Impact of Clustering on Diffusions and Contagions in Random Networks

Emilie Coupechoux, Marc Lelarge

Impact of Clustering on Diffusions and Contagions in Random Networks

Emilie Coupechoux, *INRIA - ENS*, and Marc Lelarge, *INRIA - ENS*
E-mail: Emilie.Coupechoux, Marc.Lelarge @ens.fr

***Abstract*—Motivated by the analysis of social networks, we study a model of network that has both a tunable degree distribution and a tunable clustering coefficient. We compute the asymptotic (as the size of the population tends to infinity) for the number of acquaintances and the clustering for this model. We analyze a contagion model with threshold effects and obtain conditions for the existence of a large cascade. We also analyze a diffusion process with a given probability of contagion. In both cases, we characterize conditions under which a global cascade is possible.**

***Index Terms*—Contagion threshold, diffusion, Random graphs, clustering**

I. Introduction

Most of the epidemic models [11], [15] consider a transmission mechanism which is independent of the local condition faced by the agents concerned. There is now a vast literature on epidemics in complex networks (see [12] for a review) and there is now a good understanding of the impact of the topology on the spread of an epidemic. But if there is a factor of persuasion or coordination involved, relative considerations tend to be important in understanding whether some new behavior or belief is adopted [17].

In social contexts, the diffusion of information and behavior often exhibits features that do not match well those of the SIR or SIS model [17]. In the classical (SI) diffusion model, an individual is influenced by each of her neighbors independently. For the spread of a new technology in the network, a rather appropriate model is the case where each individual adopts the technology as soon as enough of her neighbors have already adopted it: this corresponds to the basic game-theoretic contagion model proposed by Morris [10]. Consider a graph G in which the nodes are the individuals in the population and there is an edge (i, j) if i and j can interact with each other. Each node has a choice between two possible behaviors labeled A and B. On each edge (i, j), there is an incentive for i and j to have their behaviors match, which is modeled as the following coordination game parametrized by a real number $q \in (0, 1)$: if i and j choose A (resp. B), they each receive a payoff of q (resp. $(1 - q)$); if they choose opposite strategies, then they receive a payoff of 0. Then the total payoff of a player is the sum of the payoffs with each of her neighbors. If the degree of node i is d_i and S_i^B is the number of its neighbors playing B, then the payoff to i from choosing A is $q(d_i - S_i^B)$ while the payoff from choosing B is $(1 - q)S_i^B$. Hence, in best response update, i should adopt B if $S_i^B > qd_i$ and A if $S_i^B \leq qd_i$. A number of qualitative insights can be derived from a diffusion model even at this level of simplicity. Specifically, consider a network where all nodes initially play A. If a small number of nodes are forced to adopt strategy B (the seed) and we apply best-response updates to other nodes in the network, then these nodes will be repeatedly applying the following rule: switch to B if enough of your neighbors have already adopted B. There can be a cascading sequence of nodes switching to B such that a network-wide equilibrium is reached in the limit.

Large complex networks such as social contact structures, the internet and various types of collaboration networks have received a lot of attention during the last few years; [14] and the references therein. As for social networks, one of their most striking features is that they are highly clustered, meaning that there is a large number of triangles and other short cycles [12]. 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. A model (inspired from [16]) that captures this in a natural way will be described in Section II. Roughly, the idea of the model 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. As we will show this model generalizes the standard configuration model to incorporate clustering and it is still possible to derive rigorously exact formulas for the analysis of contagions and diffusions on these networks. The model has the advantage to allow any arbitrary degree distribution: in particular, it can be applied to scale-free networks that have a power law degree distribution.

An important goal of network modeling is to investigate how the structure of the network affects the behavior of various types of dynamic processes on the network. The aim of this paper is to give a rigorous analysis of how clustering in a network affects the spread of an epidemic (we study both game-theoretical contagion and classical diffusion models). For the Reed-Frost epidemic, [12] studies the effect of clustering for a different model than ours and by heuristic means. Calculations indicate that the epidemic threshold should decrease as the clustering increases. This result has been recently rigorously proven in [2]. The analysis of a contagion process with threshold effect has not been done for their model. Another model of random graphs with positive clustering (different than the one we consider here) is introduced in [13], and heuristic results on both diffusion and contagion models are present in [6], for this model. Up to our knowledge, there is no rigorous analysis for the contagion model on a random graph

with clustering.

Our main contributions are

- the study of a tractable random graph model with tunable clustering (inspired from [16]). In Section II, we introduce it and compute the asymptotic degree distribution of the graph and the clustering coefficient,
- the analysis of a contagion process with threshold effect. In Section III, we derive the contagion threshold for our random graph model with clustering extending recent results of [8], and
- the analysis of a diffusion process with given probability of infection. In Section IV, we derive the minimal value for the probability of contagion in our random graph model with clustering such that a global diffusion is possible. This proves a heuristic result of [4].

