

HOW CLUSTERING AFFECTS EPIDEMICS IN RANDOM NETWORKS

EMILIE COUPECHOUX * AND
MARC LELARGE, ** INRIA-ENS

Abstract

Motivated by the analysis of social networks, we study a model of random networks that has both a given degree distribution and a tunable clustering coefficient. We consider two types of growth process on these graphs that model the spread of new ideas, technologies, viruses, or worms: the diffusion model and the symmetric threshold model. For both models, we characterize conditions under which global cascades are possible and compute their size explicitly, as a function of the degree distribution and the clustering coefficient. Our results are applied to regular or power-law graphs with exponential cutoff and shed new light on the impact of clustering.

Keywords: Contagion threshold; diffusion; random graph; clustering

2010 Mathematics Subject Classification: Primary 60C05

Secondary 05C80; 91D30

1. Introduction

In this paper we analyze two different types of epidemic modeled by simple growth processes that we now describe. In a given network, each node can either be active or inactive. The *diffusion* process corresponds to the case where each node of the network that becomes active transmits the activation to her neighbors with a given probability, independently from each other. On the other hand, in the *symmetric threshold* model, a threshold is associated to each node, and the dynamics of the process correspond to the case where a node of the network becomes active as soon as the number of her active neighbors exceeds the threshold of the node. Thresholds are (possibly) random, with a distribution depending on the degree of the node, and such that thresholds are independent among nodes. The symmetric threshold model will allow us to analyze the contagion process [24]. For both models, we will first consider a case where there is only one initial active node and characterize conditions under which global cascades are possible (see a precise definition below). In such cases we compute the probability of a cascade and its size. Then, we consider the cascade size when a positive fraction of the population is initially active. The initial activations are random in that case, and the probability that a node belongs to the seed might depend on its degree.

We now describe the model of random graphs studied in this paper. For many real-world networks, the underlying graph G is a power-law graph, i.e. a graph whose degree distribution follows a power law. Random graphs with a given degree sequence allow such behavior to be modeled. This model is usually called the configuration model [7]. There is a vast literature on the analysis of the diffusion for such graphs [26]. The contagion process has also been studied

Received 13 March 2012; revision received 7 August 2013.

* Current address: Laboratoire I3S, CS 40121, Université Nice Sophia Antipolis, 06903 Sophia Antipolis Cedex, France. Email address: coupecho@i3s.unice.fr

This paper is part of the author's PhD thesis done at INRIA-ENS.

** Postal address: INRIA-ENS, 23 avenue d'Italie, CS 81321, 75214 Paris Cedex 13, France.

for such graphs through heuristics [30], or rigorously in [21] or [2]. Random graphs are not considered to be highly realistic models of most real-world networks, and they are used as a first approximation since they are a natural choice for sparse interaction networks in the absence of any known geometry. One essential drawback of this model is that these graphs are ‘locally tree-like’: short cycles are very rare. However, real-world networks are often highly clustered, meaning that there is a large number of triangles and other short cycles [25]. For social graphs, this is a consequence of the fact that friendship circles are typically strongly overlapping so that many of our friends are also friends of each other.

There are works in the physics and biology literature on models of random graphs with clustering [27]. Our model is inspired from [29] in which random graphs with positive clustering and possibly power-law degree distribution are modeled. The idea is to ‘add’ clustering to a standard configuration model by replacing some vertices by cliques. By choosing the fraction of vertices replaced, this leads to a graph where the amount of clustering can be tuned by adjusting the parameters of the model. This model generalizes the standard configuration model to incorporate clustering. Understanding how clustering affects diffusion and contagion remains largely an open question. Our work is a first step towards addressing this issue in a systematic and rigorous way. In particular, we are able to make a rigorous analysis of the impact of the clustering coefficient while the degree distribution in the graph is kept fixed. In [29] and [15], the diffusion process on such graphs is analyzed by a heuristic approximation through a branching process with additional cliques. We derive rigorous proofs for these results. A different random graph model with clustering, called the random intersection graph, has been studied rigorously in [13] and, in [9], the diffusion process is studied on such graphs. However, the degree distribution for this kind of graph has to be a Poisson distribution and clustering cannot vary independently of the degree distribution. A generalization of the diffusion process (with general infectious periods) in random graphs with clustering has been studied rigorously by Ball *et al.* [3], [4], and heuristically in [5] (see the end of Section 2.1 for a comparison with our model). The impact of clustering on the diffusion process is studied in [5], but only for Poisson degree distributions (using a property of this distribution). To the best of our knowledge, results on the contagion model have not been proved previously for random graphs with clustering. (A preliminary version without proofs of our work appeared in [10] and we also recently studied a simple model with overlapping communities in [12].) Recently, Acemoglu *et al.* [1] derived bounds which are valid for the contagion model [24] on deterministic networks. Our analysis in contrast gives asymptotic results as the size of the graph tends to ∞ and allows us (by looking at a more specific model) to neatly identify the impact of clustering.

The paper is organized as follows. In Section 2 we present the graph model, and compute its asymptotic degree distribution and clustering coefficient. In Section 3 we study the diffusion process on such graphs with a seed of only one vertex and with degree-based activation. Similarly, in Section 4 we deal with the symmetric threshold model. Proofs are given in Section 5 and conclusions in Section 6.

Notation. In the following, we consider asymptotics as $n \rightarrow \infty$, and we denote by ‘ $\xrightarrow{\mathbb{P}}$ ’ the convergence in probability as $n \rightarrow \infty$. The abbreviation ‘w.h.p.’ (‘with high probability’) means with probability tending to 1 as $n \rightarrow \infty$, and we use the notation $o_{\mathbb{P}}(n)$ and $\Theta_{\mathbb{P}}(n)$ in a standard way (see [18] for instance): $X = o_{\mathbb{P}}(n)$ means that, for every $\varepsilon > 0$, $\mathbb{P}(X > \varepsilon n) \rightarrow 0$ as $n \rightarrow \infty$, and $X = \Theta_{\mathbb{P}}(n)$ means that, for every $\delta > 0$, there exist constants $c > 0$ and $C > 0$ such that $\mathbb{P}(cn \leq X \leq Cn) > 1 - \delta$ for large enough n . We also set $[n] = \{1, 2, \dots, n\}$. In addition, for integers $s \geq 0$ and $0 \leq r \leq s$, let b_{sr} denote the binomial probabilities $b_{sr}(p) := \mathbb{P}(\text{Bin}(s, p) = r) = \binom{s}{r} p^r (1-p)^{s-r}$.

2. Random graph model and its basic properties

We first present the model for the random graph, compute its asymptotic degree distribution and its asymptotic clustering coefficient, and finally explain how to tune the clustering coefficient while keeping the asymptotic degree distribution fixed.

2.1. Model of random graph with clustering

We first consider the uniform random graph with fixed degree distribution: since this graph has asymptotically no clustering, we will then modify it to obtain a graph with clustering. Finally, we compare our clustered random graph model to those that are close in the literature [4], [5], [29].

Let $n \in \mathbb{N}$, and let $\mathbf{d} = (d_i^{(n)})_{i=1}^n = (d_i)_1^n$ be a sequence of non-negative integers such that $\sum_i d_i$ is even. The integer n is the number of vertices in the graph and vertex $i \in [n]$ has degree d_i in the graph. Let $G(n, \mathbf{d})$ be a graph chosen uniformly at random among all simple (i.e. with no multiedges or self-loops) graphs with n vertices and degree sequence \mathbf{d} (assuming such graphs exist) [7].

We will let $n \rightarrow \infty$, and assume that we are given \mathbf{d} satisfying the following regularity conditions which are standard in the random graph literature; see [23].

Condition 1. For each n , $\mathbf{d} = (d_i)_1^n$ is a sequence of nonnegative integers such that $\sum_i d_i$ is even. We assume that there exists a probability distribution $\mathbf{p} = (p_r)_{r=0}^\infty$ (independent of n) such that

- (i) $n_r/n = |\{i : d_i = r\}|/n \rightarrow p_r$ as $n \rightarrow \infty$ for all $r \geq 0$;
- (ii) $\lambda := \sum_r r p_r \in (0, \infty)$;
- (iii) $\sum_i d_i^3 = O(n)$.

The random graph model $G(n, \mathbf{d})$ is ‘locally tree-like’, i.e. it contains very few (i.e. $o(n)$) short cycles in its structure. We now use the idea of Trapman [29] to generalize this random graph model to incorporate clustering. We replace some vertices by a clique of size the degree r in the original graph, i.e. a vertex of degree r in the original graph $G(n, \mathbf{d})$ is replaced by r new vertices with $r(r - 1)/2$ new edges connecting them, each new vertex being connected to exactly one of the neighbors of the original vertex, as illustrated on Figure 1. Note that if $r = 0$, i.e. if the original node is isolated, this procedure removes the node. By convention, a clique of size 0 is empty.

In order to be able to tune the clustering coefficient in the graph, we will not replace all vertices by a clique but do a probabilistic choice whether to replace a vertex or not: for all $r \geq 0$, $\gamma_r \in [0, 1]$ represents the probability that a vertex of degree r in $G(n, \mathbf{d})$ is replaced by a clique of size r in the new model denoted by $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = (\gamma_r)_{r=0}^\infty$ is shorthand

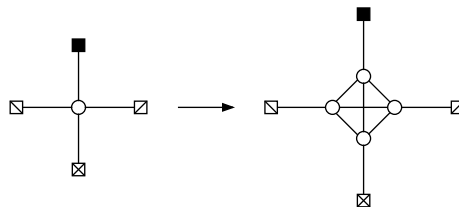


FIGURE 1: Transformation.

notation for the sequence of γ_r s. The choice of whether or not to replace a vertex by a clique is made independently at each vertex of the original graph $G(n, \mathbf{d})$. In particular, if $\gamma_r = 0$ for all $r \geq 0$ then we simply obtain $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma}) = G(n, \mathbf{d})$, whereas if $\gamma_r = 1$ for all $r \geq 0$, all vertices in $G(n, \mathbf{d})$ are replaced by cliques (and isolated vertices are removed). We write $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ for the graph $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ in which the sequence $\boldsymbol{\gamma}$ is constant and equal to γ .

In the rest of the paper, when no random graph model is specified, we tacitly assume that we consider the model $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ for a sequence \mathbf{d} satisfying Condition 1 with probability distribution $\mathbf{p} = (p_r)_{r=0}^\infty$, and clustering parameter $\boldsymbol{\gamma} = (\gamma_r)_{r=0}^\infty$.

There are several differences with the model of Trapman [29]. The main difference is that we allow the probability γ_r to depend on the vertex degree r , while $\gamma_r = \gamma$ is constant in [29]. In our model, the degree d_i of each vertex i , $1 \leq i \leq n$, is also fixed (i being a vertex of the unclustered model), and the proportion of vertices with degree r tends asymptotically to p_r , while the degree of each vertex in Trapman’s model is a random variable with distribution $(p_r)_r$, where the random variables are independent and identically distributed (i.i.d.) among vertices. Finally, a minor difference is that self-loops and multiedges are allowed in [29]: this difference is minor since we will also consider multigraphs in the proofs, and, as explained at the beginning of Section 5, theorems are valid for both simple graphs and multigraphs.

In the model of Ball *et al.* [5], the sizes of ‘cliques’ are i.i.d. (called households, and possibly of size 1). Each node v inside a household is assigned a number d_v^{ext} of ‘external’ half-edges, with all d_v^{ext} being i.i.d. among vertices. A matching of these half-edges is then chosen uniformly at random, as in the configuration model. In our model, vertices belonging to a clique have always only one ‘external’ edge, whereas, for vertices not belonging to a clique, the number of external edges equals the degree. In the particular cases where $\gamma_r = 1$ for all r or $\gamma_r = 0$ for all r , our model is a special case of that given in the work of Ball *et al.* [4], [5].

