Chapter 29
Scale-Free Network Topologies with Clustering Similar to Online Social Networks

Imre Varga

Abstract In this paper I propose a novel method to model real online social networks where the growing scale-free networks have tunable clustering coefficient independently of the average degree and the exponent of the degree distribution. Models based on purely preferential attachment are not able to describe high clustering coefficient of social networks. Beside the attractive popularity my model is based on the fact that if a person knows somebody, probably knows several individuals from his/her acquaintanceship as well. The topological properties of these complex systems were studied and it was found that in my networks the cliques are relevant independently of the system size as usual in social systems.

29.1 Introduction

While networks are present everywhere in our everyday life, these complex systems attract considerable scientific interest. Researches showed that social networks are different from other networks in some sense. The reason of this was studied by Newman and Park [1]. The biggest difference is in average clustering coefficient. In social networks there is a high probability that two friends of a given individual will also be friends of each other thus the clustering coefficient is high. Opposite to non-social networks, where these triangles are rare.

Many models of networks appeared in the last decades, but most of them are not able to describe social networks directly. Models based on "small-world" networks of Watts and Strogatz [2] do not reproduce the power law degree distribution. Most of growing scale-free network models result low clustering coefficients [3–5]. There are some trials to create scale-free networks with tunable clustering [6–9], but in these models the desired value of clustering coefficient determines other properties of the networks. Avoiding this problem I wanted to create a model for online social

I. Varga (✉)
Department of Informatics Systems and Networks, University of Debrecen, Debrecen, Hungary
e-mail: varga.imre@inf.unideb.hu

H. Takayasu et al. (eds.), *Proceedings of the International Conference on Social Modeling and Simulation, plus Econophysics Colloquium 2014*, Springer Proceedings in Complexity, DOI 10.1007/978-3-319-20591-5_29

networks in which I can set the average clustering coefficient without affecting other properties (e.g. degree distribution exponent, average degree) of the network.

29.2 Basic Model

In order to achieve my goal I generalized the well-known Barabási-Albert (BA) model [3] modifying the linking method. The growing networks start from a small fully connected network of N_0 nodes where each nodes have $N_0 - 1$ links to others. Then I start to grow the network by adding more and more new nodes to it step by step. When a new node joins it is attached by $m = N_0 - 1$ links to existing nodes. These vertices are chosen by two different ways.

(a) Some nodes are chosen based on preferential attachment. The probability of a node to be chosen is proportional to the number of existing connections of it. Thus nodes with more neighbors have larger probability to get a new one. The number of these chosen nodes is denoted by π.

(b) In the second phase the new node is linked to ν number of neighbors of each previously chosen vertices. The neighbors of popular nodes have the same probability to be linked to the new node, independently from their degree.

The exact linking algorithm has the following steps:

1. *Create a new node,* $i = 1$.
2. *If* $i > \pi$*, then the linking method of this node is over.*
3. *Link the node to a probably large degree, popular one by preferential attachment.* $i = i + 1$ *and* $j = 1$.
4. *If* $j > \nu$*, go to step 2.*
5. *Link the node to one of the neighbors of ith popular neighbor of this node with equal probability.* $j = j + 1$.
6. *Go to step 4.*

These steps are repeated until the number of nodes N reaches a desired value ($N \gg N_0$). The basic idea of this two-phase linking is that to have a popular friend is advantageous and then one gets to know some acquaintances of the popular friend. Finally the number of links of a new node can be written as $m = \pi(1 + \nu)$. This method is a kind of generalized version of BA model, if $\nu = 0$ the networks generated by these two methods are the same. Now the model has three independent parameters: N, π and ν.

