_i^2\}/n)$ in this setup, and it is $o(1)$

if $\max_i\{w_i\} = o(n^{1/\tau})$ for some $\tau > 2$. In many complex networks, $\tilde{\epsilon}_i$ becomes $o(1)$ so that ϵ_i in equation (A.1) is negligible. It means that we can solve P_i approximately from equation (3).

Appendix B: Correctness of the algorithm

To show the correctness of Algorithm 1, we prove that $\frac{1}{n}\|\hat{\mathbf{P}} - \mathbf{P}\|_1 = \frac{1}{n}\sum_i |\hat{P}_i - P_i|$ is $o(1)$. By using the triangle inequality,

$$\begin{aligned}\|\hat{\mathbf{P}} - \mathbf{P}\|_1 &= \|g^{(1)}(\hat{\mathbf{P}}) - g^{(0)}(\mathbf{P})\|_1 \\ &\le \|g^{(1)}(\hat{\mathbf{P}}) - g^{(1)}(\mathbf{P})\|_1 + \|g^{(1)}(\mathbf{P}) - g^{(0)}(\mathbf{P})\|_1 \\ &\le \|g^{(1)}(\hat{\mathbf{P}}) - g^{(1)}(\mathbf{P})\|_1 + n\bar{\epsilon}, \end{aligned} \tag{B.1}$$

since $\|g^{(1)}(\mathbf{P}) - g^{(0)}(\mathbf{P})\|_1 = \|g^{(1)}(\mathbf{P}) - \mathbf{P}\|_1 = \|\epsilon_i\|_1 = n\bar{\epsilon}$ from the definition of $g^{(1)}$ and equation (3). By the best linear approximation of $g^{(1)}$ near the point $\hat{\mathbf{P}}$, we have $g^{(1)}(\hat{\mathbf{P}}) = g^{(1)}(\mathbf{x}) + \nabla g^{(1)}(\hat{\mathbf{P}})(\hat{\mathbf{P}} - \mathbf{x}) + o(\|\hat{\mathbf{P}} - \mathbf{x}\|)$ for $\mathbf{x} \in [0,1]^n$. Thus,

$$\|\hat{\mathbf{P}} - \mathbf{P}\|_1 \le \rho\|\hat{\mathbf{P}} - \mathbf{P}\|_1 + n\bar{\epsilon} + o(\|\hat{\mathbf{P}} - \mathbf{P}\|_1), \tag{B.2}$$

where $\rho = \rho(\nabla g^{(1)}(\hat{\mathbf{P}}))$. Hence, $\frac{1}{n}\|\hat{\mathbf{P}} - \mathbf{P}\|_1 \le \frac{\bar{\epsilon}}{1-\rho} + o(1) = o(1)$ where $\bar{\epsilon} = o(1)$ since $\epsilon = o(1)$ and $1 - \rho$ is a non-zero constant. As a result, we obtain a novel iterative algorithm to compute the probability that an initially infected node i induces a global cascade on the network asymptotically.

Appendix C: Random clustered network construction

In the construction of a random clustered network, the probability that an edge $\{i, j\}$ with $1 \le i < j \le n$ is created more than once is at most the following (by the union bound method):

$$\begin{aligned}&\frac{s_i s_j}{\sum_x s_x}\sum_k \frac{t_i t_j t_k}{\sum_{x<y} t_x t_y} + \sum_{k_1<k_2} \frac{t_i t_j t_{k_1}}{\sum_{x<y} t_x t_y}\frac{t_i t_j t_{k_2}}{\sum_{x<y} t_x t_y} \\ &= \frac{s_i s_j}{\sum_x s_x}\frac{t_i t_j \sum_k t_k}{\sum_{x<y} t_x t_y} + \frac{(t_i t_j)^2}{\sum_{x<y} t_x t_y} \\ &\approx \frac{s_i s_j}{m_1}\frac{t_i t_j}{m_2/2} + \frac{(t_i t_j)^2}{m_2^2/2} \\ &= o((m_1 m_2)^{2/\tau - 1}) + o(m_2^{2(2/\tau-1)}), \end{aligned} \tag{C.1}$$

