Pioneers of Influence Propagation in Social Networks

Kumar Gaurav[1], Bartłomiej Błaszczyszyn[2], and Paul Holger Keeler[3]

[1] UPMC/Inria/ENS, 23 av. d'Italie 75214 Paris, France
Kumar.Gaurav@inria.fr
[2] Inria/ENS, 23 av. d'Italie 75214 Paris, France
Bartek.Blaszczyszyn@ens.fr
[3] Inria/ENS, 23 av. d'Italie 75214 Paris, France
Holger.Keeler@inria.fr

Abstract. In this paper, we present a diffusion model developed by enriching the generalized random graph (a.k.a. configuration model), motivated by the phenomenon of viral marketing in social networks. The main results on this model are rigorously proved in [3], and in this paper we focus on applications. Specifically, we consider random networks having Poisson and Power Law degree distributions where the nodes are assumed to have varying attitudes towards influence propagation, which we encode in the model by their transmitter degrees. We link a condition involving total degree and transmitter degree distributions to the effectiveness of a marketing campaign. This suggests a novel approach to decision-making by a firm in the context of viral marketing which does not depend on the detailed information of the network structure.

1 Introduction

The penetration of internet and the emergence of huge online social networks in the last decade has lead to a decline of the conventional channels of communication and consequently, marketing through them. This has given the firms an opportunity to reach a large subset of their customers through innovative viral marketing campaigns. But the wild uncertainty inherent in whether a marketing campaign goes viral or not, makes it markedly different from conventional marketing and calls for a fundamentally different approach to decision-making.

1.1 Results

In this paper, we introduce a generalized diffusion dynamic on configuration model. Configuration model, while lacking the *community structure* of real-world social networks, approximates the degree distribution of these networks quite well. The diffusion dynamic that we consider can be intuitively described in the following way: an influenced individual in the network influences a random subset of its neighbours, the distribution of which depends on the effectiveness of the marketing campaign.

Z. Cai et al. (Eds.): COCOON 2014, LNCS 8591, pp. 626–636, 2014.

We illustrate large-network-limit results on this model, rigorously proved in [3]. We present a condition involving the total degree and transmitter degree distribution of a uniformly chosen node which, if satisfied, will allow, with a non-negligible probability, the campaign to go viral when started from this particular node. Given this condition, we present an estimate of the fraction of the population that is reached when the campaign does go viral. We then state that under the same condition, the fraction of good pioneers in the network, i.e., the individuals who if targeted initially will lead the campaign to go viral, is non-negligible as well, and we give an estimate of this fraction. We analyze in detail the process of influence propagation on configuration model having two types of degree-distribution: Poisson and Power Law. Three examples illustrating the dynamic of influence propagation on these two networks are considered: (1) Bernoulli transmissions; (2) Node percolation; (3) Coupon-collector transmissions.

Based on the above analysis, we suggest what statistical data a firm should collect from the pioneers of its marketing campaign, and based on these, how to estimate the effectiveness of the campaign and make a cost-benefit analysis.

1.2 Related Work

The diffusion-on-random-graph models have been previously studied in the context of the spread of epidemics in population ([1]), and more recently, to understand the propagation of social and economic behavior through a social network ([7], [2], [9]).

In the context of viral marketing, one approach is to use the detailed information about the network structure and the past instances of influence propagation to come up with a predictor of the most influential individuals who should be targeted for future campaigns ([6]). Another approach is to take into account the current campaign's effectiveness based on the detailed temporal and structural information regarding the ongoing diffusion in the network ([4]). Our analysis follows the latter approach, but differs in that we require much less information regarding the network and the ongoing diffusion in it.

2 Model and Theoretical Claims

In this section, we introduce our model and informally describe the results which are rigorously proved in [3].

2.1 Model

Consider that the only information available to you about an online social network is the number of friends that a subset of network members have. We will work with a uniform random network which agrees with the statistics that you can obtain from the available information. Such a uniform random network is obtained by constructing what is known as *configuration model* (CM); cf [8]. This

random network is realized by attaching half-edges to each vertex corresponding to its degree (which represents here, the number of friends) and then uniformly pair-wise matching them to create edges. We assume this model of the social network throughout the paper and will use interchangeably the terms "social graph" and "random network" meaning precisely the CM. We call the vertices of this graph "nodes" or "users" and graph neighbours "friends".