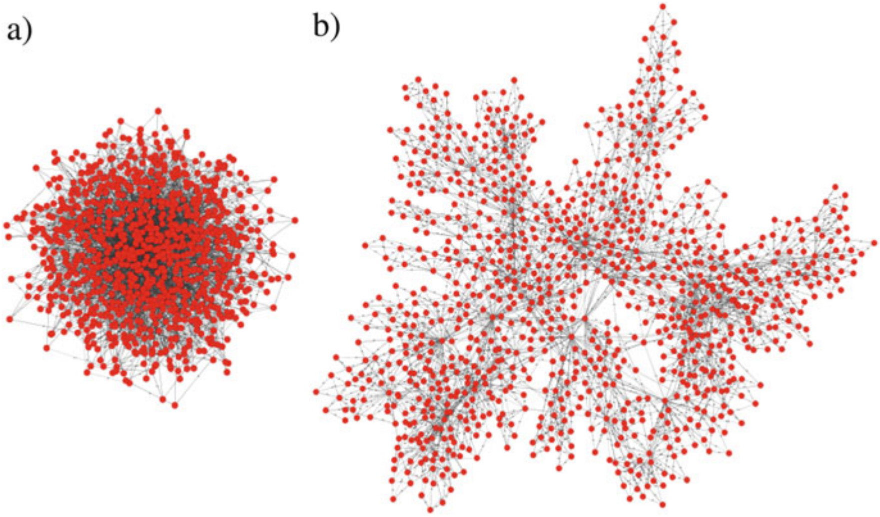

Fig. 29.1 The graphs of the model with the same number of nodes ($N = 1000$) and links ($m = 3$) using the same representation technique. On the *left side* the graph is a BA network ($\pi = 3, \nu = 0$). On the *right side* a completely different graph of my model ($\pi = 1, \nu = 2$) is presented

29.2.1 Properties of Generated Networks

This small change in the generation method leads to large differences in the network properties compared to BA-model. The differences can be seen right at the first sight even if the average degree and the density is the same (see Fig. 29.1).

In order to characterize the differences quantitatively I studied different properties first of all the average shortest path length $\langle L \rangle$ in the generated networks. It is small compared to the number of nodes and links. I found that $\langle L \rangle$ grows proportionally to the logarithm of N, so the networks have small-world property as expected. The coefficient of this proportionality depends on the parameters π and ν. BA-like networks have smaller average shortest path length than networks with high value of ν. The reason of this is the fact that in the latter case the graphs contain networks of small strongly connected groups of nodes due to the linking method. So increasing ν (at the same value of m) results networks where cliques are more important. Naturally larger number of links leads to smaller networks, where $\langle L \rangle$ obeys power law decay with parameter dependent exponent. Based on my simulation results curve fitting showed (see Fig. 29.2a) that the average shortest path length has the following functional form

$$\langle L \rangle \propto (\pi(\nu + 1))^{-F(\pi,\nu)} \ln N. \qquad (29.1)$$

However initially nodes have the same amount of neighbors finally their degree varies in a wide range. Based on the growing algorithm one can analytically

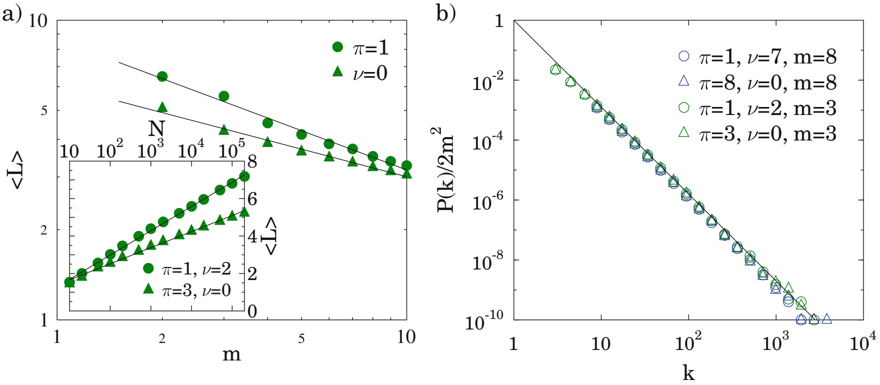

Fig. 29.2 (**a**) The average shortest path length $\langle L \rangle$ as a function of $m = \pi(\nu + 1)$. *Straight lines* indicate power law dependency on log-log scale. *Inset*: the average shortest path length $\langle L \rangle$ as a function of system size N on lin-log plot. In case of same density the BA-like graphs are smaller than generalized graphs. *Straight lines* indicate fits with Eq. (29.1). (**b**) All the graphs generated by this method have power law degree distribution. Rescaling the degree distribution data collapse occurs independently of π and ν. The exponent of the *solid line* is 2.9 as in BA model