2.2. Degree distribution in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$

As we will see in the next subsection, the procedure described above introduces clustering as soon as $\gamma_r > 0$ for some r . It also modifies the degree distribution in the graph, and we can easily derive the new degree distribution using Condition 1 (a detailed proof can be found in [11]).

Proposition 1. *For all $r \geq 0$, let \tilde{n}_r be the number of vertices with degree r in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, and let $\tilde{n} = \sum_r \tilde{n}_r$ be the total number of vertices in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$. Then we have, as $n \rightarrow \infty$, $\tilde{n}/n \xrightarrow{\mathbb{P}} \tilde{\gamma} := \sum_{d \geq 0} [d\gamma_d + (1 - \gamma_d)]p_d > 0$ and, for all $r \geq 0$, the proportion of vertices with degree r in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ has the following limit as $n \rightarrow \infty$: $\tilde{n}_r/\tilde{n} \xrightarrow{\mathbb{P}} \tilde{p}_r := [r\gamma_r + (1 - \gamma_r)]p_r/\tilde{\gamma}$.*

2.3. Clustering coefficient

We now compute the clustering coefficient of the graph $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$. The most common definition of the clustering coefficient of a finite graph is given by

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}} \in [0, 1]. \tag{1}$$

In our model of random graphs where vertices are exchangeable, this definition can also be interpreted as the conditional probability that there is an edge between two vertices j and k , given that they have a common neighbor i .

Note that the number of connected triples in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ is simply $\sum_v d_v(d_v - 1)/2$. On the other hand, for any vertex v in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, let P_v be the number of pairs of neighbors of v that share an edge together. More precisely, if \mathcal{N}_v is the set of neighbors of v (whose cardinality is

$|\mathcal{N}_v| = d_v$) then P_v is the number of pairs $\{w, w'\} \subset \mathcal{N}_v$, $w \neq w'$, such that w and w' are also neighbors of each other. Thus, three times the number of triangles in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ is given by $\sum_v P_v$. Hence, following (1), we define the clustering coefficient $C^{(n)}$ of the graph $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ by

$$C^{(n)} = \frac{2 \sum_v P_v}{\sum_v d_v(d_v - 1)} \in [0, 1].$$

Using Condition 1 and the fact that $G(n, \mathbf{d})$ has asymptotically no clustering, we can easily derive the following asymptotics for $C^{(n)}$ (see [11] for a detailed proof).

Proposition 2. *The clustering coefficient of $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ is given by*

$$C^{(n)} \xrightarrow{\mathbb{P}} C := \frac{\sum_{r \geq 2} r(r-1)(r-2)\gamma_r p_r}{\sum_{r \geq 2} ((r-1)\gamma_r + 1)r(r-1)p_r}.$$

2.4. Tunable clustering coefficient with fixed degree distribution

Following the idea of Trapman [29], we show how to use our model to generate graphs with a given degree distribution and clustering. This construction will allow us to compare graphs with a given degree distribution but with various clustering coefficients and to see the impact of clustering on the epidemic. In order to provide a graph with a given asymptotic degree distribution $\tilde{\mathbf{p}}$ and a positive clustering coefficient, we need the following assumptions on $\tilde{\mathbf{p}}$.

Condition 2. *We assume that the probability distribution $\tilde{\mathbf{p}}$ satisfies*

- (i) $\sum_r r^2 \tilde{p}_r < \infty$;
- (ii) $\sum_{r \geq 3} \tilde{p}_r > 0$;
- (iii) $\tilde{p}_0 = 0$.

Under these conditions, we have the following proposition (see [29] or [11] for a proof).

Proposition 3. *Let $\tilde{\mathbf{p}} = (\tilde{p}_r)_{r \geq 0}$ be a probability distribution satisfying Condition 2. For any value C such that $0 \leq C \leq C^{\max} := 1 - 2 \sum_{r \geq 2} (r-1)\tilde{p}_r / \sum_{r \geq 2} r(r-1)\tilde{p}_r$, there exists a sequence \mathbf{d} (satisfying Condition 1 with probability distribution $\mathbf{p} = (p_r)_{r=0}^\infty$) and a value of $\gamma \in [0, 1]$ such that the model $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ has asymptotic degree distribution $\tilde{\mathbf{p}}$ and asymptotic clustering coefficient C .*

More precisely, γ is the solution of the equation

$$C = C(\gamma) := \frac{\sum_{r \geq 3} r(r-1)(r-2)\gamma \tilde{p}_r / ((r-1)\gamma + 1)}{\sum_{r \geq 2} r(r-1)\tilde{p}_r},$$

and \mathbf{p} can be defined as follows. Let

$$F(\gamma') := \sum_{r \geq 1} \frac{r}{(r-1)\gamma' + 1} \tilde{p}_r \quad \text{for all } \gamma' \in [0, 1].$$

We set $\lambda := F(\gamma)(1-\gamma)/(1-\gamma F(\gamma))$ if $\gamma \neq 1$, and $\lambda := 1/(\sum_{r \geq 1} \tilde{p}_r/r)$ if $\gamma = 1$. Then we have $p_r := \tilde{p}_r[(\lambda-1)\gamma + 1]/((r-1)\gamma + 1)$ for all $r \geq 1$, and $p_0 := 0$.

Note that, for a fixed $\tilde{\mathbf{p}}$, the function $C(\gamma)$ is increasing in γ [11].

3. Diffusion threshold for random graphs with clustering

3.1. Diffusion model

In this section we study a simple diffusion model depending on a single parameter $\pi \in [0, 1]$. For a given graph G , the dynamics of the diffusion are as follows: some set of nodes S starts out being active; all other nodes are inactive. When a node becomes active, each of her neighbors becomes active with probability π , independently from each other. The final state of the diffusion can also be described in term of a bond percolation process in the graph G . Randomly delete each edge with probability $1 - \pi$, independently of all other edges, and denote by G_π the resulting bond percolated graph. Then any node in S will activate all nodes in its connected component in G_π .

3.2. Phase transition for the diffusion with a single activation

In this subsection we consider diffusion starting from one active node with all other nodes being inactive, and we derive conditions under which a single starting active node can activate a large fraction of the population in $G = \tilde{G}(n, d, \gamma)$. This problem corresponds to the existence of a ‘giant component’ in the random graph obtained after bond percolation.

In order to state our result, we first recall some basic results about random graphs with small order. Let K_d , $d \in \mathbb{N}$, be the complete graph on d vertices, and let $K_d(\pi)$, $\pi \in [0, 1]$, be the random graph obtained from K_d after bond percolation with parameter π . The probability that the component in $K_d(\pi)$ containing vertex 1 has k vertices is denoted by $f(d, k, \pi)$, and can be computed as follows (note that $f(d, d, \pi)$ is simply the probability that $K_d(\pi)$ is connected and has been computed in [14]):

$$\begin{aligned}
 f(d, d, \pi) &= 1 - \sum_{k=1}^{d-1} \binom{d-1}{k-1} f(k, k, \pi) (1 - \pi)^{k(d-k)}, \\
 f(d, k, \pi) &= \binom{d-1}{k-1} f(k, k, \pi) (1 - \pi)^{k(d-k)} \quad \text{for any } k \leq d.
 \end{aligned}
 \tag{2}$$

For $d \in \mathbb{N}$ and $\pi \in [0, 1]$, we define the random variable $\mathcal{K}(d, \pi, \gamma)$ by

$$\mathbb{P}(\mathcal{K}(d, \pi, \gamma) = k) = (1 - \gamma_d) \mathbf{1}(d = k) + \gamma_d f(d, k, \pi),$$

where f is defined in (2).

In addition, set, for all $k \geq 1$,

$$\varrho_k := p_k(1 - \gamma_k) + \sum_{d \geq k} \frac{d}{k} f(d, k, \pi) p_d \gamma_d,
 \tag{3}$$

$$\varrho := \sum_{\ell} \varrho_{\ell},
 \tag{4}$$

$$\mu := \sum_{\ell} \frac{\ell \varrho_{\ell}}{\varrho},
 \tag{5}$$

$$\sigma_k := p_k(1 - \gamma_k) + \sum_{d \geq k} d f(d, k, \pi) p_d \gamma_d.$$

Using the notation $\tilde{\gamma}$ of Proposition 1, we set (omitting the dependence on \mathbf{p} and \mathbf{y})

$$\begin{aligned}
 L(z) &:= \sum_s \frac{\sigma_s}{\tilde{\gamma}} [1 - (1 - \pi + \pi z)^s], \\
 h(z) &:= \sum_s s \frac{\varrho_s}{\varrho} (1 - \pi + \pi z)^s, \\
 \zeta &:= \sup\{z \in [0, 1) : \mu z(1 - \pi + \pi z) = h(z)\}.
 \end{aligned}
 \tag{6}$$

For a graph $G = (V, E)$ and a parameter $\pi \in [0, 1]$, we denote by $C^b(\pi)$ the size of the largest component in the bond percolated graph G_π .

Theorem 1. *Let D^* be a random variable with distribution p_r^* given by $p_{r-1}^* = r p_r / \lambda$ for all $r \geq 1$. We define π_c as the solution of the equation*

$$\pi \mathbb{E}[\mathcal{K}(D^* + 1, \pi, \mathbf{y}) - 1] = 1.$$

- (i) *If $\pi > \pi_c$, we have $\zeta \in (0, 1)$. In addition, the asymptotic size of the largest component of the percolated graph G_π obtained from $\tilde{G}(n, \mathbf{d}, \mathbf{y})$ is*

$$\frac{C^b(\pi)}{\tilde{n}} \xrightarrow{\mathbb{P}} L(\zeta) > 0.$$

- (ii) *If $\pi < \pi_c$, we have $C^b(\pi) = o_{\mathbb{P}}(\tilde{n})$.*

Note that in the particular case where $\gamma_r = 0$ for all r , we have $\mathcal{K}(d, \pi, 0) = d$, so we obtain $\pi_c = \mathbb{E}[D] / \mathbb{E}[D(D - 1)]$, where D is the typical degree in the random graph, and our result reduces to a standard result in the random graph literature (see Theorem 3.9 of [16]).

We have the following interpretation of the different quantities in terms of an approximating branching process. The random graph $\tilde{G}(n, \mathbf{d}, \mathbf{y})$ can be approximated by a rooted random graph G^r distributed as follows (to simplify explanations, when the root of G^r belongs to a clique, we call the ‘root’ the whole clique). The root of G^r is a single node of degree d (respectively a clique of size d) with probability proportional to $(1 - \gamma_d)p_d$ (respectively $d\gamma_d p_d$). Each subsequent ‘individual’ of the root is a single node of degree d (respectively a clique of size d) with probability proportional to $d(1 - \gamma_d)p_d$ (respectively $d\gamma_d p_d$).

We prove that the mean offspring number in G_π^r is $\pi \mathbb{E}[\mathcal{K}(D^* + 1, \pi, \mathbf{y}) - 1]$ and that $L(\zeta)$ is the survival probability in G_π^r .

