PHYSICAL REVIEW E, VOLUME 64, 025101(R)

Giant strongly connected component of directed networks

S. N. Dorogovtsev,[1,2,*] J. F. F. Mendes,[1,†] and A. N. Samukhin[2,‡]

[1]Departamento de Física and Centro de Física do Porto, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal

[2]A.F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia

(Received 30 March 2001; published 19 July 2001)

We describe how to calculate the sizes of all giant connected components of a directed graph, including the *strongly* connected one. In particular, the World Wide Web is a directed network. The results are obtained for graphs with statistically uncorrelated vertices and an arbitrary joint in and out-degree distribution $P(k_i,k_o)$. We show that if $P(k_i,k_o)$ does not factorize, the relative size of the giant strongly connected component deviates from the product of the relative sizes of the giant in- and out-components. The calculations of the relative sizes of all the giant components are demonstrated using the simplest examples. We explain that the giant strongly connected component may be less resilient to random damage than the giant weakly connected one.

DOI: 10.1103/PhysRevE.64.025101 PACS number(s): 05.10.−a, 05.40.−a, 05.50.+q, 87.18.Sn

The giant components of a network are components with relative sizes finite (nonzero) which in the large network limit. The knowledge of these sizes provides the basic information about the global topology of a network. The understanding of the topological structure of networks and its change under external action is the central problem of the statistical physics of random networks [1–8]. Actually, this is the natural generalization of the general percolation theory.

The most interesting networks in nature, including the World Wide Web (WWW), are directed graphs, i.e., their vertices are connected by directed edges [6,9–13]. In the general case, the structure of the directed graph looks as it is shown in Fig. 1 (all the notions are introduced and explained in the figure caption. In particular, the World Wide Web has such a structure [5].

In Refs. [6,7], the previous strong results of mathematicians [14,15] were developed, and the general theory of percolation phenomena in networks with arbitrary degree distributions and statistically uncorrelated (randomly connected) vertices was proposed. Of course, the last assumption is not true for most of the growing nets in nature. Nevertheless, the direct conclusions of such an approach proved to explain the behavior of some real networks [3].

Newman *et al.* [6] have shown how to find the relative sizes of the following giant components of directed graphs with statistically uncorrelated vertices: (i) *the giant weakly connected component* (GWCC), *W*; (ii) *the giant in-component* (GIN), *I*; and (iii) *the giant out-component* (GOUT), *O*. Note that, for brevity and consistency, we use the definitions of the giant in- and out-components rather than those in Refs. [5,6] (see the caption of Fig. 1).

Here we demonstrate how to calculate the relative size *S*, perhaps the most important part of the directed graph, of *its giant strongly connected component* (GSCC) (iv). In the GSCC, every pair of vertices is connected in both directions, i.e., from one of the vertices, one can approach the other by moving either along or against the edge directions. This allows us to completely describe the total structure of directed graphs with arbitrary degree distributions and statistically uncorrelated vertices. For the demonstration, we use the networks with the simplest degree distributions providing nontrivial results.

We recall briefly a very useful approach of Ref. [6]. The *Z*-transforms (or generating functions) are used. For the undirected graph, $\Phi(x)\equiv\Sigma_k P(k)x^k$, and for the directed one, $\Phi(x,y)\equiv\Sigma_{k_i,k_o}P(k_i,k_o)x^{k_i}y^{k_o}$ [17]. Here, $P(k)\equiv P^{(w)}(k)=\Sigma_{k_i}P(k_i,k-k_i)$ is the degree distribution ($k=k_i+k_o$ is the total number of connections of a node) and $P(k_i,k_o)$ is the joint distribution of in- and out-degrees. When all the connections are inside the network, the average in- and out-degrees are equal, $\partial_x\Phi(x,1)|_{x=1}=\partial_y\Phi(1,y)|_{y=1}\equiv z^{(d)}$. Therefore the average degree is $z=2z^{(d)}$. If one ignores the directedness of edges, then the degree distribution of the directed network, in the *Z*-representation, takes the form $\Phi^{(w)}(x)=\Phi(x,x)$. In this case, the distribution of the number of connections minus one of any of the end vertices of a randomly chosen edge corresponds to $\Phi_1^{(w)}(x)\equiv\Phi^{(w)\prime}(x)/z$.

