PHYSICAL REVIEW E **76**, 045101(R) (2007)

Component sizes in networks with arbitrary degree distributions

M. E. J. Newman
Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA
and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

(Received 16 July 2007; published 10 October 2007)

We give an exact solution for the complete distribution of component sizes in random networks with arbitrary degree distributions. The solution tells us the probability that a randomly chosen node belongs to a component of size s for any s. We apply our results to networks with the three most commonly studied degree distributions—Poisson, exponential, and power-law—as well as to the calculation of cluster sizes for bond percolation on networks, which correspond to the sizes of outbreaks of epidemic processes on the same networks. For the particular case of the power-law degree distribution, we show that the component size distribution itself follows a power law everywhere below the phase transition at which a giant component forms, but takes an exponential form when a giant component is present.

DOI: 10.1103/PhysRevE.76.045101 PACS number(s): 89.75.Hc, 02.10.Ox, 64.60.Ak

Methods from physics, and particularly from statistical physics, have proved invaluable for understanding the structure and behavior of networked systems such as the Internet, the world wide web, metabolic networks, protein interaction networks, and social networks of interactions between people [1,2]. In particular, by creating simple (and sometimes not-so-simple) models of network structure and formation, researchers have gained insight into the way networks behave as a function of the basic parameters governing their topology.

One of the most fundamental parameters of a network is its degree distribution. The degree of a node or vertex in a network is the number of edges connected to that vertex, and the frequency distribution of the degrees of vertices has been shown to have a profound influence on almost every aspect of network structure and function, including path lengths, clustering, robustness, centrality indices, spreading processes, and many others. Various network models have been used to illuminate the effects of the degree distribution, but perhaps the most widely studied, and certainly one of the simplest, is the so-called configuration model.

In the configuration model only the degrees of vertices are specified and nothing else; except for the constraint imposed by the degrees, connections between vertices are random. Equivalently, configuration model networks can be thought of as networks drawn uniformly at random from the set of all possible networks whose vertices have the specified degrees. Many properties of the configuration model can be calculated exactly in the limit of large system size, and for this reason the model has become one of the fundamental tools for the quantitative understanding and study of networks. Exact formulas are known for the number of vertices a given distance from a randomly chosen vertex, the existence and expected size of the giant component, the average path length in the giant component, and many other quantities [3,4].

One fundamental result that has been missing, however, is an expression for the sizes of components in the model other than the giant component. More specifically, if we choose a vertex at random from the network, what is the probability that it belongs to a component of a given size? As well as being a central structural property of the network, this distribution is directly related to important practical issues such as the distribution of the sizes of disease outbreaks for diseases spreading over contact networks [5,6].

At first sight, calculation of the component sizes appears difficult. One can derive equations that must be satisfied by the generating function for the distribution of component sizes [4], but usually these equations cannot be solved. Here we show, however, that it is nonetheless possible to derive an explicit expression for the complete distribution of component sizes in the configuration model for a general degree distribution. In particular, we show that it is possible to derive closed-form expressions for component sizes for the three most commonly studied degree distributions: the Poisson, exponential, and power-law distributions. We also show that the same techniques can be used to calculate the sizes of percolation clusters for percolation models on networks of arbitrary degree distribution, a development of some interest because of the close connection between percolation and epidemic processes. We explore this connection in the last part of the paper.

Let p_k be the degree distribution of our network—i.e., the probability that a randomly chosen vertex has degree k. If rather than a vertex we choose an edge and follow it to the vertex at one of its ends, then the number of other edges emerging from that vertex follows a different distribution, the so-called *excess degree distribution*:

$$q_k = \frac{(k+1)p_{k+1}}{\langle k\rangle}, \tag{1}$$

as shown in, for example, Ref. [4]. Here $\langle k\rangle=\Sigma_k k p_k$ is the average degree in the network.

It will be convenient to introduce the probability generating functions for the two distributions p_k and q_k, thus

$$g_0(z) = \sum_{k=0}^{\infty} p_k z^k, \quad g_1(z) = \sum_{k=0}^{\infty} q_k z^k. \tag{2}$$

Many of our results are more easily expressed in terms of these generating functions than directly in terms of the degree distributions. It will also be convenient to note that

$$\langle k\rangle = g_0'(1), \quad g_1(z) = \frac{g_0'(z)}{g_0'(1)}, \tag{3}$$

where we have made use of Eq. (1) in the second equality.