We first delete independently with probability $1 - \pi$ (in G^r) each ‘internal’ edge, i.e. each edge inside a clique, and call $G_{\pi, \text{int}}^r$ the resulting graph: the cliques in $G_{\pi, \text{int}}^r$ can be broken into several components, thus losing some of their children (see Figure 2). The probability that a vertex has degree k in $G_{\pi, \text{int}}^r$, knowing that its degree in G^r is d , is given by $(1 - \gamma_d)\mathbf{1}(d = k) + \gamma_d f(d, k, \pi) = \mathbb{P}(\mathcal{K}(d, \pi, \mathbf{y}) = k)$. Since the degree of a vertex in G^r is distributed as $D^* + 1$, the offspring number of a vertex in $G_{\pi, \text{int}}^r$ is distributed as $\mathcal{K}(D^* + 1, \pi, \mathbf{y}) - 1$ (by definition, the offspring number of a vertex inside a clique is the number of children of the clique). Equivalently, the probability that a vertex in $G_{\pi, \text{int}}^r$ has $k - 1$ children is proportional to $k \varrho_k$.

We then remove independently with probability $1 - \pi$ each ‘external’ edge, i.e. each edge outside a clique. The mean offspring number of the resulting graph G_π^r is thus $\pi \mathbb{E}[\mathcal{K}(D^* + 1, \pi, \mathbf{y}) - 1]$, and the probability that a vertex in G_π^r has $r - 1$ children is $\sum_{k \geq r} k \varrho_k b_{k-1, r-1}(\pi)$. A simple calculation shows that $z \mapsto h(z) / \mu(1 - \pi + \pi z)$ is the generating function of the

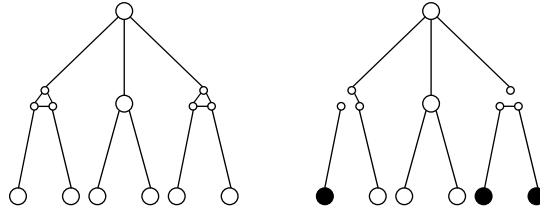


FIGURE 2: *Left*: the random graph G^r (before percolation). *Right*: the random graph $G_{\pi, \text{int}}^r$: after percolation inside cliques, some children are lost (filled circles).

offspring number in G_{π}^r . Hence, ζ is the extinction probability of G_{π}^r and, taking into account the offspring distribution of the root, we find that $L(\zeta)$ is the survival probability in G_{π}^r .

The case where a positive fraction of individuals belong to S (not only a single node) is discussed in Section 3.4. Now we study the effect of clustering on the diffusion with a single activation.

3.3. Effect of clustering on the diffusion

In this subsection we assume that $\gamma_r = \gamma$ for all $r \geq 0$.

Proposition 4. *Let \tilde{p} be a probability distribution satisfying Condition 2. For each $j = 1, 2$, let (p^j, d^j, γ_j) be chosen as described in Proposition 3 so that $\tilde{G}(n, d^1, \gamma_1)$ and $\tilde{G}(n, d^2, \gamma_2)$ both have asymptotic degree distribution \tilde{p} with clustering $C(\gamma_1) \leq C(\gamma_2)$. Let π_c^j be the diffusion threshold defined in Theorem 1 for the random graph $\tilde{G}^j = \tilde{G}(n, d^j, \gamma_j)$, $j = 1, 2$. Let S_{π}^j be the epidemic size in \tilde{G}^j , $j = 1, 2$ (i.e. $S_{\pi}^j = \lim C^b(\pi)/\bar{n}$ with the notation of Theorem 1). Then we have $\pi_c^1 \leq \pi_c^2$ and $S_{\pi}^1 \geq S_{\pi}^2$ for all $\pi \geq 0$.*

Proof. First note that $C(\gamma_1) \leq C(\gamma_2)$ implies that $\gamma_1 \leq \gamma_2$ [11].

As explained in the previous subsection, the random graph \tilde{G}^j , $j = 1, 2$, can be approximated by a branching process $G^{r,j}$; the diffusion threshold π_c^j is the solution of $M^j = 1$, where M^j is the mean offspring number in $G_{\pi}^{r,j}$, and the epidemic size is the nonextinction probability of $G_{\pi}^{r,j}$.

Let Z_0^j and Z_1^j respectively be the offspring numbers of the root and of any individual different from the root in $G_{\pi}^{r,j}$. Proving that $Z_0^1 \leq_{\text{st}} Z_0^2$ and $Z_1^1 \leq_{\text{st}} Z_1^2$ will complete the proof (where ‘ \leq_{st} ’ denotes the stochastic order).

By definition, the degree of the root is distributed in both $G^{r,1}$ and $G^{r,2}$ according to \tilde{p} . Knowing that its degree is d , it belongs to a clique with probability $p_{\text{clique}}^j(d) = d\gamma_j / (1 - \gamma_j + d\gamma_j)$, in which case Z_0^j is distributed as $\text{Bin}(K_{d,1}(\pi), \pi)$ (where $K_{d,1}(\pi)$ is the size of the connected component of vertex 1 in $K_d(\pi)$); otherwise, it is distributed as $\text{Bin}(d, \pi)$. Since $p_{\text{clique}}^1(d) \leq p_{\text{clique}}^2(d)$ and $\mathbb{P}(K_{d,1}(\pi) \leq d) = 1$, we have $Z_0^1 \leq_{\text{st}} Z_0^2$.

We set $\lambda_j := \sum_r r p_r^j$. The degree D^j of an individual different from the root in $G^{r,j}$ is given by: $\mathbb{P}(D^j = d) = d p_d^j / \lambda_j$. Using the expression of p_d^j in Proposition 3, we have

$$\frac{\mathbb{P}(D^1 = d)}{\mathbb{P}(D^2 = d)} = \frac{\lambda_2 (\lambda_1 - 1)\gamma_1 + 1 (d - 1)\gamma_2 + 1}{\lambda_1 (\lambda_2 - 1)\gamma_2 + 1 (d - 1)\gamma_1 + 1}$$

which increases with d . Thus, $D^1 \leq_{\text{lr}} D^2$ (where ‘ \leq_{lr} ’ denotes the likelihood ratio order), which implies that $D^1 \leq_{\text{st}} D^2$ (see Theorem 1.C.1 of [28]). By the same arguments as those used for the root, it follows that $Z_1^1 \leq_{\text{st}} Z_1^2$, which completes the proof.

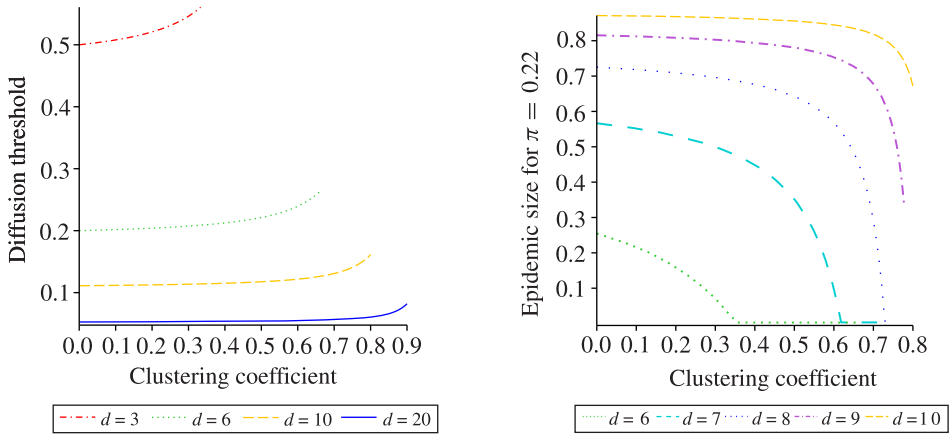


FIGURE 3: *Left*: evolution of the diffusion threshold with respect to the clustering coefficient for d -regular graphs. *Right*: evolution of the epidemic size with respect to the clustering coefficient for d -regular graphs (with infection probability $\pi = 0.22$).

In the left-hand diagram of Figure 3 we illustrate the proposition above for d -regular graphs ($\tilde{p}_r = p_r = \mathbf{1}(r = d)$ for all $r \geq 0$) and show how the diffusion threshold increases with the clustering coefficient for different values of d . In addition, the epidemic size also decreases with the clustering. In the right-hand diagram of Figure 3 we plot the ratio of the largest connected component in the percolated graph over the whole population. When the starting infected individual is a vertex chosen uniformly at random, this ratio also corresponds to the probability of explosion. Hence, as the clustering increases, it ‘inhibits’ the diffusion process. These results are in accordance with those given in [15].

These results are intuitive (for d -regular graphs) in the sense that the removal of edges inside cliques can stop the diffusion *inside* a clique in the graph $\tilde{G}(n, d, \gamma)$, while this phenomenon does not occur in the original graph $G(n, d)$.

We refer the reader to [11] for an analysis of the diffusion on graphs with power-law degree distribution.

3.4. Phase transition for the diffusion with degree-based activation

In this subsection we allow a positive fraction of nodes to be active at the beginning of the diffusion process. More precisely, on a given graph G , the set S of initial active nodes is random, and each node of degree d in G belongs to S with some probability $\alpha_d > 0$, independently for each node. We set $\alpha = (\alpha_d)_{d \geq 0}$.

Using the notation $\tilde{\gamma}$ defined in Proposition 1, and definitions (3)–(5), we define (omitting the dependence on α , p , and γ)

$$L(z) := \sum_s \frac{(1 - \gamma_s) p_s}{\tilde{\gamma}} [1 - (1 - \alpha_s)(1 - \pi + \pi \zeta)^s] + \sum_{d \geq s} df(d, s, \pi) \frac{\gamma_d p_d}{\tilde{\gamma}} [1 - (1 - \alpha_d)^s (1 - \pi + \pi \zeta)^s],$$

$$\begin{aligned}
 h(z) &:= \sum_s \frac{s(1 - \gamma_s)P_s}{\varrho} [1 - (1 - \alpha_s)(1 - \pi + \pi\zeta)^s] \\
 &\quad + \sum_{d \geq s} \frac{df(d, s, \pi)\gamma_d P_d}{\varrho} [1 - (1 - \alpha_d)^s(1 - \pi + \pi\zeta)^s], \\
 \zeta &:= \sup\{z \in [0, 1) : \mu z(1 - \pi + \pi z) = h(z)\}.
 \end{aligned}
 \tag{7}$$

Theorem 2. *We are given an activation set S drawn according to the distribution α on the random graph $G = \tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$. For the diffusion model defined in Section 3.1, if $\zeta = 0$, or if $\zeta \in (0, 1]$, and, furthermore, ζ is such that there exists $\varepsilon > 0$ with $\mu z(1 - \pi + \pi z) < h(z)$ for $z \in (\zeta - \varepsilon, \zeta)$, the size $C^b(\pi, \alpha)$ of the active nodes at the end of the diffusion satisfies $C^b(\pi, \alpha)/\tilde{n} \xrightarrow{\mathbb{P}} L(\zeta)$.*

4. Symmetric threshold model for random graphs with clustering

4.1. Symmetric threshold model

We now describe the symmetric threshold model on a finite graph $G = (V, E)$ with given thresholds $k(v)$ for $v \in V$. The progressive dynamics of the epidemic on the finite graph G operates as follows: some set of nodes S starts out being active; all other nodes are inactive. Time operates in discrete steps $t = 1, 2, 3, \dots$. At a given time t , any inactive node v becomes active if its number of active neighbors is at least $k(v) + 1$. This in turn may cause other nodes to become active. It is easy to see that the final set of active nodes depends only on the initial set S (and not on the order of the activations) and can be obtained as follows. Set $Y_v = \mathbf{1}(v \in S)$ for all v . Then, as long as there exists v such that $\sum_{w \sim v} Y_w > k(v)$, set $Y_v = 1$, where $w \sim v$ means that v and w share an edge in G . When this algorithm finishes, the final state of node v is represented by Y_v : $Y_v = 1$ if node v is active and $Y_v = 0$ otherwise. In this paper we do not analyze the dynamics of the epidemics, we concentrate on the final state only.