determine the average degree of nodes

$$\langle k \rangle = 2m = 2\pi(1 + \nu). \tag{29.2}$$

The degree distribution can be well fitted by a straight line on log-log scale indicating scale-free networks with power law degree distribution with form $P(k) \propto k^{-\gamma}$. The curves with different values of π and ν can be rescaled by $2m^2$ to get data collapse as it is shown in Fig. 29.2b. This means that the exponent is independent from m in all cases not only for BA networks. The exponent γ of the degree distribution is independent of the number of nodes connected in the first step π and in each secondary step ν as well, its value is $\gamma = 2.895 \pm 0.038$ as expected. The value of the exponent is obtained by averaging the exponents of systems at different input parameter combinations. This independence needs some explanations. Let's see for example the $\pi = 1$ and $\nu = 9$ system. Only 10 % of the links based on purely preferential attachment and 90 % just randomly connected to the neighbors of popular nodes. How can this network be scale-free? As a matter of fact the 90 % also preferred, because sooner or later these neighbors also become popular as they popular neighbor gets more and more links.

To characterize the networks from the point of view of the cliques I calculated the clustering coefficient of nodes in my undirected graphs. Local clustering coefficient C of a node is the ratio of the number of existing links between neighbors of this node and the number of possible connection between them. In a general case C is proportional to the reciprocal of the degree of node, which indicates small degree nodes are mainly members of cliques while hubs of the networks connect them together.

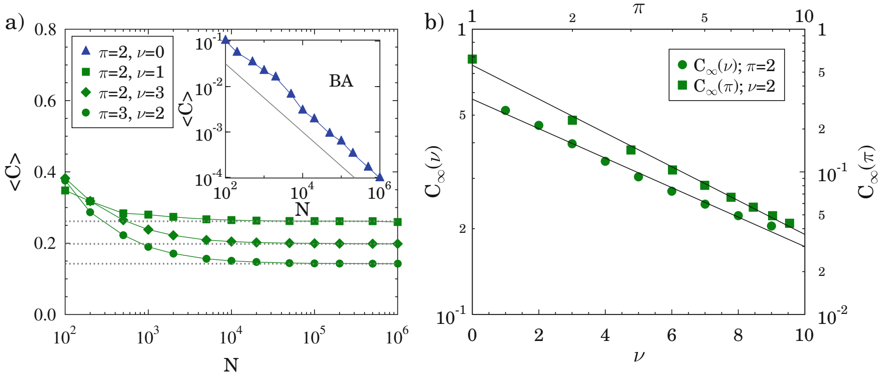

Fig. 29.3 (**a**) The average clustering coefficient $\langle C \rangle$ is decreasing with the increasing number of nodes N in the system, but it tends to zero only in BA networks (*inset*). When parameter $\nu > 0$ in a large network the value of $\langle C \rangle$ is constant. Values of C_∞ (obtained by curve fitting) which are determined by π and ν are indicated by *dashed lines*. (**b**) C_∞ as a function of ν on log-lin plot and C_∞ as a function of π on log-log plot fitted by Eq. (29.4)

The most interesting feature of my graphs can be seen if we analyze their average clustering coefficient $\langle C \rangle$. When a network is growing, $\langle C \rangle$ is decreasing. I found this can be written in the following functional form

$$\langle C \rangle \propto N^{-3/4} + C_\infty, \tag{29.3}$$

where N is the number of nodes and C_∞ is a constant at given parameter set. In case of BA network ($\nu = 0$) the value of $C_\infty = 0$, so we get back the well-known power law form. In this systems the formation of neighbor-triangles is random. Increasing the system size the degree of nodes is increasing as well so the chance of a node to belong mainly link-triangles is continually decreasing. This leads to small clustering coefficient. In generalized cases Eq. (29.3) means that $\langle C \rangle$ tends to finite values, not to zero. If $\nu > 0$, new nodes mainly compose triangles (independently from system size) due to the linking algorithm, so a given part of the system always have large clustering coefficient. One can see it on Fig. 29.3a. It indicates that when $\nu = 0$ in a large network cliques are negligible, while in the generalized networks they remains important at any system sizes. Large number of simulations were performed to discover how the constant value in $\langle C \rangle$ depends on the input parameters. I found that

$$C_\infty \propto \pi^{-A} e^{-B\nu}, \tag{29.4}$$

if $\pi > 1$ and $\nu > 0$, where A and B are constants. More links lead to smaller average clustering coefficient, where both types of linking methods (π and ν) have influence on C_∞ but they act in different ways. (See Fig. 29.3b.) Generally preferential links do not compose new triangles, so increasing π results just larger degree, but not more triangles. That is the reason why larger π leads to smaller $\langle C\rangle$. Larger value of ν creates more triangles, however these are independent, so they do not form tetrahedron-like structure. $\langle C\rangle$ is also decreasing. Practically speaking my linking method makes us able to generate large scale-free networks with different discrete values of average clustering coefficient in a wide range between 0 (BA) and the maximum at $\pi = 1, \nu = 1$ namely 0.739, however smaller values are more common. If we have maximum 15 edges to each new node ($m \leq 15$) we can create networks with 45 different values of C_∞.