where m_1 is the number of single-edges, m_2 is the number of triangle-edges, and we assume that $\max_i\{s_i\} = o(m_1^{1/\tau})$ and $\max_i\{t_i\} = o(m_2^{1/\tau})$ for some $\tau > 2$. Equation (C.1) is the sum of the probability that both single-edge and triangle-edge $\{i, j\}$ are created and the probability that triangle-edge $\{i, j\}$ is created twice. Here we use an approximation that $(\sum_k t_k)^2 \approx 2\sum_{x<y} t_x t_y$, and the resulting upper bound goes to zero as the network size grows to infinity. Therefore, with high probability, the generating procedure does not make the same edge twice. It means that each operation that decides whether there is a triangle for each three nodes i, j, k or not does not affect almost all of the others. Hence, we assume that there is at most one edge joining two nodes.

Appendix D: When c follows a power-law distribution

We consider the scenario in which the infect/receive ability c is not fixed. Let c_i be the infect/receive ability for each node i. For example, we can assume that c_i follows a power-law distribution so that the distribution of the infect/receive ability has a heavy-tail. Then, the occurrence probabilities and the expected sizes of a global cascade for model (i) and (ii) are as follows:

$$\text{(i)} \begin{cases} P \approx \quad 1 - e^{-\sum_j \left(\frac{c_j w_j}{Vol(G)}\right) P}, \\ S \approx 1 - \frac{1}{n}\sum_j e^{-\left(\frac{c_j w_j}{Vol(G)}\right) nS}. \end{cases} \tag{D.1}$$

$$\text{(ii)} \begin{cases} P \approx 1 - \frac{1}{n}\sum_i e^{-\left(\frac{c_i w_i}{Vol(G)}\right) nP}, \\ S \approx \quad 1 - e^{-\sum_i \left(\frac{c_i w_i}{Vol(G)}\right) S}. \end{cases} \tag{D.2}$$

Consider two independent continuous random variables X and Y. Suppose that X and Y follow the power-law distributions with the exponents α and β respectively (i.e., $f_X(x) \propto x^{-\alpha}$ and $f_Y(y) \propto y^{-\beta}$). The probability density function of XY is $f_{XY}(k) = \int_x f_X(x) f_Y(k/x)|x|^{-1}dx = \int_x x^{-\alpha}(k/x)^{-\beta}|x|^{-1}dx = k^{-\beta}\int_x x^{-\alpha+\beta-1}dx \propto k^{-\beta}$ with $\alpha > \beta$. Therefore, if we assume that the infect/receive ability c_i and the expected degree w_i follow the power-laws with the exponents α and β respectively and $\alpha > \beta$, then $c_i w_i$ follows the power-law distribution and the exponent of $c_i w_i$ is identical to the exponent of w_i in either equations (D.1) or (D.2). It means that if c_i satisfies the above condition, then we can use the fixed value of $c = \sum_i(\frac{c_i w_i}{Vol(G)})$, the weighted average of c_is, rather than the values of c_i to compute P and S approximately when the size of the network is sufficiently large. Indeed, we conclude that the condition needed to induce a phase transition is $\sum_i(\frac{c_i w_i}{Vol(G)}) > 1$.

For model (iii), suppose that $f(i, j) = c_{1,i}/w_i + c_{2,j}/w_j$ where $c_{1,i}$ and $c_{2,j}$ follow the power-law distributions with the exponents β_1 and β_2 respectively and that $\alpha > \max\{\beta_1, \beta_2\}$. Then, we conclude by the same argument that the phase transition point is $c = c_1 + c_2$ where $c_1 = \sum_i(\frac{c_{1,i} w_i}{Vol(G)})$ and $c_2 = \sum_j(\frac{c_{2,j} w_j}{Vol(G)})$. If we assume that $c_{1,i} = c_{2,i}$ for all $i = 1, \ldots, n$, then we obtain that the the condition needed to induce a phase transition is $\sum_i(\frac{c_i w_i}{Vol(G)}) > 1/2$.