We consider the symmetric threshold model on $G(n, \mathbf{d})$ as in [21], and define an adaptation for the random graph $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$. More precisely, let $(t_{s\ell})_{0 \leq \ell \leq s}$ be a probability distribution for all $s \geq 0$, and set $\mathbf{t} = (t_{s\ell})_{s,\ell}$. We allow the threshold $k(i)$ of a node i in $G(n, \mathbf{d})$ to be a random variable with distribution depending on the degree of i : knowing that the degree d_i of node i is s , threshold $k(i)$ is drawn according to the conditional probability distribution $(t_{s\ell})_{0 \leq \ell \leq s}$, i.e. $\mathbb{P}(k(i) = \ell \mid d_i = s) = t_{s\ell}$. We assume that thresholds are independent among the nodes of $G(n, \mathbf{d})$. Of particular interest to us will be the particular case in which $k(v) = qd_v$, where $q \in [0, \frac{1}{2}]$ and d_v is the degree of v . We refer the reader to [24] and [21] for a description of a game-theoretic framework leading to this model.

Throughout this section, we consider the following epidemic model on $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$: when a vertex i of $G(n, \mathbf{d})$ is replaced by a clique in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, we associate to each vertex inside the clique the original threshold $k(i)$, so that vertices inside a clique have the same threshold (also referred to as the ‘threshold of the clique’). We still denote by $k(v)$ the threshold of a vertex v in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$. Note that, unlike the usual symmetric threshold model, thresholds are not independent among the nodes of $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, but we can still deal with the $k(v) = qd_v$ case motivated by [24].

4.2. Phase transition for the symmetric threshold model with a single activation

For a graph $G = (V, E)$ and thresholds $\mathbf{k} = (k(v))_{v \in V}$, we consider the largest connected component of the induced subgraph in which we keep only vertices of threshold 0, and call the vertices in this component pivotal nodes. If only one pivotal node becomes active then the

whole set of pivotal nodes will eventually become active. Thus, if the set of pivotal nodes is large (in $\Theta(|V|)$ as $|V| \rightarrow \infty$) then a single pivotal node u can trigger a global cascade (i.e. the number of active nodes at the end of the epidemic starting from u is $\Theta(|V|)$).

We now consider the random graph $G = \tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, and set, for a node v , $C(v, \mathbf{t})$ for the final number of active vertices, when the initial state consists of only v active nodes with all other nodes inactive. Informally, we say that $C(v, \mathbf{t})$ is the size of the cascade induced by node v ; if $C(v, \mathbf{t}) = \Theta_{\mathbb{P}}(\tilde{n})$, we say that node v can trigger a global cascade.

Using the notation $\tilde{\gamma}$ defined in Proposition 1 and the binomial probabilities $b_{sr}(p)$ defined at the end of Section 1, we set (omitting the dependence on \mathbf{t} , \mathbf{p} , and $\boldsymbol{\gamma}$)

$$\begin{aligned}
 L(z) &:= \sum_s \frac{[s\gamma_s + (1 - \gamma_s)]p_s}{\tilde{\gamma}} t_{s0}(1 - z^s) \\
 &\quad + \sum_s \frac{(1 - \gamma_s)p_s}{\tilde{\gamma}} \left(1 - t_{s0} - \sum_{\ell \neq 0} t_{s\ell} \sum_{r \geq s-\ell} b_{sr}(z) \right), \\
 h(z) &:= \sum_s s p_s [t_{s0} z^s + \gamma_s (1 - t_{s0}) z] + \sum_s p_s (1 - \gamma_s) \sum_{s \geq \ell \neq 0} t_{s\ell} \sum_{r \geq s-\ell} r b_{sr}(z), \\
 \zeta &:= \sup\{z \in [0, 1) : \lambda z^2 = h(z)\}.
 \end{aligned} \tag{8}$$

Theorem 3. *We call the following condition the cascade condition:*

$$\sum_r r(r - 1)p_r t_{r0} > \sum_r r p_r. \tag{9}$$

(i) *If condition (9) is satisfied then there is a unique $\xi \in (0, 1)$ such that*

$$\sum_d d p_d t_{d0} (1 - \xi^{d-1}) = \lambda(1 - \xi). \tag{10}$$

Let $\tilde{\mathcal{P}}^{(n)}$ be the set of pivotal nodes in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$. Then we have

$$\frac{|\tilde{\mathcal{P}}^{(n)}|}{\tilde{n}} \xrightarrow{\mathbb{P}} \sum_d \frac{[d\gamma_d + (1 - \gamma_d)]p_d t_{d0}}{\tilde{\gamma}} (1 - \xi^d) > 0, \tag{11}$$

where $\tilde{\gamma}$ is defined in Proposition 1. Moreover, for any $u \in \tilde{\mathcal{P}}^{(n)}$, we have w.h.p.,

$$\liminf \frac{C(u, \mathbf{t})}{\tilde{n}} \geq L(\zeta) > 0,$$

where ζ is defined by (8). If, in addition, $\zeta = 0$ or ζ is such that there exists $\varepsilon > 0$ with $\lambda z^2 < h(z)$ for $z \in (\zeta - \varepsilon, \zeta)$, then we have, for any $u \in \tilde{\mathcal{P}}^{(n)}$,

$$\frac{C(u, \mathbf{t})}{\tilde{n}} \xrightarrow{\mathbb{P}} L(\zeta).$$

(ii) *If $\sum_r r(r - 1)p_r t_{r0} < \sum_r r p_r$ for a uniformly chosen node u , we have $C(u, \mathbf{t}) = o_{\mathbb{P}}(\tilde{n})$. The same result holds if $o(n)$ nodes are chosen uniformly at random.*

We use a branching process approximation to give the cascade condition and the size for the set of pivotal nodes. We approximate the random graph $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ by a rooted random graph

G^r whose root is a single node of degree d (respectively a clique of size d) with probability proportional to $(1 - \gamma_d)p_d$ (respectively $d\gamma_d p_d$), and in which each subsequent ‘individual’ of the root is a single node of degree d (respectively a clique of size d) with probability proportional to $d(1 - \gamma_d)p_d$ (respectively $d\gamma_d p_d$). The vertices in G^r that are active (after a long enough time, and when the root of G^r is initially active) are exactly those with a zero threshold. Let G_t^r be the subgraph of G^r induced by these vertices. Then the cascade condition (9) is the condition for G_t^r to be infinite with positive probability, and ξ is the extinction probability of G_t^r . In addition, if we assume that the root is a pivotal node (i.e. a node with zero threshold) then the survival probability of G_t^r is given by $\sum_d [d\gamma_d + (1 - \gamma_d)]p_d t_{d0}(1 - \xi^d)/\tilde{\gamma}$, as in the right-hand side of (11).

When $\gamma_r = 0$ for all $r \geq 0$ in Theorem 3, we recover a result of [21]. When we add cliques in the graph, the effect on the epidemic is described by the following lemma.

Lemma 1. *We consider a clique in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ where all vertices are inactive, and at least one has a neighbor outside the clique which is active. If the threshold k of the (vertices in the) clique is 0, then the epidemic will propagate to the whole clique. On the contrary, if k is positive then the clique cannot become active, even if all neighbors outside are active.*

Indeed, if $k = 0$, each vertex in the clique needs only one active neighbor to become active. If $k > 0$, each vertex in the clique needs at least two active neighbors to become active. Yet each vertex of the clique has only one (active) neighbor outside, other neighbors being (inactive ones) inside the clique. Hence, a clique with positive threshold in which all vertices are initially inactive will always stop the epidemic. This observation that cliques can stop the epidemic allows a comparison between the epidemics in $G(n, \mathbf{d})$ and $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$: if there is no global cascade in $G(n, \mathbf{d})$ then it is also the case in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ (see the proof in Section 5.4 for more details). What is remarkable is that the converse is also true: if there is a global cascade in the original graph $G(n, \mathbf{d})$ then there is one in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ (note that the cascade condition (9) depends only on the original distribution \mathbf{p} and threshold distribution \mathbf{t} , and not on $\boldsymbol{\gamma}$). More precisely, the cliques with positive threshold will reduce the *size* of the cascade, but they have no influence on whether a cascade is possible or not. It comes from the fact that a global cascade is possible if and only if the set of pivotal nodes is large (this equivalence is shown in [21] for the random graph $G(n, \mathbf{d})$).

The two graphs ($G(n, \mathbf{d})$ and $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$) do not however have the same asymptotic degree distribution. What is interesting now is to compare two graphs that have the same asymptotic degree distribution $\tilde{\mathbf{p}} = (\tilde{p}_r)_{r \geq 0}$, but different clustering coefficients.

4.3. Effect of clustering on the contagion threshold

We use our results to highlight the effect of clustering for the game-theoretic contagion model proposed by Blume [6] and Morris [24], in which the threshold distribution is given by $t_{s\ell} = \mathbf{1}(\lfloor qs \rfloor = \ell)$ for all $0 \leq \ell \leq s$. The cascade condition (9) is satisfied if and only if the parameter q of the contagion is less than the contagion threshold

$$q_c := \sup \left\{ q : \sum_{r < q^{-1}} r(r - 1)p_r > \sum_r r p_r \right\}. \tag{12}$$

We restrict ourselves to the case where $\gamma_r = \gamma$ for all $r \geq 0$ and we use Proposition 3 to construct two graphs with the same asymptotic degree distribution $\tilde{\mathbf{p}}$: one with a positive clustering coefficient and another with no clustering. We then compare the contagion thresholds in these two graphs.

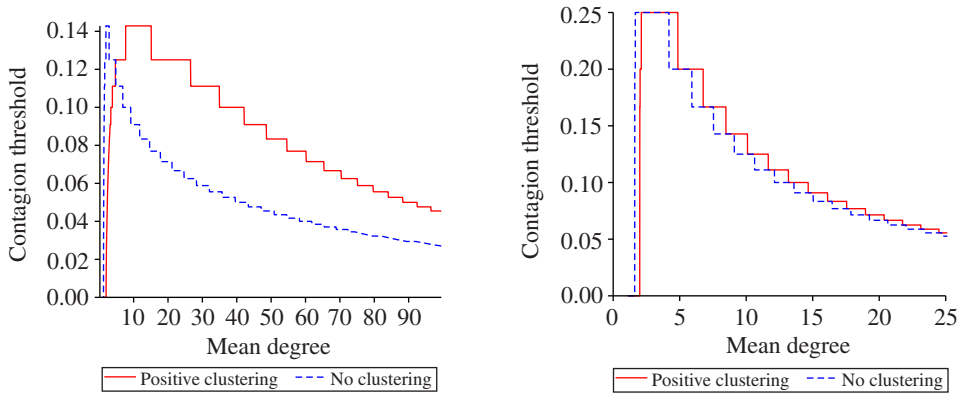


FIGURE 4: *Left*: contagion thresholds in two graphs with the same degree distribution $\tilde{p}_r \propto r^{-\tau} e^{-r/50}$. *Right*: contagion thresholds in two graphs with the same degree distribution $\tilde{p}_r = e^{-\lambda} \lambda^{r-1} / (r-1)!$.

In the left-hand diagram of Figure 4 we consider a power-law degree distribution with parameter $\tau > 0$ and exponential cutoff: for all $r \geq 1$, $\tilde{p}_r \propto r^{-\tau} e^{-r/50}$. On the one hand, we consider the graph $G^1(\tau)$ (solid line) for $C = C^{\max}$ (so that $\gamma = 1$ and the distribution in the original graph is $p_r \propto r^{-(\tau+1)} e^{-r/50}$). On the other hand, we consider the graph $G^0(\tau)$ (dashed line) for $C = 0$, i.e. $\gamma = 0$. In this subfigure we make the parameter τ vary: the solid (respectively dashed) curve corresponds to the contagion threshold $q_c^1(\tau)$ (respectively $q_c^0(\tau)$) of the graph $G^1(\tau)$ (respectively $G^0(\tau)$), defined in (12). Contagion thresholds are given with respect to the mean degree $\tilde{\lambda} = \sum_r r \tilde{p}_r$ (which is a decreasing function of τ).