We consider a marketing campaign started from some initial target called *pioneer* in this network. A person influences a subset of its friends who further propagate the campaign in the same manner. The number of friends that a person influences depends on a particular campaign. To model this dynamic, we enhance the configuration model by partitioning the half-edges into *transmitter* half-edges, those through which the influence can flow and *receiver* half-edges which can only receive influence. So, if a person A influences his friend B in the network, then in our representation, A has a transmitter half-edge matched to the transmitter or receiver half-edge of B.

Let D and $D^{(t)}$ denote the random variables corresponding to the empirically observed distributions of total degree and transmitter degree respectively. For notational convenience, we will interchangeably use random variables and their distributions to mean the same thing. Empirical receiver degree distribution, $D^{(r)}$, is $D - D^{(t)}$. Then we have the following large-network-limit results, rigorously proved in [3], but only informally stated here.

2.2 Theoretical Claims

Claim 1. *Starting from a randomly selected pioneer, the campaign can go viral, i.e., reach a strictly positive fraction of the population, with a strictly positive probability if and only if*

$$\mathbb{E}[D^{(t)}D] > \mathbb{E}[D^{(t)} + D]. \tag{1}$$

Note that $\mathbb{E}[D^{(t)}D] > \mathbb{E}[D^{(t)} + D]$ implies

$$\mathbb{E}[D(D-2)] > 0 \tag{2}$$

and recall that this latter condition is necessary and sufficient [1] for the existence of a (unique) connected component of the underlying social graph, called *big component*, encompassing a strictly positive fraction of its population; cf [5]. Obviously, our campaign can go viral only within this big component.

Call *good pioneers* the pioneers from which the campaign can go viral.

Claim 2. *If (1) is satisfied, then the population reached is, more or less, the same, irrespective of the good pioneer chosen initially.*

Let C^* denote the population reached by the campaign when started from a good pioneer and $\overline{C}^*$ the set of good pioneers.

[1] Under a few additional technical assumptions, as $0 < \mathbb{E}[D] < \infty$, $\mathbb{P}\{D = 1\} > 0$, which we tacitly assume throughout the paper.

Claim 3. *If (1) is satisfied, then the set of good pioneers* $\overline{C}^*$ *also forms a strictly positive fraction of the population.*

The next claim gives the estimates on the size of C^* and $\overline{C}^*$. Let

$$H(x) := \mathbb{E}[D]x^2 - \mathbb{E}[D^{(r)}]x - \mathbb{E}[D^{(t)}x^D] \tag{3}$$

and

$$\overline{H}(x) := \mathbb{E}[D]x^2 - \mathbb{E}[D^{(t)}x^{D^{(t)}}] - \mathbb{E}[D^{(r)}x^{D^{(t)}}]x. \tag{4}$$

If condition (1) is satisfied, then $H(x)$ and $\overline{H}(x)$ have unique zeros in $(0,1)$. Call them ξ and $\overline{\xi}$ respectively. Denote also by $G_D(x) = \mathbb{E}[x^D]$ and $G_{D^{(t)}}(x) = \mathbb{E}[x^{D^{(t)}}]$, the probability generating function (pgf) of D and $D^{(t)}$, respectively.

Claim 4. *If (1) is satisfied and* n *denotes the size of network population, then for* n *large,*

$$\frac{|C^*|}{n} \approx 1 - G_D(\xi) =: \alpha > 0 \tag{5}$$

and

$$\frac{\left|\overline{C}^*\right|}{n} \approx 1 - G_{D^{(t)}}(\bar{\xi}) =: \overline{\alpha} > 0. \tag{6}$$

Note that $\overline{\alpha}$ can be interpreted as the probability that the campaign goes viral when started from a randomly chosen pioneer.

See [3] for formal statements and proofs of the above claims. Recall also from [5] that under assumption (2), the size, $|C_0|$, of the big network component, C_0, satisfies for n large, $\frac{|C_0|}{n} \approx 1 - G_D(\xi_0) =: \alpha_0 > 0$, where ξ_0 is the unique zero of $H_0(x) := \mathbb{E}[D]x^2 - xG'_D(x)$ in $(0,1)$, with $G'_D(x)$ denoting the derivative of the pgf of D.