In the following, we consider asymptotics as $n \to \infty$, and we denote by $\to_p$ the convergence in probability as $n \to \infty$. The abbreviation 'whp' ("with high probability") means with probability tending to 1 as $n \to \infty$, and we use the notation $o_p(n)$ in a standard way: $X = o_p(n)$ means that, for every $\varepsilon > 0$, $\mathbb{P}(X > \varepsilon n) \to 0$ as $n \to \infty$.

II. Random graph model and its basic properties

We first present the model for the social graph, then its asymptotic degree distribution, and finally its clustering coefficient.

A. Model

Let $n \in \mathbb{N}$ and let $\boldsymbol{d} = (d_i^{(n)})_{i=1}^n = (d_i)_1^n$ be a sequence of non-negative integers such that $\sum_i d_i$ is even. Let $G(n, \boldsymbol{d})$ be a graph chosen uniformly at random among all graphs with n vertices and degree sequence $\boldsymbol{d}$ (assuming there exists such a graph) [1].

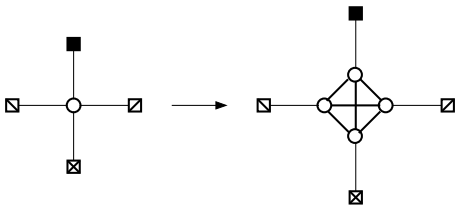
We will let $n \to \infty$ and assume that we are given $\boldsymbol{d}$ satisfying the following regularity conditions, see [9]:

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

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

If D_n is the degree of a vertex chosen uniformly at random among the n vertices of $G(n, \boldsymbol{d})$, and D a random variable with distribution $\boldsymbol{p}$, (i) is equivalent to the fact that $D_n \xrightarrow{d} D$ (convergence in distribution).

The model of random graphs $G(n, \boldsymbol{d})$ has the advantage to allow to handle arbitrary degree distributions. However, these graphs are 'locally tree-like', i.e. they contain no short loops in their structure. We now show that it is possible to generalize random graphs to incorporate clustering in simple fashion and still to derive rigorously exact formulas for diffusions and contagions. The model of random graphs is based on the model $G(n, \boldsymbol{d})$, but we 'add' clustering. The idea is to replace some vertices by a clique of size the degree in the original graph, i.e. a vertex of degree r in the original graph $G(n, \boldsymbol{d})$ is replaced by r vertices with all the $r(r-1)/2$ edges between them and each of them is connected to exactly one of the neighbors of the vertex in the original graph $G(n, \boldsymbol{d})$ as illustrated on the figure:



In order to be able to tune the clustering coefficient in the graph, we will not replace each vertex by a clique but rather do a probabilistic choice whether to replace a vertex or not. The resulting random graph will be denoted by $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = (\gamma_r)_{r=0}^\infty$ is a sequence such that: for all $r \geq 0$, $\gamma_r \in [0, 1]$ represents the probability that a vertex of degree r in $G(n, \boldsymbol{d})$ is replaced by a clique of size r in the new model $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$. More precisely, for each vertex $i \in \{1, ..., n\}$, let $X(i)$ be a Bernoulli variable with parameter γ_{d_i} (all Bernoulli variables being independent). We construct the random graph $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ the following way: start from $G(n, \boldsymbol{d})$ and, for each $i \in \{1, ..., n\}$, if $X(i) = 1$, replace i by a clique of size d_i where each vertex of the clique has exactly one neighbor outside the clique being a neighbor of i in the original graph. Note in particular that if we choose $\gamma_r = 0$ for all $r \geq 0$, then $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma}) = G(n, \boldsymbol{d})$, whereas for $\gamma_r = 1$ for all $r \geq 0$, all vertices in $G(n, \boldsymbol{d})$ have been replaced by cliques.

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

As we will see in the next subsection, the procedure described above introduces clustering at soon as $\gamma_r > 0$ for some r. It also modifies the degree distribution in the graph and we derive the new degree distribution here. Let $i \in \{1, ..., n\}$. If $X(i) = 1$, vertex i in $G(n, \boldsymbol{d})$ is replaced in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ by a clique of d_i vertices, all having degree d_i (indeed, each vertex of the clique has $d_i - 1$ edges linked with the other vertices of the clique, and one 'external' edge). So each vertex of degree r can either be replaced (with probability γ_r) by r vertices of degree r, or stays as a single vertex of degree r (with probability $1 - \gamma_r$). The following proposition gives the asymptotic degree distribution in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$.