In the right-hand diagram of Figure 4 we consider another form for the degree distribution \tilde{p} : let $\lambda > 0$, and set $\tilde{p}_r = e^{-\lambda} \lambda^{r-1} / (r-1)!$ for all $r \geq 1$. As before, we consider the graph $G^1(\lambda)$ for $C = C^{\max}$ ($\gamma = 1$ and \mathbf{p} is a Poisson distribution with parameter λ : $p_r = e^{-\lambda} \lambda^r / r!$), and the graph $G^0(\lambda)$ with $C = 0$, i.e. $\gamma = 0$. In this subfigure we plot the contagion thresholds for these two graphs, with respect to the mean degree $\tilde{\lambda} = \lambda + 1$.

Both the left- and right-hand diagrams of Figure 4 show that, when the mean degree $\tilde{\lambda}$ of the graph is low, the contagion threshold q_c^0 of the graph with no clustering is greater than the threshold q_c^1 of the graph with positive clustering. Hence, if the parameter q of the contagion process is in the interval (q_c^1, q_c^0) , a global cascade is possible only in the graph with no clustering: in that case, the clustering ‘inhibits’ the contagion process. On the contrary, for high values of the mean degree, we have $q_c^0 < q_c^1$, so the clustering increases the range of parameter q for which a global cascade is possible.

Now we study more precisely what happens if we fix the mean degree in the graph (which corresponds to a vertical cut in Figure 4), and increase the clustering coefficient between 0 and its maximal value C^{\max} .

In the left-hand (respectively middle and right-hand) diagram of Figure 5 we consider a power-law degree distribution with exponential cutoff: $\tilde{p}_r \propto r^{-\tau} e^{-r/50}$, with parameter $\tau = 2.5$ (respectively $\tau = 1.81$ and $\tau = 0.1$). We plot the contagion threshold q_c for the graph given by Proposition 3, when the degree distribution is \tilde{p} and the clustering coefficient varies from 0 to C^{\max} . We consider three different slices of Figure 4 (left), and we go from the dashed curve ($C = 0$) to the solid curve ($C = C^{\max}$), progressively increasing the clustering coefficient. For a very low value of the mean degree ($\tilde{\lambda} \approx 1.65$ —left-hand diagram of Figure 5), the contagion threshold decreases with the clustering. The opposite happens when the mean degree is very

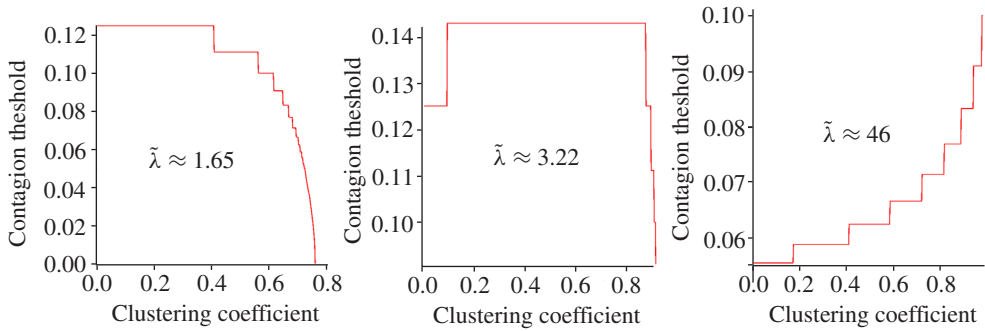


FIGURE 5: Evolution of the contagion threshold in a graph with mean degree $\tilde{\lambda}$ with respect to the clustering coefficient C (for a fixed power-law degree distribution).

high ($\tilde{\lambda} \approx 46$ —right-hand diagram of Figure 5). In addition, for some intermediate values of the mean degree, as for $\tilde{\lambda} \approx 3.22$ (middle diagram of Figure 5), low values of the clustering ‘helps’ the contagion process, but, as the clustering coefficient becomes higher, the opposite happens: it ‘inhibits’ more and more the contagion process.

4.4. Effect of clustering on the cascade size for the contagion model

We consider the game-theoretic contagion model proposed by Morris [24] and the case in which $\gamma_r = \gamma$ for all $r \geq 0$, as in the previous subsection. Now the parameter $q \in (0, 1)$ of the contagion process is fixed, and we want to highlight the effect of the clustering on the cascade size.

First we compare two graphs with the same asymptotic degree distribution \tilde{p} : one with a positive clustering coefficient and another with no clustering. In Figure 6 we plot the sizes of the cascade and the pivotal nodes set for each of these graphs.

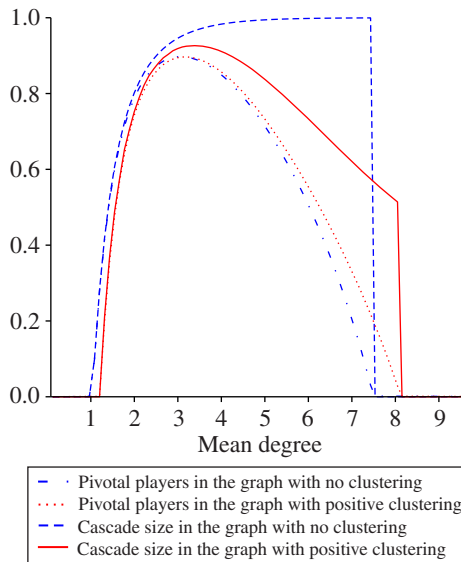


FIGURE 6: Set of pivotal nodes and cascade sizes for $q = 0.15$.

More precisely, in Figure 6 we fix $q = 0.15$. The solid curves correspond to a graph with positive clustering, constructed as follows: we start from a Poisson distribution with parameter λ for \mathbf{p} , and $\gamma = 0.2$. This gives

$$\tilde{p}_r = \frac{0.2r + 0.8 e^{-\lambda} \lambda^r}{0.2\lambda + 0.8 r!}$$

and clustering coefficient $C = 0.2\lambda / (0.2\lambda + 1.2) > 0$. The dashed curves correspond to a graph with the same asymptotic distribution $\tilde{\mathbf{p}}$, but no clustering (in this case, $\mathbf{p} = \tilde{\mathbf{p}}$ and $\gamma = 0$). We make the parameter λ vary, and the sizes of the cascade (solid and dashed lines) and the pivotal nodes set (dotted and dot-dash lines) are plotted with respect to the mean degree $\tilde{\lambda} = \sum_r r \tilde{p}_r$ in the graph.

For each graph, we observe that there is a cascade if and only if the set of pivotal nodes is large, as explained in Section 4.2. In addition, the interval of mean degrees for which a cascade is possible moves to the right when the clustering coefficient increases, which is consistent with our observations on Figure 4. Finally, we observe that the size of the cascade (when it exists) decreases with the clustering. This comes from the fact that cliques of degree $d \geq q^{-1}$ (i.e. cliques with positive threshold) stop the contagion process (as explained in Lemma 1). In the extreme case when $\gamma = 1$ (each vertex of degree d is replaced by a clique of size d), the cascade is exactly the set of pivotal nodes. When the probability γ of replacing a vertex by a clique increases, the cascade triggered by a pivotal node becomes closer and closer to the set of pivotal nodes only (until it is exactly the set of pivotal nodes). We refer the reader to [11] for more numerical results about the effect of the clustering on the cascade size.

4.5. Phase transition for the symmetric threshold model with degree-based activation

In this subsection we still consider the symmetric threshold model of Section 4.1, but we allow a positive fraction of nodes to be active at the beginning of the process. More precisely, we define an adaptation of the usual degree-based activation (defined in Section 3.4) for the random graph $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$. For all $d \geq 0$, let $\alpha_d \in (0, 1]$, and set $\boldsymbol{\alpha} = (\alpha_d)_{d \geq 0}$. We first consider the usual degree-based activation on the original graph $G(n, \mathbf{d})$: for each vertex i (of degree d_i in $G(n, \mathbf{d})$), we draw a Bernoulli random variable $a(i)$ with parameter α_{d_i} . When a vertex i of $G(n, \mathbf{d})$ is replaced by a clique in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, we associate to each vertex inside the clique the same activation variable $a(i)$ (if i is not replaced by a clique, it keeps its own activation variable). Each vertex v in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ belongs to the initial seed S if and only if $a(v) = 1$. Note that each node of degree d in $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ belongs to S with probability $\alpha_d > 0$ (since vertices inside the clique generated by i have the same degree as i). Thus, the only difference with the usual degree-based activation on $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ is that activation variables are not independent *inside* a clique.

Using the notation $\tilde{\gamma}$ defined in Proposition 1 and the binomial probabilities $b_{sr}(p)$ defined at the end of Section 1, we define (omitting the dependence on $\boldsymbol{\alpha}, \mathbf{t}, \mathbf{p}$, and $\boldsymbol{\gamma}$)

$$L(z) := \sum_s \frac{[s\gamma_s + (1 - \gamma_s)]p_s}{\tilde{\gamma}} [(1 - \alpha_s)t_{s0}(1 - z^s) + \alpha_s] + \sum_s \frac{(1 - \gamma_s)p_s}{\tilde{\gamma}} (1 - \alpha_s) \left(1 - t_{s0} - \sum_{\ell \neq 0} t_{s\ell} \sum_{r \geq s - \ell} b_{sr}(z) \right),$$

$$\begin{aligned}
 h(z) &:= \sum_s (1 - \alpha_s) s p_s [t_{s0} z^s + \gamma_s (1 - t_{s0}) z] \\
 &\quad + \sum_s (1 - \alpha_s) p_s (1 - \gamma_s) \sum_{s \geq \ell \neq 0} t_{s\ell} \sum_{r \geq s - \ell} r b_{sr}(z), \\
 \zeta &:= \sup\{z \in [0, 1) : \lambda z^2 = h(z)\}.
 \end{aligned}
 \tag{13}$$

Theorem 4. *For the symmetric threshold model defined in Section 4.1, on the random graph $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, if $\zeta = 0$, or if $\zeta \in (0, 1]$, and, furthermore, ζ is such that there exists $\varepsilon > 0$ with $\lambda z^2 < h(z)$ for $z \in (\zeta - \varepsilon, \zeta)$, then the size $C(\mathbf{t}, \boldsymbol{\alpha})$ of the active nodes at the end of the symmetric threshold process satisfies $C(\mathbf{t}, \boldsymbol{\alpha})/\tilde{n} \xrightarrow{\mathbb{P}} L(\zeta)$.*

5. Proofs

5.1. Generalities

5.1.1. *Configuration model.* In order to prove Theorem 1, it will be more convenient to work with the configuration model $G^*(n, \mathbf{d})$ (see, for instance, [7]): each vertex i , $1 \leq i \leq n$, has d_i half-edges, and the random graph $G^*(n, \mathbf{d})$ is obtained by taking a uniform matching among all possible matchings of half-edges into pairs. Conditioned on this multigraph being simple, it is distributed as $G(n, \mathbf{d})$. Condition 1 implies (in particular) that

$$\liminf \mathbb{P}(G^*(n, \mathbf{d}) \text{ is simple}) > 0
 \tag{14}$$

(see [17]), which allows us to directly transfer results that hold in probability for $G^*(n, \mathbf{d})$ to the model $G(n, \mathbf{d})$.