Appendix E: Degree correlations of the random clustered networks

In order to clarify the effect of clustering on degree correlation for Chung-Lu model, we would analyse numerically the distribution of k. For the Chung-Lu model ($\gamma = 0$),

$$k_{nn}(k) \approx \left(\sum_j w_j \frac{w_i w_j}{\sum_k w_k} \right) \Big/ \left(\sum_j \frac{w_i w_j}{\sum_k w_k} \right) = \frac{\sum_j w_j^2}{\sum_k w_k}. \tag{E.1}$$

If the expected degree sequence $\{w_i\}$ is given, therefore, the values of $k_{nn}(k)$ are the same approximately. In addition, for the random clustered network with $\gamma > 0$,

$$\begin{aligned} k_{nn}(k) &\approx \frac{\sum_j w_j \frac{s_i s_j}{\sum_k s_k} + \sum_{j<k}(w_j + w_k)\frac{t_i t_j t_k}{\sum_{j<k} t_j t_k}}{\sum_j \frac{s_i s_j}{\sum_k s_k} + 2\sum_{j<k} \frac{t_i t_j t_k}{\sum_{j<k} t_j t_k}} \\ &= \frac{s_i \sum_j w_j \frac{s_j}{\sum_k s_k} + t_i \sum_{j<k}(w_j + w_k)\frac{t_j t_k}{\sum_{j<k} t_j t_k}}{w_i} \\ &= (1-2\gamma)\frac{\sum_j w_j^2}{\sum_k w_k} + \gamma \frac{\sum_{j<k}(w_j + w_k) w_j w_k}{\sum_{j<k} w_j w_k}. \end{aligned} \tag{E.2}$$

The right term of last RHS of equation (E.2),

$$\begin{aligned} \frac{\sum_{j<k}(w_j + w_k) w_j w_k}{\sum_{j<k} w_j w_k} &\approx \sum_{j,k} \frac{(w_j + w_k) w_j w_k}{\sum_{j,k} w_j w_k} \\ &= 2\frac{\sum_{j,k} w_j^2 w_k}{\sum_{j,k} w_j w_k} = 2\frac{\sum_{j,k} w_j^2}{\sum_{j,k} w_j}. \end{aligned}$$

That is,

$$k_{nn}(k) \approx (1-2\gamma)\frac{\sum_j w_j^2}{\sum_k w_k} + 2\gamma \frac{\sum_j w_j^2}{\sum_k w_k} = \frac{\sum_j w_j^2}{\sum_k w_k}, \tag{E.3}$$

which is the same as the Chung-Lu model. Hence, $k_{nn}(k)$ has the same expected value for different k and γ.