Now let us consider the distribution of the sizes of components in our network. Every vertex belongs to a component of size at least 1 (the vertex itself), and every edge connected to the vertex adds at least one more vertex to the component, and possibly many, if there are lots of other vertices that are reachable via that edge. Let us denote by t the total number of vertices reachable via a particular edge, let the probability distribution of t be ρ_t, and let the generating function for this distribution be $h_1(z)=\Sigma_t \rho_t z^t$.

The probability that a vertex of degree k belongs to a component of size s is the probability that the numbers of vertices reachable along each of its k edges sum to $s-1$. This probability, which we will denote $P(s|k)$, is given by

$$P(s|k) = \sum_{t_1=1}^{\infty} \cdots \sum_{t_k=1}^{\infty} \delta\left(s-1, \sum_{m=1}^{k} t_m\right) \prod_{m=1}^{k} \rho_{t_m}, \tag{4}$$

where $\delta(i,j)$ is the Kronecker delta symbol. Then the probability π_s of a randomly chosen vertex belonging to a component of size s is $\pi_s=\Sigma_{k=0}^{\infty} p_k P(s|k)$ and the corresponding generating function is

$$\begin{aligned} h_0(z) &= \sum_{s=1}^{\infty} \pi_s z^s = \sum_{s=1}^{\infty}\sum_{k=0}^{\infty} p_k P(s|k) z^s \\ &= \sum_{k=0}^{\infty} p_k \sum_{s=1}^{\infty} z^s \sum_{t_1=1}^{\infty} \cdots \sum_{t_k=1}^{\infty} \delta\left(s-1, \sum_{m=1}^{k} t_m\right) \prod_{m=1}^{k} \rho_{t_m} \\ &= z\sum_{k=0}^{\infty} p_k \sum_{t_1=1}^{\infty} \cdots \sum_{t_k=1}^{\infty} z^{\Sigma_m t_m} \prod_{m=1}^{k} \rho_{t_m} = z\sum_{k=0}^{\infty} p_k \left[\sum_{t=1}^{\infty} \rho_t z^t\right]^k \\ &= z\sum_{k=0}^{\infty} p_k [h_1(z)]^k. \end{aligned} \tag{5}$$

But the final sum is simply the generating function $g_0(z)$, Eq. (2), evaluated at $h_1(z)$, and hence

$$h_0(z) = z g_0(h_1(z)). \tag{6}$$

By a similar argument the generating function $h_1(z)$ can be shown to satisfy

$$h_1(z) = z g_1(h_1(z)). \tag{7}$$

Between them, Eqs. (6) and (7) allow us, in principle, to calculate the entire distribution of cluster sizes in our network given the degree distribution p_k. Unfortunately, the self-consistent relation for $h_1(z)$, Eq. (7), is in most cases not solvable and hence we cannot calculate the value of the generating function. Surprisingly, however, we can still calculate the probabilities π_s.

Since every component is of size at least 1, the generating function $h_0(z)$ for the component sizes is of leading order z (or higher) and hence contains an overall factor of z. Dividing out this factor and differentiating, we can write the probability of belonging to a cluster of size s as

$$\pi_s = \frac{1}{(s-1)!}\left[\frac{d^{s-1}}{dz^{s-1}}\left(\frac{h_0(z)}{z}\right)\right]_{z=0}. \tag{8}$$

Using Eq. (6), this can also be written

$$\begin{aligned} \pi_s &= \frac{1}{(s-1)!}\left[\frac{d^{s-1}}{dz^{s-1}} g_0(h_1(z))\right]_{z=0} \\ &= \frac{1}{(s-1)!}\left[\frac{d^{s-2}}{dz^{s-2}}[g_0'(h_1(z))h_1'(z)]\right]_{z=0}. \end{aligned} \tag{9}$$

