# Emergence and Size of the Giant Component in Clustered Random Graphs with a Given Degree Distribution

Yakir Berchenko,[1] Yael Artzy-Randrup,[2,3] Mina Teicher,[1,4] and Lewi Stone[2,*]

[1]*Brain Research Center, Bar Ilan University, Ramat Gan, Israel*

[2]*Biomathematics Unit, Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel*

[3]*Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, USA*

[4]*Department of Mathematics, Bar Ilan University, Ramat Gan, Israel*

(Received 26 September 2007; published 30 March 2009)

Standard techniques for analyzing network models usually break down in the presence of clustering. Here we introduce a new analytic tool, the ''free-excess degree'' distribution, which extends the generating function framework, making it applicable for clustered networks ($C > 0$). The methodology is general and provides a new expression for the threshold point at which the giant component emerges and shows that it scales as $(1 - C)^{-1}$. In addition, the size of the giant component may be predicted even for more complicated scenarios such as the removal of a fixed fraction of nodes at random.

DOI: 10.1103/PhysRevLett.102.138701 PACS numbers: 89.75.Hc, 64.60.A−

Network theory is a powerful tool for describing and modeling complex systems with applications in widely differing areas, including epidemiology, neuroscience, ecology, and the Internet [1,2]. Over the last years there has been rapid progress in calculating the properties of random networks having arbitrary degree distributions, based on the analysis of branching processes and generating functions (GFs) [3,4]. This was motivated by the observation that many real-world networks are often characterized by highly heterogeneous degree distributions [5]. However, clustering also characterizes real-world networks, yet it remains far less understood. Clustering refers to the relative number of triangles in a network, commonly measured by the coefficient introduced in [3] as $C = \frac{3\times N_{\triangle}}{N_3}$. Here $N_{\triangle}$ is the number of triangles in the network, while $N_3$ is the number of connected triples of nodes. This definition has the advantage that $C$ is also the probability that two nodes which connect to a mutual node are connected themselves, thereby forming a triangle whereby ''a friend of a friend is also a friend.''

The analysis of branching processes, which are at the heart of the GF formalism of [3], is not applicable for clustered networks due to the formation of short loops, namely, triangles. Initially this problem was overlooked, and results were ''extrapolated'' from the unclustered case in a simplistic manner to the clustered one. Relevant here is the example concerning the emergence of the giant component (GC)—where it was shown [3] that in the usual unclustered case there is a GC if the mean number of nodes at a distance two ($z_2$) is larger then the mean number of nodes at a distance one ($z_1$). This is often (wrongly) taken as the criterion for clustered networks [6,7], thus initiating the quest to calculate $z_2$ in the presence of clustering [2,6,7].

A recent analysis of the problem of clustering was given in Ref. [8], and is based on an approach similar to that described here. The authors of [8]construct a branching process that applies for networks with triangles, although it has the limiting assumption that two triangles will never share an edge. Even in this limited setting the results are only applicable for relatively small levels of clustering, $C$, and difficult to interpret and to broaden for other questions such as dilution (e.g., the effects of removing randomly a fixed fraction of nodes). In this Letter we describe general methods for determining the properties of clustered networks by introducing a new GF which is not part of the classical GF approach. Our central new result concerns the critical point for emergence of the GC, but we also address the question of its size and of dilution.

We first very briefly review the application of GFs for unclustered ($C = 0$) random networks [3]. Definition of the GF: $G_0(x) = \sum_{k=0}^{\infty} p_k x^k$, where $p_k$ is the probability that a randomly chosen vertex on the graph has degree $k$. Thus, the mean degree of a node is easily calculated as $z_1 := \langle k \rangle = \sum_k k p_k = G_0'(1)$.

Another quantity of importance is the ''excess degree'' distribution. Starting at a randomly chosen node and following one of the edges at that node, we reach a neighbor $v_1$. We are interested in the distribution of the outgoing edges of $v_1$ or its ''excess degree'' (i.e., the node's degree minus one, accounting for the edge we arrived along). Since the probability, $q_k$, to have $k$ outgoing edges is $q_k = (k + 1)p_{k+1}/z_1$, the distribution of outgoing edges is generated by the function [3] $G_1(x) := \sum_k q_k x^k = \frac{G_0'(x)}{G_0'(1)} = \frac{1}{z_1} G_0'(x)$. The mean excess degree is thus $z_e = \sum_k k q_k = G_1'(1)$. Using the ''powers'' property of GFs [3] (and considering [9]), the GF for the probability distribution of the number of second-nearest neighbors of the original node can be written as

$$\sum_k p_k [G_1(x)]^k = G_0(G_1(x)). \tag{1}$$

Thus the mean number of second-nearest neighbors is

 

$$z_2 = \left[\frac{d}{dx}G_0(G_1(x))\right]_{x=1} = G_0'(1)G_1'(1) = z_1 z_e \quad (2)$$