The giant weakly connected component exists if $\Phi_1^{(w)\prime}(1)>1$, which corresponds to the well known criterium of Molloy and Reed [14],

$$\sum_k k(k-2)P(k)>0. \tag{1}$$

The size of the GWCC, *W*, can be easily obtained from the relations [6,15],

$$W=1-\Phi^{(w)}(t_c), \quad t_c=\Phi_1^{(w)}(t_c). \tag{2}$$

From Eq. (1), one sees that the existence of the GWCC crucially depends on the size of the fraction of dead ends in the network. Indeed, $P(1)$ is the only term in Eq. (1) that prevents the GWCC. In Fig. 2(a), the evolution of the giant

*Electronic address: sdorogov@fc.up.pt

†Electronic address: jfmendes@fc.up.pt

‡Electronic address: alnis@samaln.ioffe.rssi.ru

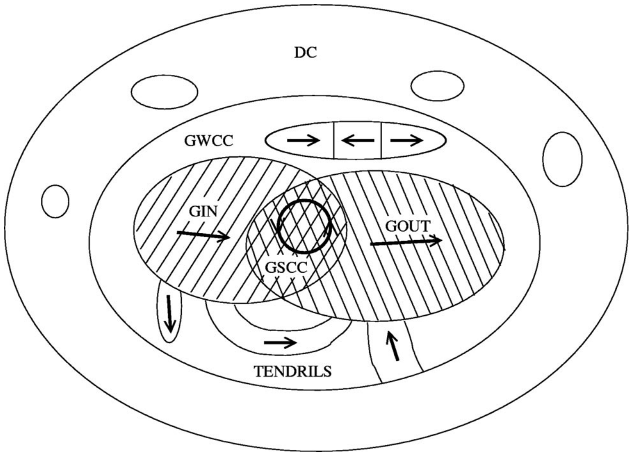

FIG. 1. General structure of a directed network in the situation when the giant strongly connected component is present. Also the structure of the WWW (compare with Fig. 9 of Ref. [5]). If one ignores the directedness of edges, the network consists of the *giant weakly connected component* (GWCC) (actually, the usual percolative cluster) and disconnect components (DC). Accounting for the directedness of edges, the GWCC contains the following components: (a) the *giant strongly connected component* (GSCC), that is the set of vertices reachable from its every vertex by a directed path. (b) the *giant out-component* (GOUT), the set of vertices approachable from the GSCC by a directed path. (c) the *giant in-component* (GIN), contains all vertices from which the GSCC is approachable. (d) the *tendrils*, the rest of the GWCC, i.e., the vertices which have no access to the GSCC and are not reachable from it. In particular, it includes something like ''tendrils'' [5] going out of GIN or coming in the GOUT, but also there are ''tubes'' going from the GIN to the GOUT without passage through GSCC and numerous clusters which are only ''weakly'' connected. Note that the definitions of the GIN and GOUT in the present paper differ from the definitions of Refs. [5,6]. Here the GSCC is included into both the GIN and GOUT, so the GSCC is the interception of the GIN and GOUT. We have to introduce the new definitions for the sake of brevity and logical presentation (see the calculations in the text).

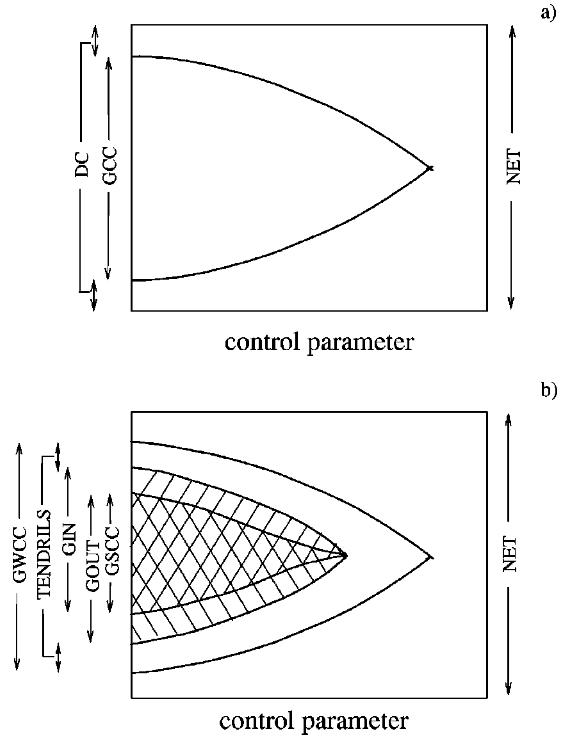

FIG. 2. Schematic plots of the variations of all the giant components vs some control parameter for the undirected network (a) and for the directed one (b). In the undirected graph, the meanings of the giant connected component (GCC), i.e., its percolative cluster, and the GWCC, coincide.

connected component of the undirected graph, induced by the change of some control parameter, is schematically shown. From Eq. (1), it is also clear that the divergency of the second moment of the degree distribution makes the GWCC extremely stable [16]. If the exponent γ of the power-law degree distribution $P(k)\sim k^{-\gamma}$ is $\leqslant 3$, one has to remove at random almost all the vertices or edges of the network to eliminate the GWCC [8].