This expression can be rewritten using Cauchy's formula for the nth derivative of a function,

$$\left.\frac{d^n f}{dz^n}\right|_{z=z_0} = \frac{n!}{2\pi i}\oint \frac{f(z)}{(z-z_0)^{n+1}}dz, \tag{10}$$

where the integral is around a contour that encloses z_0 in the complex plane but encloses no poles in $f(z)$. Applying this formula to Eq. (9) with $z_0=0$ we get

$$\pi_s = \frac{1}{2\pi i(s-1)}\oint \frac{g_0'(h_1(z))}{z^{s-1}}\frac{dh_1}{dz}dz \tag{11a}$$

$$= \frac{\langle k\rangle}{2\pi i(s-1)}\oint \frac{g_1(h_1)}{z^{s-1}}dh_1, \tag{11b}$$

where we have used Eq. (3) to eliminate g_0' in favor of g_1. In Eq. (11a) we choose the contour to be an infinitesimal loop around the origin and, since $h_1(z)$ goes to zero as $z\to 0$, the contour in Eq. (11b) is then also an infinitesimal loop around the origin.

Now regarding z as a function of h_1, rather than the other way around, we make use of Eq. (7) to eliminate z and write

$$\pi_s = \frac{\langle k\rangle}{2\pi i(s-1)}\oint \frac{[g_1(h_1)]^s}{h_1^{s-1}}dh_1. \tag{12}$$

Applying Eq. (10) again we then find that

$$\pi_s = \frac{\langle k\rangle}{(s-1)!}\left[\frac{d^{s-2}}{dz^{s-2}}[g_1(z)]^s\right]_{z=0}. \tag{13}$$

[An alternative and equivalent way to derive this formula—although a less transparent one—would be to rearrange Eq. (7) to give z as a function of h_1 and then apply the Lagrange inversion theorem [7] to derive the Taylor expansion of h_1 or h_0. Indeed, Eqs. (8)–(13) are essentially a proof of a special case of the inversion theorem, as applied to the problem in hand.]

The only exception to Eq. (13) is for the case $s=1$, for which Eq. (11) gives 0/0 and is therefore clearly incorrect. However, since the only way to belong to a component of size 1 is to have no connections to any other vertices, the probability π_1 is trivially equal to the probability of having degree zero:

$$\pi_1 = p_0. \tag{14}$$

Between them, Eqs. (13) and (14) give the entire distribution of component sizes in terms of the degree distribution. They tell us explicitly the probability that a randomly chosen vertex belongs to a component of any given size s. For any specific choice of degree distribution, the application of Eq. (13) still requires us to perform the derivatives. Any finite number of derivatives can always be carried out exactly to give expressions for π_s to finite order. It is also possible in some cases to find a general formula for any derivative and so derive a closed-form expression for π_s for general s. In particular, it turns out to be possible, as we now show, to find such closed-form expressions for the three distributions most commonly studied in the literature: the Poisson, exponential, and power-law distributions.

A network in which edges are placed between vertices uniformly at random has a Poisson degree distribution

$$p_k = e^{-c}\frac{c^k}{k!}, \tag{15}$$

where c is the distribution mean. Such networks have been studied widely for some decades, most famously by Erdős and Rényi [8]. Given Eq. (15), it is straightforward to show that $g_0(z)=g_1(z)=e^{c(z-1)}$ and the derivatives in Eq. (13) can be performed to give

$$\pi_s = \frac{e^{-cs}(cs)^{s-1}}{s!}. \tag{16}$$

(The same expression also works for the special case $s=1$.) This expression for the component size distribution of the Poisson random graph has been derived in the past by a number of other methods—see, for instance, [9] and references therein—but it is a useful check on our methods to see it appear here as a special case of the more general formulation.

Few real-world networks, however, have Poisson degree distributions. Most have highly right-skewed distributions in which most vertices have low degree and a small number of "hubs" have higher degree. A number of networks, for example, are observed to have exponential degree distributions or distributions with an exponential tail. Examples include food webs, power grids, and some social networks [10,11]. Consider the exponential distribution $p_k = C\, e^{-\lambda k}$, where C is the appropriate normalizing constant. The generating functions in this case are

$$g_0(z) = \frac{e^{\lambda}-1}{e^{\lambda}-z}, \quad g_1(z) = \left[\frac{e^{\lambda}-1}{e^{\lambda}-z}\right]^2. \tag{17}$$

Again the derivatives are straightforward to carry out and we find that

$$\frac{d^n}{dz^n}[g_1(z)]^s = \frac{(2s-1+n)!}{(2s-1)!}\frac{[g_1(z)]^s}{(e^{\lambda}-z)^n}, \tag{18}$$

and hence

$$\pi_s = \frac{(3s-3)!}{(s-1)!(2s-1)!}e^{-\lambda(s-1)}(1-e^{-\lambda})^{2s-1}. \tag{19}$$

Applying Stirling's approximation for large s we can show that this distribution behaves asymptotically as $\pi_s \sim s\, e^{-\mu s}$, where $\mu = 2\ln[\frac{3}{2}(1-e^{-\lambda})] - \lambda$. Thus the component size distribution approximately follows an exponential law itself, although with an extra leading factor of s and a different exponential constant.