*The "free-excess" degree.*—The above calculations may be modified for application to clustered networks ($C > 0$). Analogously to the excess degree, starting at a randomly chosen node $v_0$ and following one of the edges at that node, we reach a neighbor $v_1$. We are now interested in $\{e_i\}_{i=0}^{\infty}$, the distribution of the outgoing edges of $v_1$ that are not connected to a neighbor of $v_0$.

Suppose we travel from node $v_0$ along an edge to node $v_1$ having degree $d(v_1) = i + 1$ (i.e., with an excess degree of $i$). The probability that it will have $k$ neighbors that are not connected back to $v_0$ (via a triangle) is

$$\binom{i}{k}(1-C)^k C^{i-k}. \quad (3)$$

This is just the probability that of the $i$ outgoing edges of $v_1$, $i-k$ are connected in a triangular formation that includes $v_0$, while the other $k$ edges are not. Here, as before, we use the first-order approximation that $C$ is the probability of a triangular formation.

When $d(v_1)$ is not known, from (3) we obtain

$$e_k := \sum_{i=0}^{\infty} q_i \binom{i}{k}(1-C)^k C^{i-k} = \sum_{i=0}^{\infty} q_i \binom{i}{k} C^i \left(\frac{1-C}{C}\right)^k. \quad (4)$$

The GF, $G_c(x)$, for the distribution is

$$\begin{aligned} G_c(x) &= \sum_{k=0}^{\infty} e_k x^k = \sum_{k=0}^{\infty}\sum_{i=0}^{\infty} q_i \binom{i}{k} C^i \left(\frac{1-C}{C}\right)^k x^k \\ &= \sum_{i=0}^{\infty} q_i C^i \sum_{k=0}^{\infty} \binom{i}{k}\left(\frac{1-C}{C}x\right)^k \\ &= \sum_{i=0}^{\infty} q_i C^i \left(1 + \frac{1-C}{C}x\right)^i = \sum_{i=0}^{\infty} q_i [C + (1-C)x]^i \\ &= G_1[C + (1-C)x]. \end{aligned}$$

Thus, we arrive at the key relationship:

$$G_c(x) = G_1[C + (1-C)x]. \quad (5)$$

As an example of how (5) may be useful, it is possible to determine the mean free-excess degree:

$$\sum_i i e_i = \left.\frac{dG_c(x)}{dx}\right|_{x=1} = (1-C)G_1'(1) = (1-C)z_e. \quad (6)$$

It will also prove useful to calculate the mean number of edges emanating outwards from nodes at a distance one to nodes at a distance two, beginning from some arbitrary source node (note that this is not the mean number of nodes at a distance two, since there is a positive probability that two edges reach the same node at a distance two). Similarly to (1) and (2), the mean is

$$\left.\frac{dG_0(G_c(x))}{dx}\right|_{x=1} = G_0'(1)G_1'(1)(1-C) = (1-C)z_1 z_e. \quad (7)$$

This parameter was also calculated in [2] by different means, but as will be discussed shortly, its importance appears to have been overlooked.

*Emergence of the GC.*—Molloy and Reed [10] introduced the parameter $Q := \sum_i i p_i (i-2)$ that identifies the phase transition in random graphs, i.e., the point where a GC is born. Their procedure utilizes a method which can be likened to "walking through a random graph" [Fig. 1(a)] and assessing the number of unknown nodes encountered along the way. Suppose one follows a random edge to a node $v$ having degree $k$. How does this change the number of unknown vertices? First, by virtue of arriving at $v$ the number of unknown vertices decreases by one. However, because $v$ itself has degree $k$, then this leads to an increase of $(k-1)$ in the number of unknown vertices. The net number of unknown vertices increases by $(k-2)$. In order to calculate the expected change, the probability of arriving at $v$, which is proportional to the degree $k$, must also be factored in. This makes the expected increase in the number of unknown neighbors proportional to $Q = \sum_i i p_i (i-2)$. If $Q > 0$, then with each step of the walk through the graph the number of unknown vertices and the size of the component grow large—the hallmark traits of the GC. If $Q < 0$, the number of unknown neighbors reduces to zero, and we cannot be walking through a GC. Recalling earlier definitions, the condition $Q > 0$ may be stated as

$$z_e > 1. \quad (8)$$

Since in unclustered ($C = 0$) networks $z_e = z_2/z_1$, Ref. [3] advocates the following equivalent criterion.