3 Examples

Let us consider the results of Section 2 in the context of a few illustrative network examples.

3.1 Bernoulli Transmissions

Let us assume some arbitrary distribution of the degree D satisfying (2) (to guarantee the existence of the big component of the social graph). Suppose that each user decides independently for each of its friends with probability $p \in [0,1]$ whether to transmit the influence to him or not. We call this model, *CM with Bernoulli transmissions*, and p, the *transmission probability*. Note that given the total degree D, the transmitter degree D^t is Binomial(D,p) random variable.

Proposition 1. *In the CM with a general degree distribution D satisfying (2) and Bernoulli transmissions, the campaign can go viral if and only if the transmission probability p satisfies*

$$p > \frac{\mathbb{E}[D]}{\mathbb{E}[D^2] - \mathbb{E}[D]}\,. \tag{7}$$

In this latter case, the fraction of the influenced population and the fraction of good pioneers are approximately equal to each other, i.e., $|C^*|/n \approx |\bar{C}^*|/n =: \alpha$, for large n, and satisfy

$$\alpha = 1 - G_D(\xi)\,, \tag{8}$$

where ξ is the unique zero of the function $\mathbb{E}[D]((x-1)/p+1) - G'_D(x)$ in $(0,1)$.

Proof. Bernoulli transmissions with (3) and (4) imply $H(x) = \mathbb{E}[D]x^2 - (1-p)\mathbb{E}[D]x - pxG'_D(x)$ and $\overline{H}(x) = \mathbb{E}[D]x^2 - G'_D(1-p(1-x))$. Moreover $\overline{G}_{D^{(t)}}(x) = G_D(1-p(1-x))$. Dividing $H(x)$ by px and substituting $y := 1-p(1-x)$ in $\overline{H}(x)$ and $\overline{G}_{D^{(t)}}(x)$ completes the proof.

Consider two specific network degree examples.

Example 1 (Poisson degree). When D has Poisson distribution of parameter λ (in which case the CM is asymptotically equivalent to the Erdös-Rényi model) the condition (7) reduces to $\lambda p > 1$ and the fraction of the influenced population and good pioneers (8) is equal to $\alpha = (1-\xi)/p$, where ξ is the unique zero of the function $(x-1)/p + 1 - \exp(\lambda(x-1)$ in $(0,1)$.

More commonly observed degree-distributions in social networks have power-law tails.

Example 2 (Power-Law ("zipf") degree). Assume D having distribution $\mathbb{P}\{D = k\} = k^{-\beta}/\zeta(\beta) \quad k = 1, 2, \ldots$, with $\beta > 2$, where $\zeta(\beta)$ is the zeta function. Recall that the pgf of D is equal to $G_D(x) = \mathrm{Li}_\beta(x)/\zeta(\beta)$, where $\mathrm{Li}_\beta(x) = \sum_{k=1}^{\infty} k^{-\beta}x^k$ is the so-called poly-logarithmic function. Condition (2) for the existence of the big component is equivalent to $\zeta(\beta-2) - 2\zeta(\beta-1) > 0$, which is approximately $\beta < 3.48$. Condition (7) reduces to $p > \zeta(\beta-1)/(\zeta(\beta-2) - \zeta(\beta-1))$ and the fraction of the influenced population and good pioneers (8) is equal to $\alpha = 1 - \mathrm{Li}_\beta(\xi)$, where ξ is the unique zero of the function $x\zeta(\beta-1)((x-1)/p+1) - \mathrm{Li}_{\beta-1}(x)$ in $(0,1)$.

Recall from Proposition 1, that Bernoulli transmissions lead to the model where the fraction of the influenced population and the fraction of good pioneers are asymptotically equal to each other. In what follows we present two scenarios where the set of good pioneers and the influenced population have different size.

3.2 Enthusiastic and Apathetic Users or Node Percolation

Consider CM with a general degree distribution D satisfying (1), whose nodes either transmit the influence to all their friends (these are "enthusiastic" nodes)

or do not transmit to any of their friends ("apathetic" ones). Let p denote the fraction of nodes in the network which are enthusiastic. Note that this model corresponds to the *node-percolation* [2] on the CM. Thus, in this model, given D, $D^{(t)} = D$ with probability p and $D^{(t)} = 0$ with probability $1 - p$.