However, perhaps the greatest amount of attention in recent years has been focused on networks that have power-law degree distributions of the form $p_k \propto k^{-\alpha}$ for some constant exponent α [12,13]. A number of networks appear to follow this pattern, at least approximately, including the world wide web, the Internet, citation networks, and some social and biological networks [1]. The observed value of the exponent typically lies in the range $2<\alpha<3$. Equivalently, we could say that the excess degree distribution q_k—which appears in the fundamental formula (13) via its generating function—follows a power law with exponent $\alpha-1$.

In fact, in essentially all cases, the observed power law holds only in the tail of the distribution; the distribution follows some other law for small degrees. This leaves us considerable latitude about the distribution we use in our calculations. Here we use a so-called Yule distribution for q_k, with a typical real-world value of $\alpha=2.5$ for the exponent:

$$q_k = C\frac{\Gamma(k+\frac{1}{2})}{\Gamma(k+2)}, \tag{20}$$

where $\Gamma(x)$ is the standard gamma function and C is again a normalizing constant. It is straightforward to show (by Stirling's approximation) that this distribution asymptotically follows a power law $q_k \sim k^{-3/2}$, which corresponds to a raw degree distribution $p_k \sim k^{-5/2}$. The Yule distribution appears in a number of contexts in the study of networks, particularly in the solutions of preferential attachment models that may explain the origin of power laws in some networks [14,15], and is considered by some to be the most natural choice of power-law form for discrete distributions. Employing this particular choice for our configuration model gives

$$g_1(z) = \frac{1}{1+\sqrt{1-z}}, \tag{21}$$

which in turn gives

$$\left[\frac{d^n}{dz^n}[g_1(z)]^s\right]_{z=0} = \frac{2^{-(2n+s)}}{(s-1)!}\sum_{j=0}^{n-1}\frac{(n-1+j)!(s+n-1-j)!}{j!(n-1-j)!}. \tag{22}$$

Setting $n=s-2$ and substituting into Eq. (13), we can complete the remaining sum to get

$$\pi_s = [1-\ln 2]^{-1}\frac{(3s-5)!}{(s-1)!(2s-2)!}s2^{3-3s}. \tag{23}$$

In Fig. 1 we show the form of this distribution, along with those for the Poisson and exponential networks, Eqs. (16) and (19). Also shown in the figure are numerical results for the distributions of component sizes measured on computer-generated networks with the same degree distributions. As the figure shows, there is excellent agreement between the simulations and the exact calculations.

As with the exponential network, we can study the asymptotic form of the component size distribution (23) for the power-law network by making use of Stirling's approximation. We find that in the limit of large s, $\pi_s \sim s^3 e^{-\nu s}$, where $\nu = 5\ln 2 - 3\ln 3 \simeq 0.1699\ldots$. Thus again we have an exponential tail to the distribution.

This last result is at first surprising. One might imagine that the component size distribution should itself fall off as a power law or slower because the degree of a vertex provides a lower bound on the size of the component to which the vertex belongs—the fraction of vertices in components of

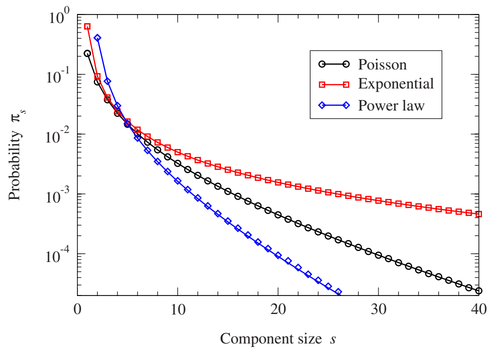

FIG. 1. (Color online) The distribution of component sizes in random graphs with Poisson (c=1.5), exponential (λ=1), and power-law (α=2.5) degree distributions. Solid lines indicate the exact solutions derived in this paper. Points are the results of computer simulations for the same degree distributions. Each point is an average over 5000 networks of 10^6 vertices each. Error bars have been omitted, but are smaller than the data points in each case.

size s or greater must be at least as large as the fraction of vertices of degree s or greater and hence the cumulative distribution of components falls off as slow or slower than the cumulative distribution of degrees.