Proposition 2. *We consider the model $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ for a sequence $\boldsymbol{d}$ satisfying Condition 1 with probability distribution $\boldsymbol{p} = (p_r)_{r=0}^\infty$, and clustering parameter $\boldsymbol{\gamma} = (\gamma_r)_{r=0}^\infty$. For all $r \geq 0$, let $\tilde{n}_r$ be the number of vertices with degree r in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$, and let $\tilde{n} = \sum_r \tilde{n}_r$ be the total number of vertices in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$. Then we have, as $n \to \infty$:*

$$\frac{\tilde{n}}{n} \xrightarrow{p} \tilde{\gamma} := \sum_{d \geq 0} \left[d\gamma_d + (1 - \gamma_d) \right] p_d$$

and, for all $r \geq 0$, the proportion of vertices with degree r in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ has the following limit, as $n \to \infty$:

$$\frac{\tilde{n}_r}{\tilde{n}} \xrightarrow{p} \tilde{p}_r := \frac{\left[r\gamma_r + (1 - \gamma_r) \right] p_r}{\tilde{\gamma}}.$$

Proof: Let $d \geq 0$, and let B_d be the number of vertices with degree d that are replaced by a clique. Then B_d follows a Binomial distribution with parameters (n_d, γ_d), where n_d is the number of vertices with degree d in $G(n, \boldsymbol{d})$. By Condition 1-*(i)*, we have: $n_d/n \to p_d$, so that the Law of Large Numbers implies: $B_d/n \to_p \gamma_d p_d$ (which is still true if $p_d = 0$).

Note that the number of vertices with degree d that are *not* replaced by a clique is $n_d - B_d$, so we can express the number $\tilde{n}$ of vertices in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ the following way:

$$\frac{\tilde{n}}{n} = \frac{1}{n}\sum_d dB_d + (n_d - B_d) \xrightarrow{p} \sum_d \left[d\gamma_d + (1-\gamma_d)\right] p_d = \tilde{\gamma}$$

which follows from the previous limits, and the uniform integrability of the random variables D_n (see Condition 1-*(iii)*).

In particular, this shows that the total number of vertices with degree r ($r \geq 0$) in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ is $rB_r + (n_r - B_r)$, so the proportion of vertices with degree r in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ is:

$$\frac{rB_r + (n_r - B_r)}{\tilde{n}} \xrightarrow{p} \frac{\left[r\gamma_r + (1-\gamma_r)\right] p_r}{\tilde{\gamma}} = \tilde{p}_r,$$

which concludes the proof. ∎

In other words, if $\tilde{D}_n$ is the degree of a vertex chosen uniformly at random in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$, then Proposition 2 implies that $\tilde{D}_n \xrightarrow{d} \tilde{D}$, where $\tilde{D}$ is a random variable with distribution $(\tilde{p}_r)_{r\geq 0}$.

In our definition of $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$, each vertex of degree 0 in $G(n, \boldsymbol{d})$ is removed from the graph with probability γ_0, and kept with probability $1 - \gamma_0$. We could have considered the case where each vertex of degree 0 is kept with probability 1. In that case, we have that $\tilde{p}_0 = p_0$ and the following value for $\tilde{\gamma}$: $\tilde{\gamma} = \sum_{d\geq 1}\left[d\gamma_d + (1-\gamma_d)\right] p_d + p_0$. To simplify notations, we keep the definition of the Proposition, but arguments are the same.

In the particular case where $\gamma_r = \gamma$ for all r, we have $\tilde{\gamma} = \gamma\lambda + 1 - \gamma$ and the mean degree of $\tilde{D}$ is then

$$\tilde{\lambda} = \mathbb{E}[\tilde{D}] = \frac{\gamma\mathbb{E}[D^2] + (1-\gamma)\lambda}{\gamma\lambda + 1 - \gamma},$$

which is a non-decreasing function of γ.

C. Clustering coefficient

The local clustering coefficient $C_v^{(n)}$ of a vertex v in a graph quantifies how close the vertex and its neighbors are to being a clique (complete graph) [18]. $C_v^{(n)}$ is defined to be the fraction of pairs of neighbors of v that are neighbors also of each other. More formally, let $\mathcal{N}_v$ be the set of neighbors of v (its cardinality $|\mathcal{N}_v| = d_v$ is the degree of v), and let P_v be the number of pairs $\{w, w'\} \subset \mathcal{N}_v$, $w \neq w'$, such that w and w' share an edge together. The total number of possible pairs is $d_v(d_v-1)/2$, so we define the local clustering coefficient of v as $C_v^{(n)} = P_v \cdot 2/[d_v(d_v-1)]$. The biased clustering coefficient for the whole network $C^{(n)}$ is defined as the average of the local clustering coefficient for each vertex: $C^{(n)} = \sum_v C_v^{(n)}/\tilde{n}$, where $\tilde{n}$ is the number of vertices in the graph, including those of degree one or zero. The local clustering coefficient of vertices with degree one or zero is null by definition, hence the clustering coefficient in the graph can be very low if the graph contains a lot of such vertices, even if other vertices are highly clustered. To overcome this, we can consider another definition of clustering coefficient (see the next paragraph).

a) Computation of the biased clustering coefficient:

Proposition 3. *We consider the model $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ for a sequence $\boldsymbol{d}$ satisfying Condition 1 with probability distribution $\boldsymbol{p} = (p_r)_{r=0}^{\infty}$. Then we have for the biased clustering coefficient of $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$:*

$$C^{(n)} \xrightarrow{p} C := \sum_{r\geq 3} p_r \frac{\gamma_r}{\tilde{\gamma}}(r-2),$$

where $\tilde{\gamma}$ is defined in Proposition 2.