Proposition 2. *Consider node-percolation on the CM with a general degree distribution D satisfying (2). The campaign can go viral if and only if the fraction p of enthusiastic users satisfies condition (1); the same as for the Bernoulli model. Moreover, in this case, the fraction α of reached population is also the same as in the network with Bernoulli transmissions, i.e., equal to (8) with ξ as in Claim 1. However, the fraction $\overline{\alpha}$ of good pioneers is equal to $\overline{\alpha} = p\alpha$.*

The proof follows easily from the general results of Section 2.2. Note that the *campaign on the network with enthusiastic and apathetic users can reach the same population as in the Bernoulli transmissions, however there are less good pioneers.*

3.3 Absentminded Users or Coupon-Collector Transmissions

Consider again CM with a general degree distribution D satisfying (1). Suppose that each user is willing (or allowed) to transmit K messages of influence. In this regard, it randomly selects, K times, one of his friends *with replacement* (as if it forgets its previous choices). An equivalent dynamic of the influence propagation can be formulated as follows: every influenced user, at all times, keeps choosing one of its friends uniformly at random and transmits the influence to him; it stops forwarding the influence after K transmissions.

In this model, the transmission degree, $D^{(t)}$, corresponds to the number of collected coupons in the classical coupon collector problem with the number of coupons being the vertex degree, D, and the number of trials, K. The conditional distribution of $D^{(t)}$, given D, can be expressed as follows: $\mathbb{P}\{\, D^{(t)} = k \,|\, D \,\} = \frac{D!}{(D-k)!D^K} \left\{ {K \atop k} \right\}$, where $\left\{ {K \atop k} \right\} = 1/k! \sum_{i=0}^{K} (-1)^i \binom{k}{i} (k-i)^K$ is the Stirling number of the second kind.

Calculating the pgf for this distribution is tedious and we do not present analytical results regarding this model but only simulations and estimation. As we shall see in Section 3.4, in this model, *the influenced population is smaller than the population of good pioneers.*

3.4 Numerical Examples

We will present now a few numerical examples of networks and diffusion models presented above.

[2] Different than edge-percolation.

Simulations. In all our examples, we simulate the enhanced configuration model on $N = 1000$ nodes assuming some particular node degree distribution, D, and influence propagation mechanism modeled by the conditional distribution of the transmitter degree, $D^{(t)}$. More precisely, we sample the individual node degrees and transmitter degrees $(D_i, D_i^{(t)})$ $i = 1 \dots N$ independently from the joint distribution of $(D, D^{(t)})$ and use these values to construct an instance of our enhanced CM by uniform pairwise matching of the half-edges. We calculate the relative size of the influenced population and the set of good pioneers through the exploration of the influenced components for all nodes. [3]

Estimation. We adopt also the following "semi-analytic" approach: Using the sample $(D_i, D_i^{(t)})$, $i = 1, \dots, N$ used to construct the CM, we consider estimators, $\hat{G}_D(x) := \frac{1}{N}\sum_{i=1}^{N} x^{D_i}$, $\hat{G}_D^{(t)}(x) := \frac{1}{N}\sum_{i=1}^{N} x^{D_i^{(t)}}$, $\hat{H}(x) := \frac{1}{N}\sum_{i=1}^{N}\Big(D_i x^2 - (D_i - D_i^{(t)})x - D_i^{(t)} x^{D_i}\Big)$, $\hat{\overline{H}}(x) := \sum_{i=1}^{N}\Big(D_i x^2 - D_i^{(t)} x^{D_i^{(t)}} - (D_i - D_i^{(t)})x^{D_i^{(t)}+1}\Big)$, of the functions, $G_D(x)$, $G_{D^{(t)}}(x)$, $H(x)$ and $\overline{H}(x)$, respectively. We calculate estimators $\hat{\alpha}$ and $\hat{\overline{\alpha}}$ of the fraction of the influenced population, α, and of good pioneers, $\overline{\alpha}$, using Claim 5 and the estimated functions, $\hat{G}_D(x)$, $\hat{G}_{D^{(t)}}(x)$, $\hat{H}(x)$ and $\hat{\overline{H}}(x)$. (That is, we find numerically, zeros, $\hat{\xi}$ and $\hat{\overline{\xi}}$ of $\hat{H}(x)$ and $\hat{\overline{H}}(x)$, respectively, and plug them into (5) and (6), with $\hat{G}_D(x)$ and $\hat{G}_{D^{(t)}}(x)$ replacing $G_D(x)$ and $G_{D^{(t)}}(x)$.)