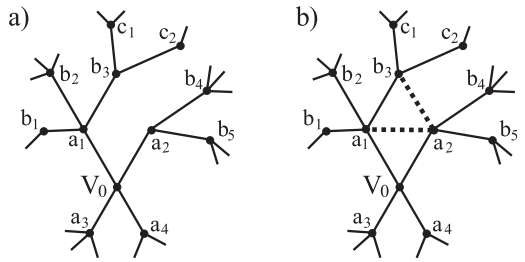


FIG. 1. Graphical illustration of the exposure procedure. Choose a node at random, say, $V_0$, and start diffusing from it and counting the nodes encountered on the way. (a) When $C = 0$ and the network is treelike [9], after counting the new nodes ($a_1 - a_4$) we pick one of them at random, say, $a_1$, and count its new neighboring nodes ($b_1 - b_3$), which are distributed according to $\{q_i\}_{i=0}^{\infty}$. In the next step, we randomly choose one of the nodes ($a_2 - a_4$, $b_1 - b_3$) and continue until the entire component is exposed. (b) When $C > 0$, two modifications are required to deal with cycles due to triangles (the dashed edges): we use $\{e_i\}_{i=0}^{\infty}$ and diffuse depthwise. After counting $a_1 - a_4$, when we count the neighbors of $a_1$, we avoid overcounting $a_2$ because $\{e_i\}_{i=0}^{\infty}$ govern the distribution of the solid-black edges. In the next step if we go from $a_1$ to $b_3$ in order to count the neighbors of $b_3$, again we avoid overcounting $a_2$ (because it is connected to $a_1$). The depthwise exposure, which is a permissible scheme [10], is made use of to avoid dependencies.

Criterion A—There is a GC in random networks if $z_2 > z_1$; i.e., the mean number of second-nearest neighbors is greater than the mean number of neighbors.

This has the intuitive epidemiological interpretation; if the mean number of infected individuals grows with distance from the source, an epidemic outbreak will occur.

The procedure of Molloy and Reed may be adapted for clustered networks. Again, suppose we follow a random edge that begins from a source node and ends at some node $v$. Previously, if $v$ had degree $k$, the number of "unknown" neighbors increased by $k-2$. However, with triangles there is a possibility that some of the $k-1$ outgoing edges will return to nodes that are already known [via dashed edges in Fig. 1(b)]. It is possible to avoid counting these nodes twice, by counting them in a manner that considers the free-excess degree distribution $e_k$. Thus, when a node $v$ of free-excess degree $i$ is encountered, the number of "unknown" neighbors increases by $i-1$, and the expected increase in the number of unknown neighbors is thus proportional to $Q_c = \sum_i e_i(i-1)$. The criterion for the GC in a clustered network is precisely $Q_c > 0$.

However, by (6), this condition becomes

$$(1-C)z_e > 1, \tag{9}$$

which differs from (8) by the scale factor $(1-C)$. Multiplying both sides by $z_1$, we obtain $(1-C)z_1 z_e > z_1$. Recalling (7), this may be interpreted as the following criterion.

Criterion B—There is a GC if the mean number of edges emanating outwards from nodes at a distance one to nodes at a distance two (beginning from some arbitrary source node) is larger than the mean degree.

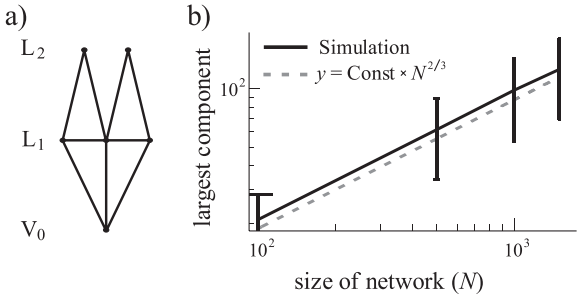


FIG. 2. The difference between the new criterion B and the conventional criterion A. (a) Consider the following example: a typical node has a neighborhood similar to $V_0$—3 nodes at a distance one at the first layer, $L_1$, and 2 nodes at a distance two at the second layer, $L_2$, but 4 edges to the second layer (from $L_1$ to $L_2$). Criterion A would not predict a GC, while criterion B would predict it. (b) The size of the largest component plotted vs $N$ for Poisson networks having mean degree $z_1 = 1.25$ and $C = 0.2$ (i.e., at the critical point according to criterion B). Indeed the size at the critical point correctly scales as $\sim N^{2/3}$ as is known for the case $z_1 = 1$ $C = 0$ (see reference in [2]). Note that criterion A would wrongly predict this regime to be below the critical point (since $z_2 \approx 1.19 < z_1$), and would suggest that all components should scale as $O(\log N)$.