In a similar way, it is easy to study the GIN and GOUT components of the directed network [6]. One introduces the Z-transform of the out-degree distribution of the vertex, approachable by following a randomly chosen edge when one moves *along* the edge direction, $\Phi_1^{(o)}(y) \equiv \partial_x \Phi(x,y)|_{x=1}/z^{(d)}$. Also, $\Phi_1^{(i)}(x) \equiv \partial_y \Phi(x,y)|_{y=1}/z^{(d)}$ corresponds to the in-degree distribution of the vertex which one can approach moving *against* the edge direction. The GIN and GOUT are present if $\Phi_1^{(i)\prime}(1)=\Phi_1^{(o)\prime}(1) = \partial_{xy}^2 \Phi(x,y)|_{x=1,y=1}/z^{(d)}>1$, that is [6],

$$\sum_{k_i,k_o}(2k_ik_o-k_i-k_o)P(k_i,k_o) = 2\sum_{k_i,k_o}k_i(k_o-1)P(k_i,k_o) = 2\sum_{k_i,k_o}k_o(k_i-1)P(k_i,k_o)>0. \tag{3}$$

In this case, there exist the nontrivial solutions of the equations

$$x_c=\Phi_1^{(i)}(x_c), \quad y_c=\Phi_1^{(o)}(y_c). \tag{4}$$

They have the following meanings: $x_c<1$ is the probability that the connected component, obtained by moving *against* the edge directions starting from a randomly chosen edge, is finite. $y_c<1$ is the probability that the connected component, obtained by moving *along* the edge directions starting from a randomly chosen edge, is finite. Then $P(k_i,k_o)x_c^{k_i}$ and $P(k_i,k_o)y_c^{k_o}$ are the probabilities that a vertex with k_i incoming and k_o outgoing edges have finite in- and out-components, respectively. The in- and out-components of a vertex are sets of vertices that are approachable from this vertex moving against and along the edges, respectively, plus the vertex itself. Summation of these expressions over (k_i,k_o) yields the total probabilities that the in- and out-components of a randomly chosen vertex are finite, respectively. Therefore, they are equal to $\Phi(x_c,1)$ and $\Phi(1,y_c)$, respectively. Thus, the relative sizes of the GIN and GOUT are

$$I=1-\Phi(x_c,1), \quad O=1-\Phi(1,y_c). \tag{5}$$

Here we show that from Eq. (4), it is possible to find not only I and O, but also the relative size S of the GSCC using the considerations similar to Ref. [6]. Suppose that a vertex has k_i incoming and k_o outgoing edges. They are assumed to be statistically independent. Then the probability that all the incoming edges come from finite in-components is $x_c^{k_i}$. The probability that this vertex has the infinite in-component is equal to $1-x_c^{k_i}$, that is, at least one of the k_i incoming edges has to come from the GIN. Similarly, $1-y_c^{k_o}$ is the probability that the vertex has the infinite out-component. The vertex belongs to the GSCC if its in- and out-components are both infinite; the corresponding probability is equal to $(1-x_c^{k_i})\times(1-y_c^{k_o})$. Then the total probability that a vertex belongs to the GSCC is equal to $\Sigma_{k_i,k_o}P(k_i,k_o)(1-x_c^{k_i})(1-y_c^{k_o})$. Finally, the relative size of the GSCC takes the form

$$S=\sum_{k_i,k_o}P(k_i,k_o)(1-x_c^{k_i})(1-y_c^{k_o})$$
$$=1-\Phi(x_c,1)-\Phi(1,y_c)+\Phi(x_c,y_c). \tag{6}$$

Therefore, $\Phi(x_c,y_c)$ is equal to the probability that both the in- and out-components of a vertex are finite. One can write $\Phi(x_c,y_c)=1-W+T$. Knowing W, S, I, and O, it is easy to obtain the relative size of the *tendrils*,

$$T=W+S-I-O. \tag{7}$$

Equations (2) and (4)–(7) allow us to obtain all the giant components of the directed networks with arbitrary joint in- and out-degree distributions and statistically uncorrelated vertices. It is useful to rewrite the main Eq. (6) in the form

$$S=IO+\Phi(x_c,y_c)-\Phi(x_c,1)\Phi(1,y_c). \tag{8}$$

If the joint distribution of in- and out-degrees factorizes, $P(k_i,k_o)=P^i(k_i)P^o(k_o)$, Eq. (8) takes the simple form $S=IO$, otherwise, such factorizing of S is impossible. At the threshold, $x_c=y_c=1$, and I, O, and S simultaneously approach zero.