Note that in the semi-analytic approach, we do not need to know/construct the realization of the underlying model. This observation is a basis of a *campaign evaluation method* that we propose in Section 4. In fact, in reality one usually does not have the complete insight into the network structure and needs to rely on statistics collected from the initially contacted pioneers.

Analytic Evaluation. Finally, for all models, except the "coupon-collector" one of Section 3.3, we calculate numerically, the values of α and $\overline{\alpha}$ using the explicit forms of all the involved functions. (For the coupon-collector model, we obtained the "true" values of α and $\overline{\alpha}$ from a sample of $(D_i, D_i^{(t)})$ of a larger size N.)

When comparing these analytic solutions to the simulation and semi-analytic estimates, we see that in some cases, $N = 1000$ is not big enough to match the theoretical values. One can easily consider larger samples, however, we decided to stay with $N = 1000$ to show how the quality of the estimation varies over different model assumptions. Also, $N = 1000$ seems to be near the lower range of the number of initial pioneers one needs to contact to produce a reasonable prognosis for the development of the campaign.

[3] The simulations are run in *python* by modifying code from the *networkx* package. Remark that the *directed configuration model* in networkx package is a completely different model despite superficial similarity.

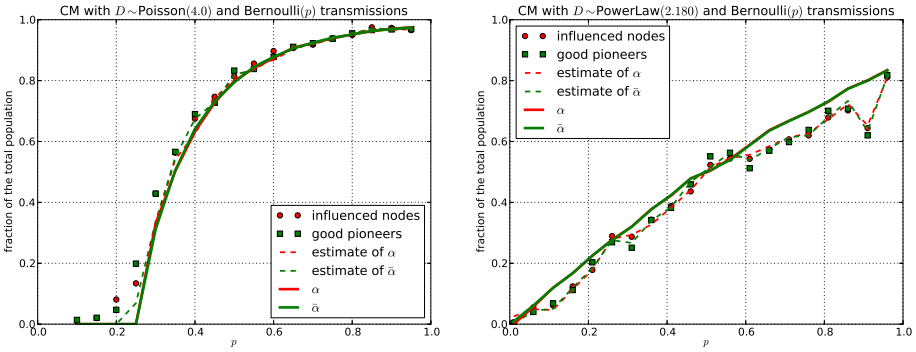

Fig. 1. CM with Poisson degree of mean $\lambda = 4$ and Power-Law degree of parameter $\beta = 2.180$ (corresponding to $\mathbb{E}[D] \approx 4$), both with Bernoulli transmissions with probability p. The set of good pioneers and the influenced population are of the same size. In the Poisson case their fraction is strictly positive for $p > 1/\lambda$ while in the Power-Law case it is so for all $p > 0$ whenever $\beta \leq 3$.

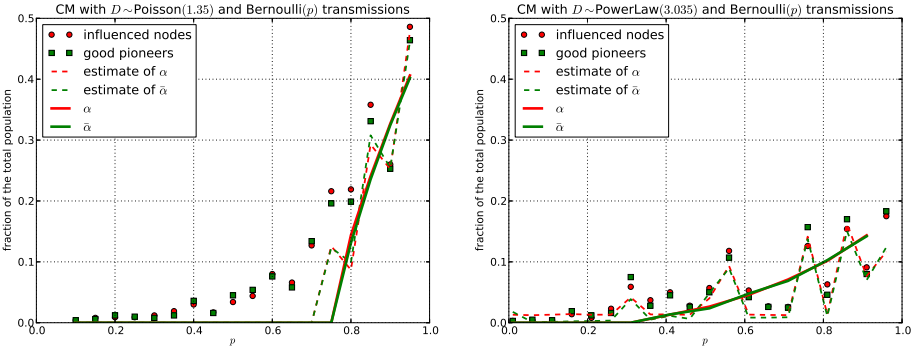

Fig. 2. CM with Poisson and Power-Law degree of mean $\mathbb{E}[D] \approx 1.35$ ($\lambda = 1.35$ and $\beta = 3.035$) and Bernoulli transmissions. The set of good pioneers and the influenced population are of the same size for each model. One observes the phase transition in both models, at $p = 1/\lambda$ and $p = \zeta(\beta-1)/(\zeta(\beta-2) - \zeta(\beta-1)$, respectively.