As for the simple graph, we consider the model $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$: we associate to each $i \in [n]$ a Bernoulli variable $X(i)$ with parameter γ_{d_i} , all variables being independent. If $X(i) = 1$, we replace node i by a clique of size d_i in which each vertex has exactly $d_i - 1$ neighbors inside the clique, and one half-edge outside. Then we match half-edges as for $G^*(n, \mathbf{d})$. Hence, $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$ is simple if and only if $G^*(n, \mathbf{d})$ is. So, conditioned on $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$ being simple, it is distributed as $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$, and (14) implies that $\liminf \mathbb{P}(\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma}) \text{ is simple}) > 0$. Thus, Theorems 1–3 can be proved for either the model $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$ or $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ (and they will be true for both).

5.1.2. *Definitions and additional notation.* Let \tilde{G} be distributed as $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$. We say that a vertex in \tilde{G} has *parent* $i \in [n]$ if it belongs to a clique that replaces the vertex i of $G(n, \mathbf{d})$ (when $X(i) = 1$) or if it is i (when $X(i) = 0$). For any subgraph $\tilde{H} \subset \tilde{G}$, we obtain the graph $\phi(\tilde{H})$ by identifying in \tilde{H} the vertices that have the same parent and that are connected in \tilde{H} (see Figure 7). We use the same definitions when \tilde{G} is distributed as $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$.

For any graph G , set $\nu(G)$ for the number of vertices in G and $\nu_r(G)$, $r \geq 0$, for the number of vertices with degree r in G .

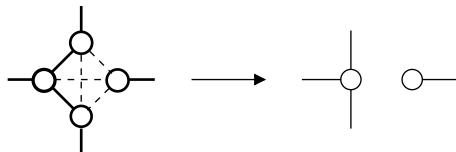


FIGURE 7: Transformation of a subgraph $\tilde{H} \subset \tilde{G}$ by ϕ . *Left:* four vertices of $\tilde{H} \subset \tilde{G}$ with the same parent; thick lines are in both \tilde{H} and \tilde{G} , while dashed lines are in \tilde{G} only. *Right:* these vertices in $\phi(\tilde{H})$.

5.1.3. *Link between the graph $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ and the original graph $\tilde{G}(n, \mathbf{d})$.* The next lemma will be useful in several proofs.

Lemma 2. *Let \tilde{H} be a subgraph of $\tilde{G} = \tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ such that $v(\phi(\tilde{H})) = o_{\mathbb{P}}(n)$. Then we have $v(\tilde{H}) = o_{\mathbb{P}}(\tilde{n})$.*

Proof. Set $H = \phi(\tilde{H})$. Using Cauchy-Schwarz inequality, we have

$$\frac{v(\tilde{H})}{n} \leq \frac{\sum_r r v_r(H)}{n} \leq \sqrt{\frac{\sum_i d_i^2}{n}} \sqrt{\frac{v(H)}{n}}.$$

The result follows from Condition 1(iii) and Proposition 1.

5.2. Proof of Theorem 1

We first give the idea of the proof. Let \tilde{G} be distributed as $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$ and $\pi \in [0, 1]$. In the percolated graph \tilde{G}_π , the removal of some edges inside a clique can split the clique into several connected components. We thus proceed in three steps.

Step 1. Let $\tilde{G}_\pi^{(1)}$ be the random graph constructed from \tilde{G} by deleting, independently and with probability $1 - \pi$, only the edges that are *inside* a clique. Conditioned on its number of vertices n' and its degree sequence \mathbf{d}' , the projected graph $G' = \phi(\tilde{G}_\pi^{(1)})$ is distributed as $G^*(n', \mathbf{d}')$. In this step, we compute the asymptotic distribution of the degree sequence \mathbf{d}' .

Step 2. We apply results of [16] in order to study the component sizes in the percolated graph G'_π (i.e. we now delete independently and with probability $1 - \pi$ only the edges that are *outside* a clique).

Step 3. By construction, G'_π is distributed as $\phi(\tilde{G}_\pi)$. We deduce the component sizes in \tilde{G}_π from those of $\phi(\tilde{G}_\pi)$ computed in the previous step.

In the following, when we consider the model $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$, we take the multiplicity of edges into account when we compute the degree of a vertex. More precisely, we say that a vertex in $G^*(n, \mathbf{d})$ or $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$ has ‘degree’ d if it has d (simple) half-edges. For instance, each loop of a given vertex has contribution 2 in its degree.

Step 1. For $d \geq 1$, let $V_d^{(n)}$ be the set of vertices i in $G(n, \mathbf{d})$ with degree d and such that $X(i) = 1$: i is replaced by a clique $K(i)$ of size d in \tilde{G} . Let $K(i, \pi)$ be the subgraph of $K(i)$ obtained after a bond percolation of parameter π . Let $\tilde{F}_d(\pi)$ be the subgraph of \tilde{G}_π containing the percolated version of the cliques with initial size d : $\tilde{F}_d(\pi) = \bigcup_{i \in V_d^{(n)}} K(i, \pi)$. The next lemma gives the limit, as $n \rightarrow \infty$, for the number $N^{(n)}(d, k, \pi)$ of connected components in $\tilde{F}_d(\pi)$ whose size is $k \leq d$.

Lemma 3. *For any $d \geq 1$ and $k \leq d$, we have $N^{(n)}(d, k, \pi)/n \xrightarrow{\mathbb{P}} (d/k)f(d, k, \pi)p_d\gamma_d$, where $f(d, k, \pi)$ is given by (2).*

Proof. For each vertex i in $V_d^{(n)}$, we label the vertices of $K(i, \pi)$ from 1 to d . We look at all the vertices with label 1, and we let $M^{(n)}(d, k, \pi)$ be the number of such vertices whose connected component in $K(i, \pi)$ has size k . Using the law of large numbers and the fact that $|V_d^{(n)}|/n \xrightarrow{\mathbb{P}} p_d\gamma_d$, we have $M^{(n)}(d, k, \pi)/n \xrightarrow{\mathbb{P}} f(d, k, \pi)p_d\gamma_d$, where $f(d, k, \pi)$ is by definition the probability that the component of 1 contains k vertices. So the total number of vertices in $\tilde{F}_d(\pi)$ that belongs to a component of size k is $df(d, k, \pi)p_d\gamma_d n + o_{\mathbb{P}}(n)$ and, in order to have the number of such components, we have to divide by k , which proves the lemma.

Let $\tilde{G}_\pi^{(1)}$ be the graph obtained from $\tilde{G}^*(n, \mathbf{d}, \boldsymbol{\gamma})$ when we replace each vertex i such that $X(i) = 1$ by the percolated clique $K(i, \pi)$. For any $k \geq 0$, let n'_k be the number of vertices with ‘degree’ k in the projected graph $G' = \phi(\tilde{G}_\pi^{(1)})$. In order to compute n'_k , we have to consider the vertices i such that $X(i) = 0$ (there are $n_k - |V_k^{(n)}|$ such vertices, where n_k is the number of vertices with ‘degree’ k in $G^*(n, \mathbf{d})$), and the vertices that come from a clique of initial size d for some $d \geq k$ (each such vertex corresponds to a component of size k in $\tilde{F}_d(\pi)$, so there are $N^{(n)}(d, k, \pi)$ such vertices). This gives the following relation for all $k \geq 0$: $n'_k = n_k - |V_k^{(n)}| + \sum_{d \geq k} N^{(n)}(d, k, \pi)$. So Lemma 3 gives the following asymptotic distribution for the degree sequence \mathbf{d}' .

Lemma 4. *Let $n' := \sum_k n'_k$ be the total number of vertices in G' . Then the proportion of vertices with degree k in G' has the limit*

$$\frac{n'_k}{n'} \xrightarrow{\mathbb{P}} p'_k := \frac{\varrho_k}{\sum_\ell \varrho_\ell} \text{ as } n \rightarrow \infty,$$

where $\varrho_k := p_k(1 - \gamma_k) + \sum_{d \geq k} (d/k) f(d, k, \pi) p_d \gamma_d$.

In addition, the uniform summability of kn_k/n implies the uniform summability of kn'_k/n' , so $\sum_k kn'_k/n' \xrightarrow{\mathbb{P}} \sum_k kp'_k$.

Step 2. We apply Theorem 3.9 of [16] to the random graph G' . Indeed, we can assume without loss of generality that the previous convergences (Lemma 4) hold almost surely, and not just in probability (as in [16]: using the Skorokhod coupling theorem, see [19, Theorem 3.30] for instance, or arguing by selecting suitable subsequences). Then there is a giant component in the percolated graph G'_π if and only if $\pi \sum_d d(d-1)p'_d > \sum_d dp'_d$, which is equivalent to the fact that $\pi \mathbb{E}[\mathcal{K}(D^* + 1, \pi, \boldsymbol{\gamma}) - 1] > 1$.

Step 3. The proof of (ii) follows easily from the previous step, and Lemma 2. We give the main lines of the proof of (i). Assume that $\pi > \pi_c$, which corresponds to $\pi \sum_d d(d-1)p'_d > \sum_d dp'_d$. Let \mathcal{C}_1 be the largest connected component in $G'_\pi = \phi(\tilde{G}_\pi)$, and let $\check{\mathcal{C}}_1$ be the connected component of \tilde{G}_π such that $\phi(\check{\mathcal{C}}_1) = \mathcal{C}_1$.

We compute the limit of $v(\check{\mathcal{C}}_1)/\tilde{n}$ as $n \rightarrow \infty$. The results in [16, see Section 2.2 and the proof of Theorem 3.9] show that, for $r \geq 0$,

$$\frac{v_r(\mathcal{C}_1)}{n'} \xrightarrow{\mathbb{P}} \sum_{\ell \geq r} b_{\ell r}(\sqrt{\pi}) p'_\ell (1 - (1 - \pi^{-1/2} + \pi^{-1/2} \zeta)^r), \tag{15}$$

where ζ is defined in (6) and is the unique solution in $(0, 1)$ of $\mu \zeta(1 - \pi + \pi \zeta) = h(\zeta)$. Computing $v(\check{\mathcal{C}}_1)$ requires a little more precision than the computation in (15). We set $\xi := 1 - \pi^{-1/2} + \pi^{-1/2} \zeta$, and let $v_r^0(\mathcal{C}_1)$ and $v_r^1(\mathcal{C}_1)$ be the numbers of vertices i with degree r in \mathcal{C}_1 such that $X(i) = 0$ and $X(i) = 1$, respectively. Then we have

$$\begin{aligned} \frac{v_r^0(\mathcal{C}_1)}{n'} &\xrightarrow{\mathbb{P}} \sum_{\ell \geq r} b_{\ell r}(\sqrt{\pi}) p_\ell (1 - \gamma_\ell) \frac{1 - \xi^r}{\varrho}, \\ \frac{v_r^1(\mathcal{C}_1)}{n'} &\xrightarrow{\mathbb{P}} \sum_{\ell \geq r} b_{\ell r}(\sqrt{\pi}) \sum_{d \geq \ell} \frac{d}{\ell} f(d, \ell, \pi) p_d \gamma_d \frac{1 - \xi^r}{\varrho}. \end{aligned}$$

To summarize, d represents the degree of vertices in the initial graph \tilde{G} , ℓ represents the degree of vertices in $\tilde{G}_\pi^{(1)}$ (after the percolation *inside* cliques), and r represents the degree of vertices after

the percolation on *external* edges. In order to recover $v(\check{\mathcal{C}}_1)$, we multiply each term in $v_r^1(\mathcal{C}_1)$ by ℓ , and then sum over all r , which gives $v(\check{\mathcal{C}}_1)/n' \xrightarrow{\mathbb{P}} (1/\varrho) \sum_{k \geq 1} \sigma_k (1 - (1 - \pi^{1/2} + \pi^{1/2} \xi)^k)$ (exchanging summations on r and ℓ). Using the facts that $n'/n \xrightarrow{\mathbb{P}} \varrho$ and $\tilde{n}/n \xrightarrow{\mathbb{P}} \tilde{\gamma}$, we obtain $v(\check{\mathcal{C}}_1)/\tilde{n} \xrightarrow{\mathbb{P}} L(\zeta)$.