We have no intention to calculate the sizes of the giant components for real networks, for instance, with the WWW, for the following reasons: (i) They are nonequilibrium (growing); (ii) there are some correlations between their vertices; and (iii) their joint degree distributions are unknown yet (nevertheless, see the attempt of the calculation of I and O for the WWW in Ref. [6]). Instead of this, for demonstration, we consider two of the simplest nontrivial equilibrium nets. In the first of them, the joint in- and out-degree distribution factorizes: $P(k_i,k_o)=P(k_i)P(k_o)$, where $P(k)=[p(\delta_{k,0}+\delta_{k,1})+(1-2p)\delta_{k,3}]$.

The results are the dependence of the sizes of the giant components on p and are shown in Fig. 3 and, schematically, in Fig. 2(b). The curves $I(p)=O(p)$ approach the threshold linearly, and $S(p)$ quadratically, but the range of the quadratic dependence is narrow. Over a wide range of p, $S(p)\approx 1/2-p$ (see also the next case in Fig. 4).

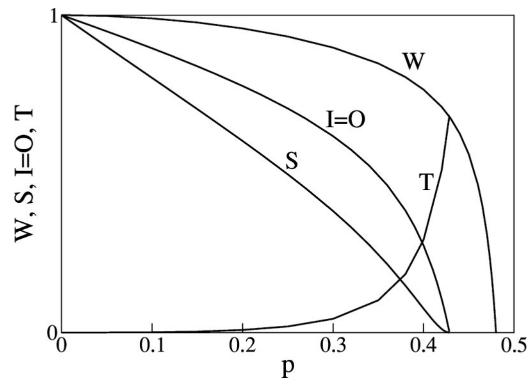

FIG. 3. Relative sizes of the GWCC (W), GSCC (S), GIN (I), GOUT (O), and TENDRILS (T) vs the parameter p for the directed graph with the factorizable joint in- and out-degree distribution $P(k_i,k_o)=P(k_i)P(k_o)$, where $P(k)=[p(\delta_{k,0}+\delta_{k,1})+(1-2p)\delta_{k,3}]$ In this case, $S=IO$, $I=O$.

In real growing nets, the joint in- and out-degree distributions do not factorize just because of their growth [12,18,19]. Therefore, for comparison, we calculate the sizes of the giant components for the network with the distribution $P(k_i,k_o)=p(\delta_{k_i,0}\delta_{k_o,1}+\delta_{k_i,1}\delta_{k_o,0})+(1-2p)\delta_{k_i,3}\delta_{k_o,3}$. This means that large in- and out-degrees correlate, as well as small in- and out-degrees. The results are plotted in Fig. 4 [see also the schematic plot in Fig. 2(b)] [20]. One sees that, in this case, the size of the GSCC noticeably differs from the product IO. We should note that a similar deviation is present in the WWW. From the data of Ref. [5] for the WWW, $I\approx 0.490$, $O\approx 0.489$, so $IO\approx 0.240$, which is less than the measured value $S\approx 0.277$, but is not far from it.

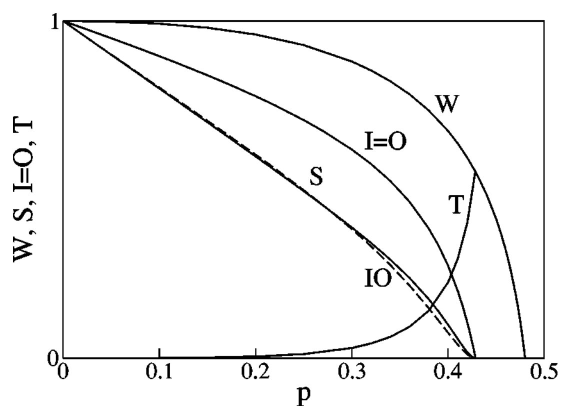

FIG. 4. Relative sizes of the GWCC (W), GSCC (S), GIN (I), GOUT (O), and TENDRILS (T) vs the parameter p for the directed graph with the joint in- and out-degree distribution $P(k_i,k_o)=p(\delta_{k_i,0}\delta_{k_o,1}+\delta_{k_i,1}\delta_{k_o,0})+(1-2p)\delta_{k_i,3}\delta_{k_o,3}$ that does not factorize. This form of the distribution means that if a node is of large in-degree, then its out-degree is also large. Also, if the degree of a node is small, the out-degree is small too. The dashed curve shows the product IO (compare with the curve for S). In the particular case that we consider here, $I=O$.