The proof is given at the end of the subsection (together with the proof of Proposition 4).

In the particular case where $\gamma_r = \gamma \geq 0$ for all $r \geq 0$, we get for the asymptotic biased clustering coefficient:

$$C = \frac{\lambda - 2 + p_1 + 2p_0}{\lambda - 1 + \frac{1}{\gamma}}.$$

If $\gamma = 0$, there is no clustering and as γ increases, the biased clustering coefficient also increases to $1 - \frac{2-(p_1+2p_0)}{\lambda}$.

b) Another definition of clustering coefficient: We keep the notations of the previous paragraph: d_v is the degree of vertex v, and P_v is the number of pairs of neighbors of v that share an edge together. Then we define the clustering coefficient $C_2^{(n)}$ of the graph by:

$$C_2^{(n)} = \frac{2 \cdot \sum_v P_v}{\sum_v d_v(d_v-1)}.$$

It is the mean probability that three given vertices constitute a triangle conditional on that two of the three possible edges between them exist, and it corresponds to the notion of clustering studied in [2].

Proposition 4. *We consider the model $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ for a sequence $\boldsymbol{d}$ satisfying Condition 1 with probability distribution $\boldsymbol{p} = (p_r)_{r=0}^{\infty}$. Then we have for the clustering coefficient of $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$:*

$$C_2^{(n)} \xrightarrow{p} C_2 := \frac{\sum_{r\geq 2} r(r-1)(r-2)\gamma_r p_r}{\sum_{r\geq 2}((r-1)\gamma_r + 1)r(r-1)p_r}.$$

Proofs of Propositions 3 and 4: We have the following result, that follows from the fact that $G(n, \boldsymbol{d})$ converges locally to a tree, when $n \to \infty$:

Lemma 5. *Let $\overline{C}^{(n)}$ be the biased clustering coefficient in $G(n, \boldsymbol{d})$. Then we have: $\overline{C}^{(n)} \xrightarrow{p} 0$.*

The same result holds for the clustering coefficient $\overline{C}_2^{(n)}$ of the graph $G(n, \boldsymbol{d})$.

We say that a vertex in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ has *parent* $i \in \{1, ..., n\}$ if it belongs to a clique that replaces the vertex i of $G(n, \boldsymbol{d})$ (when $X(i) = 1$) or if it is i (when $X(i) = 0$).

We first consider a vertex v in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ whose parent i is such that $X(i) = 1$ (vertex i of $G(n, \boldsymbol{d})$ is replaced by a clique K in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$). In this case we can directly compute

the local clustering coefficient $C_v^{(n)}$. Indeed, vertex v has $d_i - 1$ neighbors inside K, that are all linked together (which gives $\frac{(d_i-1)(d_i-2)}{2}$ edges in total), and one neighbor v' outside K, which is not linked to the other neighbors of v (if it were the case, there would be several edges between i and the parent j of v', which is not the case in the simple graph $G(n, \boldsymbol{d})$). Hence

$$P_v = \frac{(d_i-1)(d_i-2)}{2} \quad \text{and} \quad C_v^{(n)} = \frac{2P_v}{d_i(d_i-1)} = \frac{d_i-2}{d_i},$$

provided that $d_i \geq 2$. If $d_i \in \{0, 1\}$, then $C_v^{(n)} = 0$.

We first prove Proposition 3. Since there are d_i such vertices inside a clique, the contribution of clique K in the total clustering $C^{(n)} = \sum_v C_v^{(n)}/\tilde{n}$ is equal to $d_i C_v^{(n)}/\tilde{n} = (d_i - 2)/\tilde{n}$. This leads to the following:

$$\frac{\tilde{n}}{n} C^{(n)} = \frac{1}{n}\sum_{d\geq 2}(d-2)B_d + \frac{1}{n}\sum_{i:X(i)=0} C_i^{(n)}$$

where B_d is the number of vertices with degree d that are replaced by a clique, as in the proof of Proposition 2. Using that $B_d/n \to_p \gamma_d p_d$, and that $\sum_{i:X(i)=0} C_i^{(n)}/n \to_p 0$ (as a consequence of Lemma 5), we obtain: $\frac{\tilde{n}}{n} C^{(n)} \xrightarrow{p} \sum_{d\geq 3}(d-2)\gamma_d p_d$. Proposition 3 follows, applying Proposition 2.