Let $\check{\mathcal{C}}_1$ be the largest component in \tilde{G}_π . We prove that $\check{\mathcal{C}}_1 = \tilde{\mathcal{C}}_1$ w.h.p. (adding cliques changes the sizes of the connected components). Let $\check{\mathcal{C}}$ be any other component of \tilde{H} different from $\check{\mathcal{C}}_1$. Its projection $\mathcal{C} = \phi(\check{\mathcal{C}})$ is different from \mathcal{C}_1 , so Theorem 3.9 of [16] implies that $v(\mathcal{C})/n \xrightarrow{\mathbb{P}} 0$. Using Lemma 2 with $H = \mathcal{C}$ shows that $v(\check{\mathcal{C}})/\tilde{n} \xrightarrow{\mathbb{P}} 0$. Hence, $\check{\mathcal{C}}_1$ is the largest connected component of \tilde{G} w.h.p., which completes the proof.

5.3. Proof of Theorem 2

This proof differs from the previous in that instead of using Theorem 3.9 of [16] in steps 2 and 3, we use Theorem 10 of [21].

Indeed, the first step is the same: the graph $G' = \phi(\tilde{G}_\pi^{(1)})$ (where $\tilde{G}_\pi^{(1)}$ is the graph obtained from $\tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ after a bond percolation on the edges *inside* cliques only) has asymptotic degree distribution $\mathbf{p}' = (p'_k)_k$, with $p'_k = \varrho_k/\varrho$.

We apply (a slight extension of) Theorem 10 of [21] for the graph G' (with $t_{s\ell} = \mathbf{1}_{\{\ell=0\}}$). Let v_s^0 be the number of vertices i such that $X(i) = 0$, and assume that the degree of i in G' (that is to say before the bond percolation in G') is s and that i is active at the end of the process. Let v_{ds}^1 be the number of vertices i such that $X(i) = 1$, and assume that the degree of i in the original graph $\phi(\tilde{G})$ is d , that the degree of i in G' is $s \leq d$, and that i is active at the end of the process. The probability that such a node i (with degree d in $\phi(\tilde{G})$ and s in G') does not belong to the original seed S is $(1 - \alpha_d)^s$ (and initial activations are independent among nodes). Hence, we have

$$\begin{aligned} \frac{v_s^0}{n'} &\xrightarrow{\mathbb{P}} \frac{p_s(1 - \gamma_s)}{\varrho} [1 - (1 - \alpha_s)(1 - \pi + \pi \zeta)^s], \\ \frac{v_{ds}^1}{n'} &\xrightarrow{\mathbb{P}} \frac{d}{s} \frac{f(d, s, \pi) \gamma_d p_d}{\varrho} [1 - (1 - \alpha_d)^s (1 - \pi + \pi \zeta)^s], \end{aligned}$$

where ζ is given by (7). The result follows easily (using the fact that $C^b(\pi, \boldsymbol{\alpha}) = \sum_s v_s^0 + \sum_{d,s} s v_{ds}^1$).

5.4. Proof of Theorem 3

For any graph G and any vertex v of G , let $D(v, \mathbf{t})$ be the subgraph of G induced by the final set of active vertices, when v is the only vertex in the initial seed. When H is a subgraph of G , we set $D(H, \mathbf{t})$ for the subgraph induced by the final set of active vertices in G , when the initial vertices in the seed are those of H .

In the whole proof, we set $\tilde{G} := \tilde{G}(n, \mathbf{d}, \boldsymbol{\gamma})$ and $G := \phi(\tilde{G})$ (distributed as $G(n, \mathbf{d})$).

We first prove case (ii) using the next lemma.

Lemma 5. *Let u be a vertex of \tilde{G} , and let i be its parent in G . Then we have*

$$\phi(D(u, \mathbf{t})) \subset D(i, \mathbf{t}). \tag{16}$$

Assume that $\sum_r r(r - 1)p_r t_{r0} < \sum_r r p_r$ and that u is chosen uniformly at random among the vertices of \tilde{G} . Then we have

$$C(i, \mathbf{t}) = v(D(i, \mathbf{t})) = o_{\mathbb{P}}(n). \tag{17}$$

Before proving this lemma, we conclude the proof of case (ii). Using (16) and (17), we obtain $v(\phi(D(u, \mathbf{t}))) = o_{\mathbb{P}}(n)$. We then apply Lemma 2 with $\tilde{H} = D(u, \mathbf{t})$.

Proof of Lemma 5. Let K be the clique generated by i if $X(i) = 1$ (otherwise, set $K = \{u\}$). Then we have $\phi(D(u, \mathbf{t})) \subset \phi(D(K, \mathbf{t}))$, and (16) follows from Lemma 1.

To prove (17), we apply Theorem 10 of [21], with a parameter $\alpha = (\alpha_d)_{d=0}^{\infty}$ that satisfies $\alpha_d = (d\gamma_d + 1 - \gamma_d)\alpha$ for all d , where α is a positive constant. Then the same arguments as used in the proof of Theorem 11(ii) of [21] apply.

We now assume that the cascade condition (9) is satisfied and prove case (i). We first prove (11). Let \tilde{H} and H respectively be the subgraphs of \tilde{G} and G induced by the vertices of threshold 0, and let $\tilde{\mathcal{C}}_1$ and \mathcal{C}_1 respectively be the largest connected components in \tilde{H} and H . Note that $\phi(\tilde{H}) = H$. The number $v_r(\mathcal{C}_1)$ of vertices with degree r in \mathcal{C}_1 is computed in the proof of Theorem 11 of [21] (applied on the graph G , and with $\pi = 1$): $v_r(\mathcal{C}_1)/n \xrightarrow{\mathbb{P}} p_r t_{r0}(1 - \xi^r)$, where ξ is defined in (10). Hence, we can deduce the size of the connected component $\tilde{\mathcal{C}}_1$ in \tilde{H} such that $\phi(\tilde{\mathcal{C}}_1) = \mathcal{C}_1$: $v(\tilde{\mathcal{C}}_1)/\tilde{n} \xrightarrow{\mathbb{P}} \sum_d [d\gamma_d + (1 - \gamma_d)] p_d t_{d0}(1 - \xi^d)/\tilde{\gamma}$. We show that $\tilde{\mathcal{C}}_1 = \tilde{\mathcal{C}}_1$ w.h.p. by a similar argument as used at the end of the proof of Theorem 1. This proves (11).

Now the idea is to make a coupling between the epidemic on \tilde{G} (with threshold distribution \mathbf{t}), and an epidemic on G , with a threshold distribution $\mathbf{t}' = (t'_{s\ell})_{s,\ell}$ that we define below.

Proposition 5. *Assume that the epidemic on \tilde{G} starts from a vertex u that has threshold 0, and let i be the parent of u in G . We consider the following distribution of thresholds $\mathbf{t}' = (t'_{s\ell})_{0 \leq \ell \leq s}$ for each $s \geq 0$: $t'_{s0} = t_{s0}$; $t'_{s\ell} = (1 - \gamma_s)t_{s\ell}$ for all $0 < \ell < s$; $t'_{ss} = (1 - \gamma_s)t_{ss} + \gamma_s(1 - t_{s0})$. Then there exist random thresholds $(k'(j))_{1 \leq j \leq n}$ with this distribution \mathbf{t}' and such that $\phi(D(u, \mathbf{t})) = D(i, \mathbf{t}')$, where $D(i, \mathbf{t}')$ is the subgraph induced by the final set of active vertices in the symmetric threshold model starting from i in G , with threshold distribution \mathbf{t}' .*

Proof. As explained in Lemma 1, we can have $\phi(D(u, \mathbf{t})) \subsetneq D(i, \mathbf{t})$, since a clique K of \tilde{G} stops the epidemic: if j is the parent of K in G , we will change its original threshold $k(j)$ into a new threshold $k'(j)$ so high that j cannot become active in the new epidemic that we define on G .

More precisely, we start the epidemic in G from the parent i of u (if u belongs to a clique then the whole clique becomes active at the next step). Let v be the neighbor of u outside the clique (any neighbor if $X(i) = 0$), and let j be the parent of v . If $k(v) = 0$, then v (and its whole clique if it has one) becomes active (Lemma 1): we set $k'(j) := 0$, so that j also becomes active in G . If $k(v) > 0$, we distinguish two cases.

- If $X(j) = 1$, vertex v and its clique stay inactive (Lemma 1): we set $k'(j) := d_j$ (so that j stays inactive).
- If $X(j) = 0$, vertex v becomes active if and only if it has at least $k(v) + 1$ active neighbors: we set $k'(j) := k(v) = k(j)$.

Since the random variables $X(j)$ for j in G are independent, the thresholds we associate to each node are also independent. In addition, we can easily verify that the conditional probability distribution of thresholds (knowing that the degree of the node is s) is given by $(t'_{s\ell})_{0 \leq \ell \leq s}$.

More precisely, let $C_{s\ell}(u, \mathbf{t})$ (respectively $C'_{s\ell}(i, \mathbf{t}')$) be the final number of active vertices with degree $s \geq 0$ and threshold ℓ at the end of the symmetric threshold epidemic on \tilde{G} (respectively G), with threshold parameter \mathbf{t} (respectively \mathbf{t}'), when the only vertex in the initial

seed is u (respectively i). Then, using the coupling described above, we have the following results, for each degree $s \geq 0$.

- $C_{s0}(u, \mathbf{t}) = C'_{s0}(i, \mathbf{t}')[sY_s + (1 - Y_s)]$, where Y_s is the proportion of vertices j in G such that $X(j) = 1$, among those that have degree s and that belong to the cascade triggered by i .
- For all $\ell \neq 0$, we have $C_{s\ell}(u, \mathbf{t}) = C'_{s\ell}(i, \mathbf{t}')$, since the vertices of positive threshold that belong to the cascade triggered by u are exactly those that are not replaced by a clique.

We have $Y_s/n \xrightarrow{\mathbb{P}} \gamma_s$ for all s , and the limit for $C'_{s\ell}(i, \mathbf{t}')$ is given by the following lemma.

Lemma 6. *Assume (using the notation of Theorem 3) that $\zeta = 0$ or ζ is such that there exists $\varepsilon > 0$ with $\lambda z^2 < h(z)$ for $z \in (\zeta - \varepsilon, \zeta)$. Then, for any i that belongs to the set of pivotal nodes in G , we have*

$$C'_{s\ell}(i, \mathbf{t}')/n \xrightarrow{\mathbb{P}} p_s t'_{s\ell} (1 - \sum_{r \geq s-\ell} b_{sr}(\zeta)).$$

In particular, for $\ell = 0$, we have

$$C'_{s0}(i, \mathbf{t}')/n \xrightarrow{\mathbb{P}} p_s t'_{s0} (1 - \zeta^s).$$

Proof. By slight extension of Theorem 11 of [21], the number of inactive nodes with degree s in the original graph, degree r in the graph of inactive nodes and threshold ℓ tends to

$$\sum_{i \geq s-r-\ell} p_s t'_{s\ell} b_{sr}(\zeta) b_{s-r,i}(0) = p_s t'_{s\ell} b_{sr}(\zeta) \mathbf{1}\{r \geq s - \ell\}.$$

Hence, summing over r it follows that the number of inactive nodes with original degree s and threshold ℓ tends to $p_s t'_{s\ell} \sum_{r \geq s-\ell} b_{sr}(\zeta)$, which completes the proof.