Note that in the epidemiological sense, the emphasis is on the growth in the number of outward edges or transmission routes from a typical source node to its neighbors, and then to its neighbors' neighbors [Fig. 2(a)].

Although previously criterion A was used for a clustered network without any proper justification [2,6,7], Fig. 3 shows that it provides poor predictions of the critical mean degree $z_1^*$ as a function of the clustering, $C$. (Predictions were made using estimates of $z_2$ in the presence of clustering as detailed in [2,6].) The accuracy of the prediction can be assessed against simulations [11] (Fig. 3, circles and inset). In contrast, criterion B is a much better predictor as shown in Figs. 2(b) and 3. The latter plots the analytic result for a Poisson degree distribution where $z_1 = z_e$ [3] and $z_1^* = (1-C)^{-1}$ [from (9)].

*The size of the GC.*—In order to find the size of the GC Andersson [12] examined the probability of extinction in a two-phase branching process that mimics a construction of a random graph (with $C = 0$). In this branching process the source node has a number of direct descendants distributed according to $\{p_i\}_{i=0}^{\infty}$ (the first phase), while each of its descendants has a number of direct descendants distributed according to $\{q_i\}_{i=0}^{\infty}$ (the second phase). First, consider the probability $u$ for a lineage of a single branch that arrives at some node, $v_1$, to eventually die out. This necessitates that all $k$ branches leaving $v_1$ die out, an event that occurs with probability $u^k$. Since the degree of $v_1$ is unspecified, we obtain the self-consistency condition $u = \sum_{k=0}^{\infty} q_k u^k = G_1(u)$, which may be solved to find $u$.

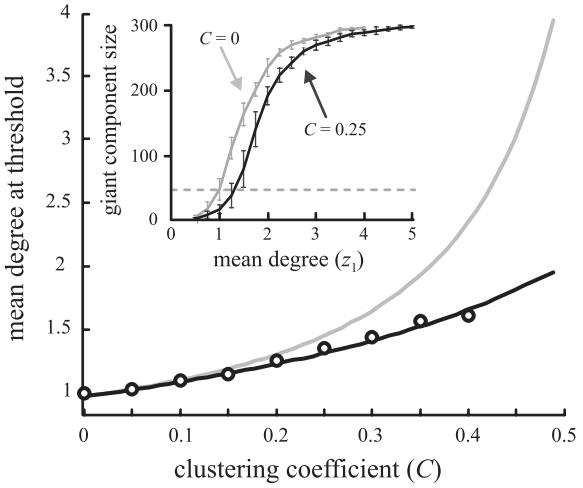


FIG. 3. The critical mean degree $z_1^*$ for the formation of a GC, plotted as a function of $C$, for Poisson degree distribution. Predictions of criterion A (grey line; $z_2$ estimated as in [6]). Predictions of criterion B [black line; $z_1^* = (1-C)^{-1}$ (see text)]. Empirical estimates of $z_1^*$ (circles) were obtained through the following procedure in order to overcome finite size effects: First the value of the size of the largest component was found for networks with $C = 0$ at the known threshold $z_1^* = 1$ (inset; horizontal grey line). This value was used to identify the critical threshold in comparable networks with $C > 0$.

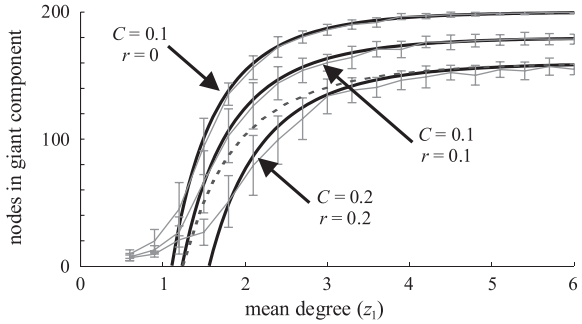


FIG. 4. The size of the GC, after a fraction $r$ of the nodes were removed randomly, as a function of the mean degree for $C = 0.1$ and $C = 0.2$. Black (grey) solid lines are predictions (simulations). The dashed line is the case $C = 0$, $r = 0.2$ given for comparison.