29.3 Extended Model

At this point we are able to adjust the average clustering coefficient by the input parameters. However the values of π and ν determine the average degree of nodes as well. In order to model different real world networks we must tune $\langle C\rangle$ and $\langle k\rangle$ independently. That is the reason why my model has been extended. To change the number of links a reduction process is applied. After the growing period the system undergoes a destroying procedure where independently chosen nodes and their connections are removed. I used the so called *general attack* process [5] which means that all the nodes has the same probability to be removed. The strength η of this reduction process can be characterized by the ratio of number of removed nodes ΔN and the original number of nodes at the end of growing phase, so $\eta = \Delta N/N$. Thus finally the extended model has four parameters: N, π, ν and η. This reduction process has significant influence to the topological properties of the network.

29.3.1 Properties of the Reduced Networks

Remaining nodes loose connections by removing their neighbors. The final average degree in the system is determined by three things which can be expressed as

$$\langle k\rangle = \frac{\sum_i k_i - \sum_j k_j - \sum_l k_l}{N - \Delta N}, \tag{29.5}$$

where $i = 1, 2, \ldots, N$, j runs over removed nodes and l runs over the remained neighbors of removed vertices. The first term in the numerator is the sum of original degree of nodes before reduction. The second one is the loss of degree of the removed nodes. The third term describes the loss of degree due to the fact that remained nodes lose the links to removed neighbors. While removed nodes can have links to other removed nodes as well, the last two terms are not equal, their ratio is $(1 - \eta)$. In this way the Eq. (29.5) can be written as follow using mean field

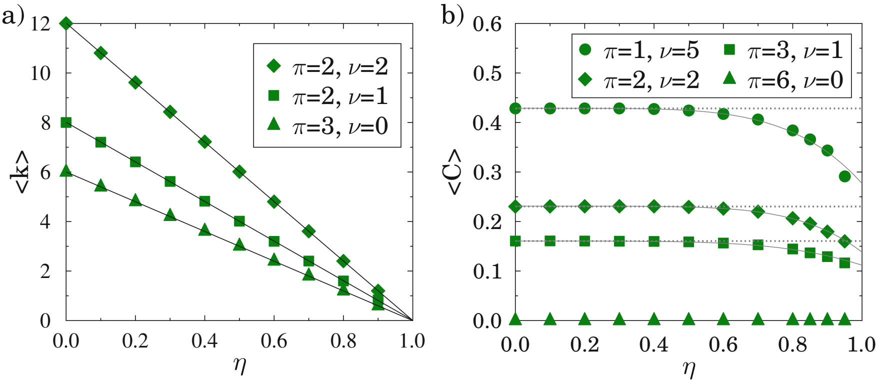

Fig. 29.4 (**a**) The average degree $\langle k\rangle$ is decreasing linearly with the reduction strength η. (Fitted by Eq. (29.7).) (**b**) The average clustering coefficient is decreasing very slowly during the reduction process. For small reduction it remains almost constant. In case of BA network ($\nu = 0$) $\langle C\rangle$ is always close to zero. *Dotted lines* denote C_∞ and *grey fitted curves* represent Eq. (29.9), where R^2 coefficient is above 0.96 for all $\nu > 0$ data sets

approximation

$$\langle k\rangle = \frac{2mN - 2m\Delta N - 2m\Delta N(1-\eta)}{N-\Delta N}. \tag{29.6}$$

Using Eq. (29.2) and the definition of η the Eq. (29.6) can be simplified to

$$\langle k\rangle = \frac{2m(1-\eta-\eta(1-\eta))}{1-\eta} = 2m(1-\eta) = 2\pi(1+\nu)(1-\eta). \tag{29.7}$$

In my simulations the average number of links of nodes decreases linearly with increasing reduction strength as predicted analytically. The effect of the reduction process on $\langle k\rangle$ is illustrated in Fig. 29.4a.