The end of the proof for Proposition 4 is similar, and follows from the fact that:

$$2\sum_v P_v/n \to_p \sum_d d(d-1)(d-2)\gamma_d p_d,$$
$$\sum_v d_v(d_v-1)/n \to_p \sum_d d(d-1)[d\gamma_d + (1-\gamma_d)]p_d.$$

■

In the next paragraph, we use the second definition of clustering coefficient, but results are true for both.

D. Link between clustering coefficient and mean degree in the graph $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$

Both asymptotic clustering coefficient C_2 (Proposition 4) and asymptotic degree distribution $(\tilde{p}_r)_{r\geq 0}$ (Proposition 2) depend on the following parameters:

- the asymptotic degree distribution $\boldsymbol{p}$ of the original graph $G(n, \boldsymbol{d})$,
- the clustering parameter $\boldsymbol{\gamma}$.

We will see how the clustering coefficient C_2 and the mean degree $\tilde{\lambda} = \sum_r r\tilde{p}_r$ vary when $\boldsymbol{p}$ and $\boldsymbol{\gamma}$ vary. We focus on the case when $\gamma_r = \gamma$ for all $r \geq 0$, *i.e.* each vertex in $G(n, \boldsymbol{d})$ is replaced by a clique with probability γ. In addition, we impose some conditions on the probability distribution $\boldsymbol{p}$.

Indeed, in figures 1 and 2, we assume that the initial graph $G(n, \boldsymbol{d})$ has power law degree distribution with exponential cutoff: there exists a power $\tau > 0$ and a cutoff $\kappa > 0$ such that, for all $r \geq 1$, $p_r = c(\tau, \kappa) \cdot r^{-\tau} e^{-r/\kappa}$, where $c(\tau, \kappa) = 1/(\sum_s s^{-\tau} e^{-s/\kappa})$ is a normalizing constant. This cutoff κ allows Condition 1 to be satisfied for any power $\tau > 0$: in figures, we take $\kappa = 50$. If we compute (using Proposition 2) the asymptotic degree distribution in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ when each vertex is replaced by a clique ($\gamma = 1$), we obtain a power law distribution with parameter $\tau - 1$: $\tilde{p}_r \propto r^{-(\tau-1)} e^{-r/\kappa}$. More generally, for any given value of γ, the degree distribution in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ is a linear combination between a power law of parameter τ, and a power law of parameter $\tau - 1$.

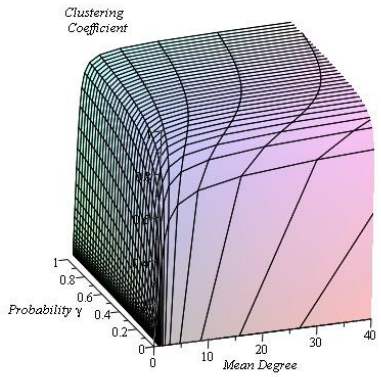


Fig. 1. Correlation between the clustering coefficient C_2 and the mean degree $\tilde{\lambda}$ in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ (when the initial graph $G(n, \boldsymbol{d})$ has power law degree distribution $p_r \propto r^{-\tau} e^{-r/50}$, and varying parameters are τ and γ)

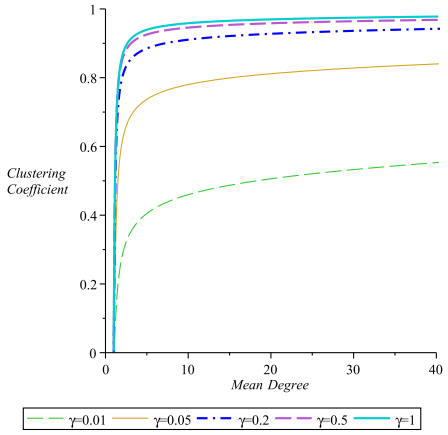


Fig. 2. Evolution of the clustering coefficient C_2 with respect to the mean degree $\tilde{\lambda}$ in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ (when the initial graph $G(n, \boldsymbol{d})$ has power law degree distribution $p_r \propto r^{-\tau} e^{-r/50}$)

Once we impose this form for the probability distribution $\boldsymbol{p}$, we are left with two degrees of freedom: the power τ and the probability γ. In figure 1, we make these two parameters vary, and plot the correlation between the clustering coefficient C_2, the mean degree $\tilde{\lambda}$ and the probability γ.

Figure 2 represents several slices of figure 1, for different values of the clique probability γ. Increasing the mean degree in the graph $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ also increases the clustering. With our graph model, we are not able to reach a clustering coefficient of 1, especially for low values of the mean degree $\tilde{\lambda}$. Yet the maximal clustering coefficient we can obtain (for $\gamma = 1$) is greater than 0.9 as soon as the mean degree $\tilde{\lambda}$ is greater than 2.5.