Case Study. Figure 1 presents Bernoulli influence propagation on the CM with Poisson and Power-Law degree distribution of mean $\mathbb{E}[D] = 2$. Bernoulli transmissions imply that the sets of good pioneers and influenced population are of the same size. The Power-Law degree with $\beta < 3$ leads to a positive fraction of good pioneers and influenced component for all $p > 0$, while for the Poisson degree distribution one observes the phase transition at $p = 1/\lambda$. That is, the fractions of good pioneers and the influenced component are strictly positive, if and only if, $p > 1/\lambda$.

Figure 2 shows again the model with Bernoulli transmissions on CM with Poisson and Power-Law degree distribution, this time for $\mathbb{E}[D] \approx 1.35$ for which both models exhibit the phase transition in p.

A general observation is that the Power-Law degree distribution gives smaller critical values of p for the existence of a positive fraction of influenced popula-

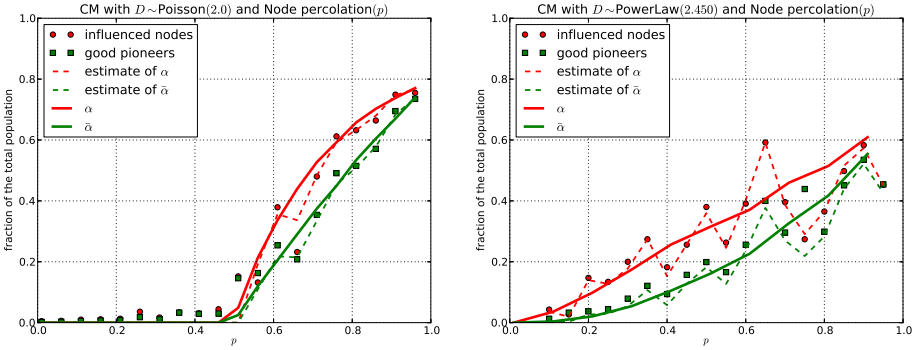

Fig. 3. Node percolation ("apathetic and enthusiastic users") on CM with Poisson and Power-Law degree of mean $\mathbb{E}[D] \approx 2$ ($\lambda = 2$ and $\beta = 2.45$). The influenced component and the critical values for p are equal to these for the CM with Bernoulli transmissions. The set of good pioneers is smaller than the influenced population. We do not observe the phase transition for the Power-Law model since $\beta < 3$.

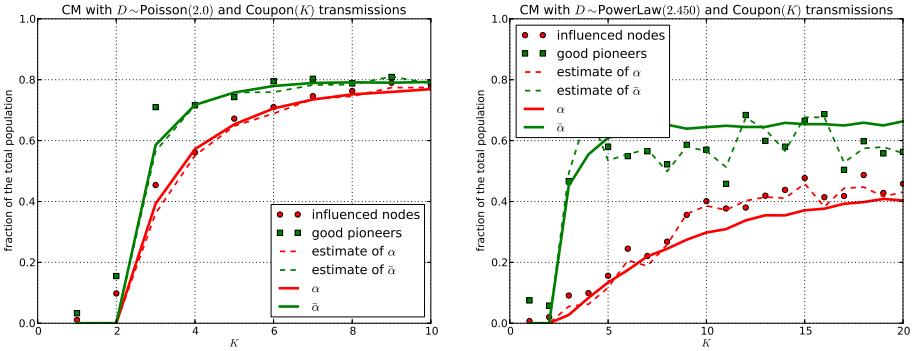

Fig. 4. Coupon collector dynamics "absentminded users") on CM with Poisson and Power-Law degree of mean $\mathbb{E}[D] \approx 2$ ($\lambda = 2$ and $\beta = 2.45$). The set of good pioneers is bigger than the influenced population.

tion and good pioneers, however for these, the size of these sets increase more slowly with the transmission probability, p, in the Bernoulli model. Obviously, the values of $\alpha = \overline{\alpha}$ at $p = 1$ correspond to the size of the biggest connected component of the underlying CM.