The reduction has only minor influence on average clustering coefficient, which is negligible even if half of nodes are removed. Stronger reduction leads to a bit smaller value of $\langle C\rangle$. I determined the functional form of this dependency which can describe as

$$C_\infty - \langle C\rangle \propto \eta^D \tag{29.8}$$

for large networks, where exponent D determines how fast the average clustering coefficient decreasing. (See Fig. 29.4b.) Using Eqs. (29.3), (29.4), and (29.8) finally we can write the average clustering coefficient as a function of input parameters of the model if $\pi > 1$ and $\nu > 0$

$$\langle C\rangle \propto KN^{-3/4} + K'\pi^{-A}e^{-B\nu} - K''\eta^D, \tag{29.9}$$

where K, K', K'', A, B and D are coefficients and exponents of the model.

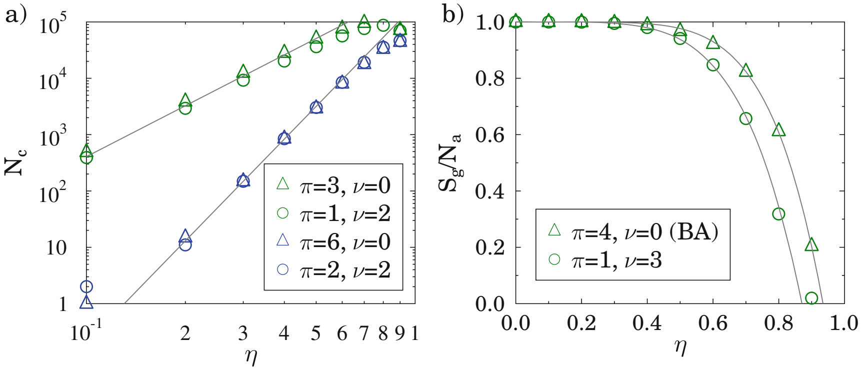

Fig. 29.5 (**a**) Number of clusters N_c as a function of reduction strength η on log-log scale. *Straight lines* indicate power law behavior, where the exponent depends only m, but independent from π and ν. (**b**) Strong reduction destroys giant component, it disappears faster in generalized networks. The decay can be described by Eq. (29.11) illustrated by *grey curves*

The values of $\langle k \rangle$ and $\langle C \rangle$ in my network are independently tunable with the reduction process, which has other side effects. The originally connected networks fall into pieces. Separate clusters appear, which are smaller networks without connections to other parts of the system. Increasing the reduction strength η the number of clusters N_c is increasing according to power law, where the exponent depends on the number of links only, independently from their role in the growing process (Fig. 29.5a). Large number of clusters can occur depending on η and the system size N. Based on the simulation results the value of N_c can be characterized by the following form

$$N_c \propto \frac{N}{\pi(\nu+1)} \eta^{\pi(\nu+1)}, \tag{29.10}$$

if the reduction is not negligible. When the reduction is very strong the number of clusters N_c saturates.

If the reduction strength is smaller than approximately 0.4 clusters are negligible except one which gives almost 100 % of the system. It is called giant component in the literature. It can be still dominant even if more than 75 % of the nodes are removed. After this the dominancy of giant component disappears fast in case of strong reduction. The speed of this process depends on the growing period. Not only the number of links of a new node m are important, but also the parameters π and ν separately. The size of giant component S_g can be written by the form

$$S_g \propto N_a(1-\eta^{E(\pi,\nu)}) = N(1-\eta)(1-\eta^{E(\pi,\nu)}), \tag{29.11}$$

where $N_a = N - \Delta N$ is the number of nodes in the reduced system. (See Fig. 29.5b.) The exponent E depends not only on the value of m, but also π and ν, however larger m results smaller exponent, so larger giant component. In BA networks ($\nu = 0$) the giant component is always larger than in generalized networks at a given link number. This shows that BA networks are strongly connected while if $\nu > 0$ the system is a weakly connected set of densely linked groups of nodes. Since the number of clusters is independent from π and ν at a given value of m, but the size of giant component is smaller for larger ν, clusters (excluding the giant component) are larger. The average cluster size is much smaller in BA networks then in the generalized case. These are also proofs of presence and importance of cliques. These clusters have a power law size distribution with a parameter dependent exponent. Number of clusters $n(S)$ of size S can be expressed as

$$n(S) \propto S^{-\tau(\pi,\nu)}. \tag{29.12}$$

29.4 Model of Real Online Social Network

Due to the discussed topological properties my networks are appropriate candidates for modeling real world online social networks. I managed to get a set of data of almost 60 million Facebook users [10]. This network has small world property, its degree distribution can be characterized by two power law regimes (see Table 29.1), so it is a kind of scale-free network. The quite high average clustering coefficient indicates the presence of cliques of users.