Now we are interested in the evolution of the clustering coefficient C_2 with respect to γ, when the mean degree $\tilde{\lambda}$ is fixed (which corresponds to other slices of figure 1). In order

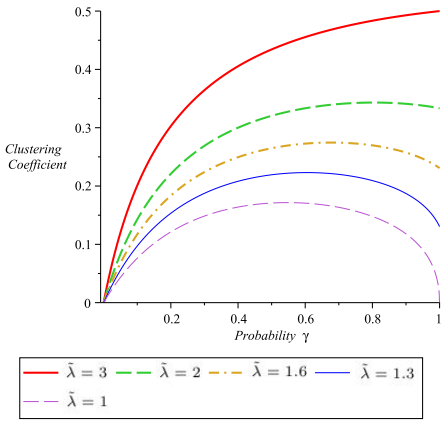


Fig. 3. Evolution of the clustering coefficient C_2 with respect to the clique probability γ, when the mean degree $\tilde{\lambda}$ in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ is fixed (for initial graph $G(n, \boldsymbol{d})$ distributed as Erdős-Renyi)

to keep $\tilde{\lambda}$ fixed when γ varies, we need to adjust the value of the power τ: this changes both the probability distribution $\boldsymbol{p}$ and the asymptotic degree distribution $\tilde{\boldsymbol{p}}$ (whose mean is kept fixed).

For technical reasons, it will be more convenient to work with a Poisson distribution with mean $\lambda > 0$ for the probability distribution $\boldsymbol{p}$: for all $r \geq 0$, $p_r = e^{-\lambda}\lambda^r/r!$, that is to say $G(n, \boldsymbol{d})$ is distributed as an Erdős-Renyi graph. If $\gamma = 1$, the asymptotic degree distribution in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ is a Poisson variable (with parameter λ) shifted by 1, so the mean degree is $\tilde{\lambda} = \lambda + 1$. In that case, the degrees of freedom are parameters λ and γ. In order to keep $\tilde{\lambda}$ fixed when γ varies, we have to adjust the value of λ. If the mean degree $\tilde{\lambda}$ in $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ is fixed and high enough, the clustering coefficient C_2 increases with γ (figure 3), but, for low values of $\tilde{\lambda}$, the clustering coefficient C_2 is not a non-decreasing function of γ (distributions $\boldsymbol{p}$ and $\tilde{\boldsymbol{p}}$ vary).

The advantage of this model is that it allows to consider any degree distribution, contrary to random intersection graphs [2] that are restricted to the Poisson distribution. Yet we are limited by the correlation between the clustering coefficient and the mean degree in the graph $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$: in [2], the clustering coefficient can vary between 0 and 1, even for low values of the mean degree in the graph.

III. Contagion threshold for random graphs with clustering

A. Contagion model

Motivated by the game-theoretic contagion model proposed by Morris [10] and described in the introduction, we now describe formally our model of contagion on any finite graph G. The progressive dynamic of the diffusion 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 $t = 1, 2, 3, \ldots$. At a given time t, any inactive node i becomes active if its number of active neighbors is at least $\lfloor qd_i \rfloor + 1$. This in turn may cause other nodes to become active. It is easy to see that the final set of active nodes (after n time steps if the network is of size n) only depends on the initial set S (and not on the order of the activations) and can be obtained as follows: set $Y_i = \mathbb{1}(i \in S)$ for all i. Then as long as there exists i such that $\sum_{j\sim i} Y_j > qd_i$, set $Y_i = 1$, where $j \sim i$ means that i and j share an edge in G. When this algorithm finishes, the final state of node i is represented by Y_i: $Y_i = 1$ if node i is active and $Y_i = 0$ otherwise.

We see that the lower q is, the easier the diffusion spreads. In [10], the contagion threshold of a connected infinite network is defined as the maximum threshold q_c at which a finite set of initial adopters can cause a complete cascade, i.e. the resulting cascade of adoptions of B eventually causes every node to switch from A to B. In this section, we restrict ourselves to the model where the initial adopters are forced to play B forever. In this case, the contagion is monotone and the number of nodes playing B is non-decreasing. We say that this case corresponds to the permanent adoption model: a player playing B will never play A again.

B. Phase transition for the contagion

We now compute the contagion threshold for a sequence of random networks. Since a random network is finite and not necessarily connected, we first need to adapt the definition of contagion threshold to our context as was done in [8]. For a graph $G = (V, E)$ and a parameter q, we consider the largest connected component of the induced subgraph in which we keep only vertices of degree strictly less than q^{-1}. We call the vertices in this component pivotal players: if only one pivotal player switches from A to B then the whole set of pivotal players will eventually switch to B in the permanent adoption model. For a player $v \in V$, we denote by $C(v, q)$ the final number of players B in the permanent adoption model with parameter q, when the initial state consists of only v playing B, all other players playing A. Informally, we say that $C(v, q)$ is the size of the cascade induced by player v.