So how is it possible that we have an exponential distribution of component sizes in the present case? The answer is that we are studying a network that has a giant component. Vertices not in the giant component—which make up almost all of the component size distribution—have a different degree distribution from the graph as a whole because the probability of not being in the giant component dwindles exponentially with increasing degree [6]. This creates an exponential cutoff for the degree distribution, and hence we are back to the situation we had for the exponential network, which gave an exponential component size distribution.

Thus in a power-law network we expect π_s to have an exponential tail whenever there is a giant component in the network, but a power-law tail when there is no giant component. This contrasts with the case for essentially every other degree distribution, where we expect a power-law distribution of component sizes only precisely at the phase transition where the giant component forms; everywhere else, we expect the distribution to fall off exponentially or faster [4].

The methods described here can be extended to the calculation of cluster sizes for percolation processes on networks also. Of particular interest is the bond percolation process, whose cluster sizes give the distribution of outbreaks for a standard susceptible-infective-recovered (SIR) epidemiological process on the same network [5]. Bond percolation can be framed in the same language as the calculation of component sizes above by considering the network formed by just the occupied edges. If the occupation probability is ϕ, then it is straightforward to show [6] that the generating functions for the degree distribution and excess degree distribution of this latter network are $g_0(1-\phi+\phi z)$ and $g_1(1-\phi+\phi z)$, with g_0 and g_1 defined as before. Substituting into Eq. (13), we then find

$$\pi_s(\phi) = \frac{\phi^{s-1}\langle k\rangle}{(s-1)!}\left[\frac{d^{s-2}}{dz^{s-2}}[g_1(z)]^s\right]_{z=1-\phi}. \qquad (24)$$

This implies, contrary to what one might at first expect, that the distribution of cluster sizes need not fall off faster when $\phi<1$: although removing edges from the network necessarily makes clusters smaller, it can paradoxically also make the cluster distribution decay more slowly. As an example, consider again the Poisson degree distribution, Eq. (15), for which Eq. (24) takes the form

$$\pi_s(\phi) = \frac{e^{-cs\phi}(cs\phi)^{s-1}}{s!} = \frac{\pi_s(1)}{\phi}\exp\{[c(1-\phi)+\ln\phi]s\}. \qquad (25)$$

Thus $\pi_s(\phi)$ decays either slower or faster than the distribution for the underlying network depending on whether $c(1-\phi)+\ln\phi$ is positive or negative, with a ϕ-dependent transition between the two regimes when $c=\ln\phi/(\phi-1)$.

To conclude, we have given an exact solution for the distribution of component sizes in random graphs with arbitrary degree distributions and applied it to networks with Poisson, exponential, and power-law distributed degrees. In the latter case we find that though the network has a power-law distribution of component sizes when there is no giant component, the distribution develops an exponential tail once a giant component appears.

The author thanks Cris Moore for useful conversations. This work was funded in part by the National Science Foundation under Grant No. DMS-0405348 and by the Santa Fe Institute.

[1] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[2] M. E. J. Newman, A.-L. Barabási, and D. J. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, 2006).
[3] M. Molloy and B. Reed, Combinatorics, Probab. Comput. **7**, 295 (1998).
[4] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[5] P. Grassberger, Math. Biosci. **63**, 157 (1982).
[6] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).
[7] edited by M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, (Dover, New York, 1974).
[8] P. Erdős and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).
[9] B. Bollobás, *Random Graphs*, 2nd ed. (Academic, New York, 2001).
[10] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11149 (2000).
[11] J. A. Dunne, R. J. Williams, and N. D. Martinez, Proc. Natl. Acad. Sci. U.S.A. **99**, 12917 (2002).
[12] R. Albert *et al.*, Nature (London) **401**, 130 (1999).
[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[14] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000).
[15] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).