Based on my presented results I found a set of input parameters which leads to a very similar network. The values of input parameters in my Facebook model are: $\boldsymbol{\pi = 3}$, $\boldsymbol{\nu = 1}$ and $\boldsymbol{\eta = 0.72}$ ($N = 10{,}750{,}000$). This final sample contains more than three million nodes. In this size scale N has not got influence to the network properties, so not necessary to create larger system. The properties of the real social network and my model network are summarized in Table 29.1. As one can see the values of the main quantities $\langle C\rangle$ and $\langle k\rangle$ well describe the real case and other properties give quite good qualitative description (e.g. presence of separate clusters or power law degree distribution) as well.

Table 29.1 Comparison of my extended model network and the Facebook data set

		Facebook	Extended model
Average shortest path length	$\langle L\rangle$(N)	Logarithmic	Logarithmic
Degree distribution	$P(k)$	Power law	Power law
Degree distribution exponent	γ	1.32, 3.38	2.96
Average degree	$\langle k\rangle$	**3.13**	**3.24**
Dominance of giant component	S_g/N_a	0.99	0.90
Cluster size distribution	$N(S)$	Power law	Power law
Average clustering coefficient	$\langle C\rangle$	**0.16**	**0.15**

The features of the two networks are in good agreement

29.5 Conclusion

In summary, I proposed a simple method for generating scale-free networks where the average clustering coefficient is tunable in a broad range and determined by the input parameters π and ν. The method is a kind of generalized version of growing Barabási-Albert model where the links of a new node play different roles. Beside the preferential attachment some links obey the so called "friend of my friend is my friend" philosophy. After the growing process a reduction process was used in order to create large variety of networks changing $\langle k \rangle$ and $\langle C \rangle$ independently. This reduction process means random removal of nodes. The strength of reduction is characterized by parameter η. A detailed study of the model was presented proofing that in these scale-free networks the cliques have very important role which cannot be described by the original BA model. Comparing a real online social network and the graphs generated by the proposed algorithm I found very good agreement. For clarity my model does not describe the time evolution of real social networks just generate graphs topologically similar to a given state of real online social networks. In the near future the model networks are being subjected to agent-based simulation of information spreading using the model of Kocsis and Kun [11]. This model can be a good base of later study of effectiveness of advertising in online social networks.

Acknowledgements The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

The computations were performed on the HPC centre provided by the Hungarian National Information Infrastructure Development Institute at the University of Debrecen.

Special thanks to my colleague Gergely Kocsis for fruitful discussions.

References

1. Newman MEJ, Park J (2003) Why social networks are different from other types of networks. Phys Rev E 68:036122
2. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393(6684):440–442
3. Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
4. Lee HY, Chan HY, Hui PM (2004) Scale-free networks with tunable degree distribution exponents. Phys Rev E 69:067102
5. Varga I, Németh A, Kocsis G (2013) A novel method of generating tunable underlying network topologies for social simulation. In: Proceedings of the 4th IEEE international conference on cognitive infocommunicaitons, pp 71–74
6. Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. Phys Rev E 65:026107
7. Newman MEJ (2009) Random graphs with clustering. Phys Rev Lett 103:058701

8. Heath LS, Parikh N (2011) Generating random graphs with tunable clustering coefficients. Phys A 390:4577–4587
9. Samalam VK (2013) A model for generating tunable clustering coefficients independent of the number of nodes in scale free and random networks [arXiv:1311.6401]
10. Gjoka M, Kurant M, Butts CT, Markopoulou A (2010) Walking in Facebook: a case study of unbiased sampling of OSNs. In: Proceedings of the IEEE INFOCOM ’10, pp 1–9
11. Kocsis G, Kun F (2011) Competition of information channels in the spreading of innovations. Phys Rev E 84:026111