Theorem 6. *Consider the random graph $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ for a sequence $\boldsymbol{d}$ satisfying Condition 1 with probability distribution $\boldsymbol{p} = (p_r)_{r=0}^{\infty}$, and define*

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

Let $\tilde{\mathcal{P}}^{(n)}$ be the set of pivotal players of $\tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$.

(i) *If $q < q_c$, then there is a unique $\xi \in (0, 1)$ such that*

$$\sum_{d<q^{-1}} dp_d(1 - \xi^{d-1}) = \lambda(1 - \xi)$$

and we have:

$$|\tilde{\mathcal{P}}^{(n)}|/\tilde{n} \xrightarrow{p} \sum_{d<q^{-1}} \frac{[d\gamma_d + (1-\gamma_d)]\, p_d}{\tilde{\gamma}}(1 - \xi^d) > 0,$$

where $\tilde{\gamma}$ is defined in Proposition 2. Moreover, for any $u \in \tilde{\mathcal{P}}$, we have whp

$$\liminf \frac{C(u, q)}{\tilde{n}} \geq \lim \frac{|\tilde{\mathcal{P}}^{(n)}|}{\tilde{n}} > 0$$

(ii) *If $q > q_c$, for an uniformly chosen player u, we have $C(u,q) = o_p(\tilde{n})$. The same result holds if $o(n)$ players are chosen uniformly at random.*

When $\gamma_r = 0$ for all $r \geq 0$, we recover the result of [8]. If $\gamma_r = 1$ for all $r \geq 1$ (which means that we systematically replace each vertex i of degree d_i by a clique of size d_i) we can even be more precise in the case $q < q_c$: we have in this case, $C(u,q) = |\tilde{\mathcal{P}}^{(n)}|$ for any $u \in \tilde{\mathcal{P}}^{(n)}$ since in this case, the contagion starting from a pivotal player will propagate to the pivotal players only.

From these results, we see that the impact of clustering is different for low values of the mean degree and for high values of the mean degree. In the low values regime, as the clustering increases, the contagion threshold decreases whereas in the high values regime, the opposite happens. We see that in the low values regime for the mean degree, the clustering makes the contagion more difficult whereas it 'helps' the contagion in the high values regime.

IV. Diffusion threshold for random graphs with clustering

A. Diffusion model

In this section, we study a simple diffusion model depending on a parameter $\pi \in [0,1]$ which can be described in term of a bond percolation process in a general graph G. Randomly delete each edge with probability $1-\pi$ independently of all other edges. denote by G_π the resulting graph. Then any active node will activate all nodes in its component in G_π. As in previous section, we will derive conditions under which a single starting active node can activate a large fraction of the population in $G = \tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$. This problem corresponds to the existence of a 'giant component' in the random graph obtained after bond percolation. Note that this model of diffusion corresponds to a simple epidemics with probability of contagion given by the parameter $\pi \in [0,1]$.

B. Phase transition for the diffusion

In order to state our result, we first need to recall some basic results about random graphs with small order. For $d \in \mathbb{N}$, let K_d be the complete graph on d vertices denoted $\{1, \ldots, d\}$, with $d(d-1)/2$ edges. For $\pi \in [0,1]$, we denote by $K_d(\pi)$ the random graph obtained from K_d after bond percolation with parameter π, i.e. each edge of K_d is kept independently of the others with probability π, otherwise it is removed.

We need to compute the probability that the component in $K_d(\pi)$ containing vertex 1 has k vertices denoted by $f(d,k,\pi)$. Note that $f(d,d,\pi)$ is simply the probability that $K_d(\pi)$ is connected and has been computed in [3]. Indeed simple computations show that we have the simple recurrence relation

$$\begin{aligned} f(d,d,\pi) &= 1 - \sum_{k=1}^{d-1} \binom{d-1}{k-1} f(k,k,\pi)(1-\pi)^{k(d-k)}, \\ f(d,k,\pi) &= \binom{d-1}{k-1} f(k,k,\pi)(1-\pi)^{k(d-k)}, \end{aligned} \tag{1}$$

for any $k \leq d$.

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

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

where f is defined in (1).