Figure 3 shows the node percolation (or the case of "apathetic and enthusiastic users") on CM with Poisson and Power-Law degree distribution of mean $\mathbb{E}[D] \approx 4$. Note that the influenced components have the same size as for Bernoulli transmissions, however good components are smaller. The critical values of p for the phase transition are also the same as for Bernoulli transitions. Note that the estimation of the node percolation model is more difficult than the Bernoulli transmissions because of higher variance of the estimators.

Finally, Figure 4 shows that the coupon collector dynamic (the case of "absentminded users") on CM produces bigger sets of good pioneers than the influenced population.

4 Application to Viral Campaign Evaluation

From the point of view of a firm which has no prior information about the network structure and the campaign effectiveness, it could be useful to assume that the network is a uniform random network, with the total degree and transmitter degree distributions estimated using the information collected from the initial set of pioneers targeted. The collected information, denote it by $(D_i, D_i^{(t)})$, $i = 1, \ldots, N$, allows to estimate various quantities relevant to the potential development of the ongoing campaign, as we did in 3.4. Particularly relevant are the estimates for the following.

Network fragmentation If the value of the estimator, $\mathbb{E}[D^2-2D] \approx \frac{1}{N}\sum_{i=1}^{N}\Big(D_i^2-2D_i\Big)$, is not sharply larger than zero, then the firm must assume that the network is too fragmented to allow for viral marketing (condition (2)).

Effectiveness of the campaign If one estimates that the network is not too fragmented, then the firm can evaluate the effectiveness of the ongoing campaign using an estimate of $\mathbb{E}[DD^{(t)} - D^{(t)} - D] \approx \frac{1}{N}\sum_{i=1}^{N}\Big(D_i D_i^{(t)} - D_i - D_i^{(t)}\Big)$ (condition (1)). If the value of this estimator is sharply larger than zero, then the firm can assume that there is a realistic chance of picking a good pioneer via random sampling and make the campaign go viral.

The estimates of the fraction of good pioneers and the vulnerable population can also be considered to make a cost-benefit analysis of the marketing campaign.

5 Conclusion

Diffusion studies on networks generally tend to focus on the component that can be reached starting from an initial target. Our work in [3], over and above, focuses on the set of good pioneers based on a new approach which consists of identifying this subset as the big component of a *reverse dynamic* in which an "acknowledgment" message is sent in the reversed direction on every edge thus allowing to trace all the possible sources of influence of a given vertex.

In this paper, we consider what insight the graph-theoretical results obtained through this approach provide about the phenomenon of viral marketing on online social networks, particularly to a firm trying to decide how much to spend on a marketing campaign which might go viral or not.

References

1. Bailey, N.: The Mathematical Theory of Infectious Diseases. Books on cognate subjects. Griffin (1975)
2. Banerjee, A.V.: A simple model of herd behavior. The Quarterly Journal of Economics 107(3), 797–817 (1992)

3. Błaszczyszyn, B., Gaurav, K.: Viral marketing on configuration model. arxiv 1309.5779 (2013) (submitted)
4. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted? In: Proc. of WWW, pp. 925–936 (2014)
5. Janson, S., Luczak, M.J.: A new approach to the giant component problem. Random Structures and Algorithms 34(2), 197–216 (2008)
6. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proc. of ACM SIGKDD, KDD 2003, pp. 137–146. ACM, New York (2003)
7. Moore, C., Newman, M.E.J.: Epidemics and percolation in small-world networks 61, 5678–5682 (May 2000)
8. Van Der Hofstad, R.: Random graphs and complex networks (2009), `http://www.win.tue.nl/rhofstad/NotesRGCN.pdf`
9. Yagan, O., Qian, D., Zhang, J., Cochran, D.: Conjoining speeds up information diffusion in overlaying social-physical networks. IEEE JSAC 31(6), 1038–1048 (2013)