References

1. P.S. Dodds, D.J. Watts, Phys. Rev. Lett. **92**, 218701 (2004)
2. S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, Rev. Mod. Phys. **80**, 1275 (2008)
3. J. Kleinberg, Commun. ACM **51**, 66 (2008)
4. D. Easley, J. Kleinberg, *Netwoks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge University Press, 2010)
5. D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information Diffusion Through Blogspace, in *Proc. of the 13th WWW, 2004*, pp. 491–501
6. D. Liben-Nowell, J. Kleinberg, Proc. Natl. Acad. Sci. USA **105**, 4633 (2008)
7. J.J. Cheng, Y. Liu, B. Shen, W.G. Yuan, Eur. Phys. J. B **86**, 29 (2013)
8. R. Pastor-Satorras, A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001)
9. R.M. Anderson, R.M. May, *Infectious Diseases of Humans* (Oxford University Press, 1991)
10. W.O. Kermack, A.G. McKendrick, Proc. R. Soc. Lond. A **115**, 700 (1927)
11. J.C. Miller, A.C. Slim, E.M. Volz, J. R. Soc. Interface **9**, 890 (2011)
12. E. Kenah, J.M. Robins, J. Theor. Biol. **249**, 706 (2007)
13. E. Kenah, J.M. Robins, Phys. Rev. E **76**, 036113 (2007)
14. J.C. Miller, Phys. Rev. E **76**, 010101(R) (2007)
15. M.E.J. Newman, Phys. Rev. E **66**, 016128 (2002)
16. P. Rattana, J. Miller, I. Kiss, Math. Modell. Nat. Phenom. **9**, 58 (2014)
17. M.E.J. Newman, SIAM Rev. **45**, 167 (2003)
18. E. Volz, Eur. Phys. J. B **63**, 381 (2008)
19. P. Erdős, A. Rényi, Publ. Math. **6**, 290 (1959)
20. R.M. D'Souza, Nat. Phys. **5**, 627 (2009)
21. L.A. Meyers, B. Pourbohloul, M.E.J. Newman, D.M. Skowronski, R.C. Brunham, J. Theor. Biol. **232**, 71 (2005)
22. A.L. Barabasi, Nature **435**, 207 (2005)
23. J.L. Iribarren, E. Moro, Phys. Rev. Lett. **103**, 038702 (2009)
24. M.E.J. Newman, Phys. Rev. Lett. **103**, 058701 (2009)
25. J.C. Miller, Phys. Rev. E **80**, 020901(R) (2009)
26. F. Chung, L. Lu, Ann. Combin. **6**, 125 (2002)
27. F. Chung, L. Lu, SIAM J. Discrete Math. **20**, 395 (2006)
28. B. Barry, I.S. Fulmer, Acad. Manag. Rev. **29**, 272 (2004)
29. L. Weng, A. Flammini, A. Vespignani, F. Menczer, Sci. Rep. **2**, 335 (2012)
30. R. Dunbar, How Many "Friends" Can You Really Have?, in *IEEE Spectrum, 2011*, Vol. 48, pp. 81–83
31. L.A. Adamic, R.M. Lukose, A.R. Puniyani, B.A. Huberman, Phys. Rev. E **64**, 046135 (2001)
32. M.W. Hirsch, S. Smale, R.L. Devaney, in *Differential Equations, Dynamical Systems, and an Introduction to Chaos* (Academic press, 2004), Vol. 60
33. B. Bollobás, *Random Graphs* (Oxford University Press, 2001)
34. M. Molloy, B. Reed, Random Struct. Alg. **6**, 161 (1995)
35. M. Molloy, B. Reed, Combinatorics, Probab. Comput. **7**, 295 (1998)
36. M.E.J. Newman, S.H. Strogatz, D.J. Watts, Phys. Rev. E **64**, 026118 (2001)
37. E. Kenah, J.C. Miller, Interdisciplinary Perspectives on Infectious Diseases **2011**, 543520 (2011)
38. D. Wang, Z. Wen, H. Tong, C.Y. Lin, C. Song, A.L. Barabasi, Information Spreading in Context, in *Proc. of the 20th WWW, 2011*, pp. 735–744
39. D. Kempe, J. Kleinberg, E. Tardos, Maximizing the Spread of Influence through a Social Network, in *Proc. of the 9th ACM SIGKDD, 2003*, pp. 137–146
40. D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, C. Faloutsos, ACM Trans. Inform. Syst. Security **10**, 1 (2008)
41. A. Ganesh, L. Massoulie, D. Towsley, The effect of network topology on the spread of epidemics, in *Proc. of 24th IEEE INFOCOM, 2005*, pp. 1455–1466

42. B.A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, C. Faloutsos, Threshold Conditions for Arbitrary Cascade Models on Arbitrary Networks, in *Proc. of 11th IEEE ICDM, 2011*, pp. 537–546
43. B. Karrer, M. Newman, L. Zdeborová, Phys. Rev. Lett. **113**, 208702 (2014)
44. B. Viswanath, A. Mislove, M. Cha, K.P. Gummadi, On the Evolution of User Interaction in Facebook, in *Proc. of the 2nd ACM WOSN, 2009*, pp. 37–42
45. Y.Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, Analysis of topological characteristics of huge online social networking services, in *Proc. of the 16th WWW, 2007*, pp. 835–844
46. A.L. Barabasi, Philos. Trans. Roy. Soc. London A **371**, 20120375 (2013)
47. J.P. Gleeson, S. Melnik, A. Hackett, Phys. Rev. E **81**, 066114 (2010)
48. E. Coupechoux, M. Lelarge, Adv. Appl. Probab. **46**, 985 (2014)