We assume that $\zeta = 0$ or ζ is such that there exists $\varepsilon > 0$ with $\lambda z^2 < h(z)$ for $z \in (\zeta - \varepsilon, \zeta)$. Let u be a vertex in \tilde{G} whose parent i belongs to the set of pivotal nodes in G . Let $C_s(u, \mathbf{t})$ be the final number of active vertices with degree $s \geq 0$ at the end of the symmetric threshold epidemic on \tilde{G} , with threshold parameter \mathbf{t} , when the only vertex in the initial seed is u . Then we have

$$\frac{C_s(u, \mathbf{t})}{\tilde{n}} \xrightarrow{\mathbb{P}} \frac{[s\gamma_s + (1 - \gamma_s)]p_s}{\tilde{\gamma}} t'_{s0} (1 - \zeta^s) + \frac{p_s}{\tilde{\gamma}} \sum_{\ell \neq 0} t'_{s\ell} \left(1 - \sum_{r \geq s-\ell} b_{sr}(\zeta)\right).$$

Using the definition of \mathbf{t}' , we have $t'_{s0} = t_{s0}$ and

$$\sum_{\ell \neq 0} t'_{s\ell} \sum_{r \geq s-\ell} b_{sr}(\zeta) = \sum_{\ell \neq 0} (1 - \gamma_s) t_{s\ell} \sum_{r \geq s-\ell} b_{sr}(\zeta) + \gamma_s (1 - t_{s0}),$$

which finally gives

$$\frac{C_s(u, \mathbf{t})}{\tilde{n}} \xrightarrow{\mathbb{P}} \frac{[s\gamma_s + (1 - \gamma_s)]p_s}{\tilde{\gamma}} t_{s0} (1 - \zeta^s) + \frac{(1 - \gamma_s)p_s}{\tilde{\gamma}} \left(1 - t_{s0} - \sum_{\ell \neq 0} t_{s\ell} \sum_{r \geq s-\ell} b_{sr}(\zeta)\right).$$

Then, by an argument similar to that used at the end of the proof of Theorem 1 or (11), u belongs to the set of pivotal nodes in \tilde{G} , which completes the proof.

5.5. Proof of Theorem 4

We use the same idea as in the previous proof. The same statement as for Proposition 5 holds when the epidemic starts from a set (instead of a single vertex u). Indeed, let S be the initial seed in G . By definition, the initial seed \tilde{S} in \tilde{G} consists of the vertices whose parent belongs to S .

Let $C_{s\ell}(\tilde{S}, t)$ (respectively $C'_{s\ell}(S, t')$) be the final number of active vertices with degree $s \geq 0$ and threshold ℓ at the end of the symmetric threshold epidemic on \tilde{G} (respectively $\phi(\tilde{G})$), with threshold parameter t (respectively t' , defined in Proposition 5), when the initial seed is \tilde{S} (respectively S).

Using a slight extension of Theorem 10 of [21], we have, for all $s \geq 0$ and $\ell \geq 0$,

$$\frac{C'_{s\ell}(S, t')}{n} \xrightarrow{\mathbb{P}} p_s t'_{s\ell} \left(\alpha_s + (1 - \alpha_s) \left(1 - \sum_{r \geq s - \ell} b_{sr}(\zeta) \right) \right),$$

where ζ is defined in (13). More precisely, the first term $p_s t'_{s\ell} \alpha_s$ comes from the vertices that belong to the initial seed S , and the second term $p_s t'_{s\ell} (1 - \alpha_s) (1 - \sum_{r \geq s - \ell} b_{sr}(\zeta))$ comes from those that are activated during the process. In order to obtain the asymptotic for $C_{s\ell}(\tilde{S}, t)/n$, we have to multiply the first term by $(s\gamma_s + 1 - \gamma_s)$. The multiplicative constant for the second term depends on the value of the threshold ℓ : if $\ell = 0$, we multiply the second term by $(s\gamma_s + 1 - \gamma_s)$, and if $\ell > 0$, we multiply it by 1 (since the vertices with positive threshold that are activated during the process necessarily do not belong to a clique). Summing over s and ℓ , and replacing t' by its expression gives the following limit as $n \rightarrow \infty$:

$$\begin{aligned} \frac{C(t, \alpha)}{n} &\xrightarrow{\mathbb{P}} \sum_s p_s t_{s0} (s\gamma_s + 1 - \gamma_s) [\alpha_s + (1 - \alpha_s) (1 - \zeta^s)] \\ &\quad + \sum_s p_s (1 - \gamma_s) \alpha_s (s\gamma_s + 1 - \gamma_s) (1 - t_{s0}) \\ &\quad + \sum_s p_s (1 - \gamma_s) (1 - \alpha_s) \left[(1 - t_{s0}) - \sum_{\ell \neq 0} t_{s\ell} \sum_{r \geq s - \ell} b_{sr}(\zeta) \right] \\ &\quad + \sum_s p_s \gamma_s (1 - t_{s0}) \alpha_s (\gamma_s + 1 - \gamma_s). \end{aligned}$$

Gathering some terms and using the fact that $\tilde{n}/n \xrightarrow{\mathbb{P}} \tilde{\gamma}$ completes the proof of Theorem 4.

6. Conclusions

Our analysis is one of the first systematic studies of random graphs with both a tunable asymptotic degree distribution and a clustering coefficient. Our model allows a rigorous analysis of diffusion and symmetric threshold models.

For both models, we were able to derive explicit formulae for the cascade condition, i.e. the condition under which a single infected individual can turn a positive fraction of the population into infected individuals. When such a cascade was possible, we provided an analytic expression of its size. In the case of random regular graphs, we proved that the clustering ‘inhibits’ the diffusion process. Numerical evaluations also showed that clustering decreases the cascade size of the diffusion process for regular graphs, and ‘inhibits’ the diffusion process for power-law graphs. The impact of clustering on the symmetric threshold model was studied in the particular case of the contagion model [24]: numerical evaluations showed that the effect of clustering on

the contagion process depends on the value of the mean degree in the graph: while clustering ‘inhibits’ the contagion for a low mean degree, the contrary happens in the high-value regime. When a cascade was possible, we observed that clustering decreases its size.

As discussed in [22], the impacts of clustering and degree correlation also have to be separated, and it seems important, as a future work, to separate the impacts of these different features on the processes. In addition, we can also compute explicitly the cascade size for a degree-based activation, in both diffusion and symmetric threshold models. This theoretical analysis paves the way to a possible control of such epidemic processes, as done in [8] or [20].

Acknowledgements

We thank an anonymous referee whose comments helped to improve the paper, especially Proposition 4. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under grant ANR-11-JS02-005-01 (GAP project).

References

- [1] ACEMOGLU, D., OZDAGLAR, A. AND YILDIZ, E. (2011). Diffusion of innovations in social networks. In *Proc. 50th IEEE Conf. Decision Control*, IEEE, New York, pp. 2329–2334.
- [2] AMINI, H. (2010). Bootstrap percolation and diffusion in random graphs with given vertex degrees. *Electron. J. Combin.* **17**, paper 25.
- [3] BALL, F. AND SIRL, D. (2012). An SIR epidemic model on a population with random network and household structure, and several types of individuals. *Adv. Appl. Prob.* **44**, 63–86.
- [4] BALL, F., SIRL, D. AND TRAPMAN, P. (2009). Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. Appl. Prob.* **41**, 765–796.
- [5] BALL, F., SIRL, D. AND TRAPMAN, P. (2010). Analysis of a stochastic SIR epidemic on a random network incorporating household structure. *Math. Biosci.* **224**, 53–73.
- [6] BLUME, L. E. (1995). The statistical mechanics of best-response strategy revision. Evolutionary game theory in biology and economics. *Games Econom. Behav.* **11**, 111–145.
- [7] BOLLOBÁS, B. (2001). *Random Graphs* (Camb. Stud. Adv. Math. **73**), 2nd edn. Cambridge University Press.
- [8] BORGS, C., CHAYES, J., GANESH, A. AND SABERI, A. (2010). How to distribute antidote to control epidemics. *Random Structures Algorithms* **37**, 204–222.
- [9] BRITTON, T., DEIJFEN, M., LAGERÅS, A. N. AND LINDHOLM, M. (2008). Epidemics on random graphs with tunable clustering. *J. Appl. Prob.* **45**, 743–756.
- [10] COUPECHOUX, E. AND LELARGE, M. (2011). Impact of clustering on diffusions and contagions in random networks. In *Proc. 5th Internat. Conf. Network Games Control Optim.*, IEEE, New York, pp. 1–7.
- [11] COUPECHOUX, E. AND LELARGE, M. (2012). How clustering affects epidemics in random networks. Preprint. Available at <http://uk.arxiv.org/abs/1202.4974v1>.
- [12] COUPECHOUX, E. AND LELARGE, M. (2014). Contagions in random networks with overlapping communities. Preprint. Available at <http://uk.arxiv.org/abs/1303.4325v2>.
- [13] DEIJFEN, M. AND KETS, W. (2009). Random intersection graphs with tunable degree distribution and clustering. *Prob. Eng. Inf. Sci.* **23**, 661–674.
- [14] GILBERT, E. N. (1959). Random graphs. *Ann. Math. Statist.* **30**, 1141–1144.
- [15] GLEESON, J. P., MELNIK, S. AND HACKETT, A. (2010). How clustering affects the bond percolation threshold in complex networks. *Phys. Rev. E* **81**, 066114.
- [16] JANSON, S. (2009). On percolation in random graphs with given vertex degrees. *Electron. J. Prob.* **14**, 87–118.
- [17] JANSON, S. (2009). The probability that a random multigraph is simple. *Combin. Prob. Comput.* **18**, 205–225.
- [18] JANSON, S., ŁUCZAK, T. AND RUCINSKI, A. (2000). *Random Graphs*. Wiley-Interscience, New York.
- [19] KALLENBERG, O. (2002). *Foundations of Modern Probability*, 2nd edn. Springer, New York.
- [20] LELARGE, M. (2009). Efficient control of epidemics over random networks. In *SIGMETRICS '09*, ACM, New York, pp. 1–12.
- [21] LELARGE, M. (2012). Diffusion and cascading behavior in random networks. *Games Econom. Behav.* **75**, 752–775.
- [22] MILLER, J. C. (2009). Percolation and epidemics in random clustered networks. *Phys. Rev. E* **80**, 020901.
- [23] MOLLOY, M. AND REED, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures Algorithms* **6**, 161–179.
- [24] MORRIS, S. (2000). Contagion. *Rev. Econom. Stud.* **67**, 57–78.

- [25] NEWMAN, M. E. J. (2003). Properties of highly clustered networks. *Phys. Rev. E* **68**, 026121.
- [26] NEWMAN, M. E. J. (2003). The structure and function of complex networks. *SIAM Rev.* **45**, 167–256.
- [27] NEWMAN, M. E. J. (2009). Random graphs with clustering. *Phys. Rev. Lett.* **103** 058701.
- [28] SHAKED, M. AND SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. Springer, New York.
- [29] TRAPMAN, P. (2007). On analytical approaches to epidemics on networks. *Theoret. Pop. Biol.* **71**, 160–173.
- [30] WATTS, D. J. (2002). A simple model of global cascades on random networks. *Proc. Nat. Acad. Sci. USA* **99**, 5766–5771.