Figures 2–4 demonstrate that, in the wide enough range of parameters, the following situation may be realized. The directed graphs may have the GWCC and, simultaneously, may not have the GSCC. Only the stability of the GWCC to random damage was discussed yet [8]. Nevertheless, just the stability of the GSCC is the most important, e.g., for the WWW. Is it possible that the GSCC are less resilient to failures than the GWCC? Let us briefly discuss this problem.

One can see from Eq. (3) that the GSCC is extremely resilient if the average $\langle k_i k_o \rangle$ diverges. Let us consider two limiting situations. In the first one, the joint in- and out-degree distribution factorizes, so $\langle k_i k_o \rangle = \langle k_i \rangle^2$. In this case, if the distributions are of a power-law form then, for the robustness of the GSCC, the corresponding exponent γ_i or γ_o should be $\leqslant 2$. This is a very strong requirement. Here, the smallest exponent of γ_i and γ_o is also equal to the exponent γ of the degree distribution. For the resilience of the GWCC to random damage, it should be $\gamma \leqslant 3$ [8], which ensures the divergence of $\langle k^2 \rangle$. Therefore, for such distributions, when $2 < \min(\gamma_i, \gamma_o) = \gamma \leqslant 3$, random damage can destroy the GSCC, but cannot eliminate the GWCC.

In the other limiting case, $P(k_i, k_o) = P(k_i)\delta_{k_i, k_o}$, the correlations between in- and out-degrees are very strong. This form more resembles the joint degree distributions of the real growing networks. In such an event, $\langle k_i k_o \rangle = \langle k_i^2 \rangle$, and the conditions for the resilience of the GSCC and GWCC, $\gamma \leqslant 3$, coincide. One should note that the real distributions are between the considered limiting cases.

In summary, we have shown how to obtain the size of the giant strongly connected component of the directed network with the arbitrary degree distribution and statistically uncorrelated vertices. This allows us to find all the giant components of such a graph and to describe its basic structure. Using the simplest examples and the general considerations we have demonstrated that the correlations between in- and out-degrees subsequently influence the global topology of the network.

S.N.D. thanks PRAXIS XXI (Portugal) for research Grant No. PRAXIS XXI/BCC/16418/98. S.N.D. and J.F.F.M. were partially supported by Project No. POCTI/1999/FIS/33141. We also thank P. L. Krapivsky for useful discussions.

[1] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) (London) **401**, 130 (1999).
[2] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[3] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) (London) **406**, 378 (2000).
[4] R. Albert and A.-L. Barabási, Phys. Rev. Lett. **85**, 5234 (2000).
[5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, in *Proceedings of the 9th WWW Conference* (Elsevier, Amsterdam, 2000), p. 309.
[6] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, e-print cond-mat/0007235.
[7] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. Lett. **85**, 5468 (2000).
[8] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).
[9] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000).
[10] S. N. Dorogovtsev and J. F. F. Mendes, Phys. Rev. E **63**, 056125 (2001); e-print cond-mat/0012009.
[11] B. Tadic, Physica A **293**, 273 (2001).
[12] P. L. Krapivsky, G. J. Rodgers, and S. Redner, Phys. Rev. Lett. **86**, 5401 (2001).
[13] G. Ergun and G. J. Rodgers, e-print cond-mat/0103423.
[14] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).
[15] M. Molloy and B. Reed, Combinatorics, Probab. Comput. **7**, 295 (1998).
[16] Rigorously speaking, to study the resilience of the giant connected components to random damage, one has to consider networks with randomly deleted vertices or edges (see Refs. [7,8]). Nevertheless, the divergence of the moments of the undamaged network already indicates its anomalous resilience, since it cannot be removed by random damage [7,8].
[17] S. Janson, T. Luczak, and A. Rucinski, *Random Graphs* (Wiley, New York, 2000).
[18] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).
[19] P. L. Krapivsky and S. Redner, e-print cond-mat/0011094.
[20] Note that even if the average degree z is large, the size W of the GWCC, in principle, may subsequently deviate from 1. One can easily check this using, e.g., degree distributions similar to the distributions considered above. The Poisson distribution with large z produces extremely small values $1-W \cong \exp[-z]$ since there are few dead ends in this case.