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

Theorem 7. *Consider the random graph $G = \tilde{G}(n, \boldsymbol{d}, \boldsymbol{\gamma})$ for a sequence $\boldsymbol{d}$ satisfying Condition 1 with probability distribution $\boldsymbol{p} = (p_r)_{r=0}^{\infty}$. Let D^* be a random variable with distribution p^*_r given by $p^*_{r-1} = \frac{rp_r}{\lambda}$ for all $r \geq 1$. We define π_c as the solution of the equation:*

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

- *if $\pi > \pi_c$, we have whp that $\liminf \frac{C(u,\pi)}{\tilde{n}} > 0$, i.e. there exists a 'giant component' in the percolated graph G_π.*
- *if $\pi < \pi_c$, for an uniformly chosen player u, we have $C(u,\pi) = o_p(\tilde{n})$. The same result holds if $o(n)$ players are chosen uniformly at random, i.e. there is no 'giant component' in the percolated graph G_π.*

Note that in the particular case where $\gamma_r = 0$ for all r, we have $\mathcal{K}(d,\pi,0) = d$ so that we get $\pi_c = \frac{\mathbb{E}[D]}{\mathbb{E}[D(D-1)]}$ where D is the typical degree in the random graph and we recover a standard result in the random graphs literature (see Theorem 3.9 in [7]).

We can guess the value of the diffusion threshold π_c using a branching process approximation. Indeed the random graph $G(n, \boldsymbol{d})$ can be approximated by a branching process in which each node (except the root) has a number of offspring distributed as D^*. The degree of a node v in the corresponding random tree is thus distributed as D^*+1. Let us assume $D^*+1 = d$. If we replace v by a clique K of size d with probability γ_d, and delete independently each edge inside the clique with probability $1-\pi$, then the probability that the component of v inside K contains k vertices is given by $f(d,k,\pi)$. Hence the probability that v is linked to k vertices is: $(1-\gamma_d)\mathbb{1}(d=k) + \gamma_d f(d,k,\pi) = \mathbb{P}\left(\mathcal{K}(D^*+1,\pi,\boldsymbol{\gamma}) = k\right)$. The new distribution of offspring is thus $\mathcal{K}(D^*+1,\pi,\boldsymbol{\gamma}) - 1$. Finally, we remove each edge with probability π, which gives $\pi\mathbb{E}\left[\mathcal{K}(D^*+1,\pi,\boldsymbol{\gamma}) - 1\right]$ for the expected number of offspring.

For regular graphs, we obtain that the diffusion threshold increases as the clustering increases, as it was already observed in [5].

V. Further remarks and conclusions

If the characteristic parameter of the epidemic is above some threshold for the diffusion model, or under a certain threshold for the contagion model, then a global cascade is possible, starting from a single infected individual. An interesting question would be to study the effect of clustering on the cascade size, especially for the contagion model. The cascade size for the diffusion model is derived by heuristic means in [4].

The diffusion threshold increases when the clustering increases (for random regular graphs), which makes the diffusion

more difficult to spread. The effect of clustering on the contagion threshold depends on the value of the mean degree in the graph: for low values of the mean degree, we observe that clustering inhibits the contagion, and the contrary happens in the high values regime. One can wonder what is the effect of clustering on contagion threshold when degree distribution and degree-degree correlations are fixed: note that the simple consideration of random regular graphs does not provide significant information here, since increasing clustering in a random regular graph (adding cliques) does not change the contagion threshold with our graph model.

REFERENCES

[1] B. Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.

[2] T. Britton, M. Deijfen, A. N. Lagerås, and M. Lindholm. Epidemics on random graphs with tunable clustering. *J. Appl. Probab.*, 45(3):743–756, 2008.

[3] E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30:1141–1144, 1959.

[4] J. P. Gleeson and S. Melnik. Analytical results for bond percolation and k-core sizes on clustered networks. *Physical Review E*, 80, 2009.

[5] J. P. Gleeson, S. Melnik, and A. Hackett. How clustering affects the bond percolation threshold in complex networks. *Physical Review E*, 81, 2010.

[6] A. Hackett, S. Melnik, and J. P. Gleeson. Cascades on a class of clustered random networks. *Physical Review E*, 83, 2011.

[7] S. Janson. On percolation in random graphs with given vertex degrees. *Electron. J. Probab.*, 14:no. 5, 87–118, 2009.

[8] M. Lelarge. Diffusion and cascading behavior in random networks. *under revision for Games and Economic Behavior*, arxiv/1012.2062, 2010.

[9] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures Algorithms*, 6(2-3):161–179, 1995.

[10] S. Morris. Contagion. *Rev. Econom. Stud.*, 67(1):57–78, 2000.

[11] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):016128, Jul 2002.

[12] M. E. J. Newman. Properties of highly clustered networks. *Phys. Rev. E*, 68(2):026121, Aug 2003.

[13] M. E. J. Newman. Random graphs with clustering. *Phys. Rev. Lett.*, 2009.

[14] M. E. J. Newman, A. L. Barabási, and D. J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[15] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, New York, NY, USA, 2004.

[16] P. Trapman. On analytical approaches to epidemics on networks. *Theoretical Population Biology*, 71(2):160–173, 2007.

[17] F. Vega-Redondo. *Complex social networks*, volume 44 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 2007.

[18] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.