The second step takes into consideration that the branching process begins from some arbitrary source node. Because all branches originating from the source must die out in order for the process to become extinct, the probability of extinction (which is equivalent to belonging to a small component) is equal to $G_0(u)$, while the probability of persistence (or belonging to a GC) is $S = 1 - G_0(u)$, which is also the size of the GC.

The above argument needs to be modified for clustered networks. For these, the probability, $u$, for the lineage of a single branch to die out no longer fulfills the condition $u = G_1(u)$, because the progeny in the second phase are no longer distributed by $\{q_i\}_{i=0}^{\infty}$. Instead it is necessary to replace $q_i$ with $e_i$ so that the self-consistency condition is, to a close approximation, $u = G_c(u)$.

The main error remaining is due to higher order correlations between nodes in the branching process that occur with probability of $\sim C^2$ (and smaller when there are no triangles sharing an edge, as assumed in [8]).

The size of the GC may also be estimated in the presence of dilution, i.e., when a fraction $r$ of the nodes has been randomly removed [13,14]. In the solutions found by [14], it is only necessary to replace the $\{q_i\}_{i=0}^{\infty}$ with the free-excess probabilities $\{e_i\}_{i=0}^{\infty}$ (or $G_1$ with $G_c$). It is thus only necessary to (a) solve for $u$, such that $r + (1 - r)G_c(u) = u$, and (b) calculate GC size as $S = 1 - r - (1 - r)G_0(u)$. When $C = 0$, these equations coincide with those in [14]. Simulations and theory for node removal are in excellent agreement (Fig. 4), and even better for edge removal or joint edge-nodes removal [15].

*Discussion.*—According to [4], one can study random graphs with two different techniques, either by ''growth as an epidemic''—where the layers away from an origin node are fully exposed one at a time (e.g., [3] with the GF approach)—or by a ''random walk,'' which exposes nodes one at a time (e.g., [10]). The first approach was adopted in [8] to treat all realizations with a given clustering, which do not include triangles sharing a common edge (''low clustering''), since in such cases there is a possible correlation between different nodes in the same layer. In this Letter we succeeded in analyzing realizations having high clustering, by adapting the second approach. We overcame the correlation problem by exposing one node at a time (in particular, diffusing depthwise according to the free-excess degree). We were able to identify the emergence of the giant component by establishing the existence of a large ''scaffold'' suggesting it (on the assumption that the probability of a triadic closure is $C$). In order to calculate the actual size of the GC, we approximate with a $\sim C^2$ error and take the growth of the GC as an ''epidemic.'' However, since $C^2 \ll 1$ in many real-world networks, we expect this discovery to be useful as well. In addition our simulations show that the method is robust and accurate for random networks with arbitrary degree sequences; indeed, these results have also been shown to apply to exponential and scale-free networks, as will be explored in more detail elsewhere.

Y. A.-R. and L. S. were supported by the Yeshaya Horowitz Center for Complexity Science, Israel Science Foundation, and an EU-ICT epiwork grant. M. T. was supported by DIP ''Compositionality'' DIP-F 1.2 and the EC NEST program ''MATHfSS''.

*Corresponding author.
lewi@post.tau.ac.il

[1] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[2] M. E. J. Newman, in *Handbook of Graphs and Networks*, edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, Berlin, 2003).
[3] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[4] R. Durrett, *Random Graph Dynamics* (Cambridge University Press, Cambridge, England, 2006).
[5] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[6] E. Volz, Phys. Rev. E **70**, 056115 (2004).
[7] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
[8] M. A. Serrano and M. Boguna, Phys. Rev. Lett. **97**, 088701 (2006).
[9] When $C = 0$ the probability that any of the outgoing edges connects to the original node that we began at, or to any of its other immediate neighbors, scales as $N^{-1}$ and hence can be neglected for large $N$. Similarly, when $C > 0$, the probability to have a cycle of length four, which is not composed of two triangles, scales as $N^{-1}$ and hence can also be neglected for large $N$.
[10] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).
[11] Clustered networks were generated with a number of methods, all giving similar results. Essentially, we first generated a random network with an appropriate degree sequence. Edges were then selectively switched until a specified clustering was reached. B. J. Kim, Phys. Rev. E **69**, 045101(R) (2004); Y. Artzy-Randrup and L. Stone, Phys. Rev. E **72**, 056708 (2005).
[12] H. Andersson, Ann. Appl. Probab. **8**, 1331 (1998).
[13] R. Cohen *et al.*, Phys. Rev. Lett. **85**, 4626 (2000).
[14] D. S. Callaway *et al.*, Phys. Rev. Lett. **85**, 5468 (2000).
[15] Y. Berchenko (to be published).