Centrality Scaling in Large Networks

Mária Ercsey-Ravasz[*] and Zoltán Toroczkai[†]

*Interdisciplinary Center for Network Science and Applications (iCeNSA), Department of Physics, University of Notre Dame,
Notre Dame, Indiana, 46556 USA*
(Received 2 March 2010; published 16 July 2010)

Betweenness centrality lies at the core of both transport and structural vulnerability properties of complex networks; however, it is computationally costly, and its measurement for networks with millions of nodes is nearly impossible. By introducing a multiscale decomposition of shortest paths, we show that the contributions to betweenness coming from geodesics not longer than L obey a characteristic scaling versus L, which can be used to predict the distribution of the full centralities. The method is also illustrated on a real-world social network of 5.5×10^6 nodes and 2.7×10^7 links.

Many complex networks are organically evolving without any centralized control or design, and for this reason intense research has been devoted to understand their performance properties and more importantly, their vulnerabilities and failure modes. In these studies, a fundamental role is played by centrality measures (originally introduced in social sciences [1–3]), and, in particular, betweenness centrality [4–7]. Betweenness centrality (BC) of a node (edge) is defined as the fraction of all geodesics (shortest paths) passing through that node (edge). Since transport tends to minimize the cost or time of the route from source to destination, geodesics, and hence centrality measures and their distributions will strongly determine overall transport performance. Interestingly, geodesics are not only important for network flows but also for structural connectivity: removing nodes (edges) with high centrality one obtains a rapid increase in diameter, and eventually the structural breakup of the graph. Analysis of traffic, or information flow [5,7–11], network vulnerability in face of attacks [12], cascading failures [13] or epidemics [14], all involve betweenness calculations.

Unfortunately, computation of betweenness is very costly [10,11,15–18] and for large networks with millions to billions of nodes it is nearly impossible; hence, approximation methods are needed. Existing approximations [19], however, are sampling based, and ill controlled.

Here we show that when geodesics are restricted to a maximum length L, the corresponding range-limited L betweenness (introduced by Borgatti and Everett as bounded-distance betweenness [3]) for large graphs assumes a characteristic scaling form as function of L. This scaling can then be used to predict the betweenness distribution in the (usually unattainable) diameter limit, and with good approximation, to predict the ranking of nodes or edges by betweenness. Additionally, the range-limited method generates l-betweenness values *for all* nodes and edges and *for all* $1 \leq l \leq L$, providing systematic information on geodesics on all length scales. This is of interest in its own right, when the transported entity has a small transmission probability (rumors, viruses) and thus high

attrition rate, not exploring longer geodesics. As we show, the L-betweenness scaling is already achieved for relatively small L values and there is increasingly less new information obtained on BC distribution and ranking when going from L to $L + 1$. The computational overhead, however, involved in the $L \mapsto L + 1$ step is usually immense. The range-limited centrality algorithm presented here, even in the diameter limit ($L = D$), has no larger complexity than the currently known fastest algorithms by Brandes [15] and Newman [16], that is $O(NM)$, where N is the number of nodes and M is the number of (directed) edges, and it is fully parallelizable. For $L < D$ our algorithm runs *sublinearly* in $O(NM)$, making it possible to study networks with millions of nodes. As an illustration, we analyzed a social network (SocNet) inferred from mobile phone trace logs [20] having a giant cluster with $N = 5\,568\,785$ and $M = 26\,822\,764$. For this network we calculated all L-betweenness centralities (L-BCs) for all nodes and edges up to $L = 5$ in 6 days, on 10 processors. With increasing L the ranking of the highest BC nodes freezes and one can predict the top nodes early. The number of geodesics running through these nodes, however, explodes with L. For example, while the node with highest centrality for $L = 4$ has $40\,084\,702$ geodesics, for $L = 5$ it has $500\,903\,498$ of them passing through.

Calculating betweenness centrality of a node or edge in a directed graph $G(V, E)$ requires us to count the number of all-pair shortest directed paths incident on it. Here we include end points; however, the algorithm can easily be changed to exclude them, or produce other variants. The stress centrality (SC) $S(i)$ of a node $i \in V$ is simply the sum of the total number $\sigma_{mn}(i)$ of shortest directed paths from node m to n going through i, $S(i) = \sum_{m,n \in V} \sigma_{mn}(i)$. Betweenness centrality (BC) [4,6] normalizes the number of paths through a node by the total number of paths (σ_{mn}) for a given source-destination pair (m, n): $B(i) = \sum_{m,n \in V} \sigma_{mn}(i)/\sigma_{mn}$. Similar quantities can be defined for an edge $(j, k) \in E$: $S(j, k) = \sum_{m,n \in V} \sigma_{mn}(j, k)$ and $B(j, k) = \sum_{m,n \in V} \sigma_{mn}(j, k)/\sigma_{mn}$.

In order to define range-limited quantities, let $s_l(j)$ and $b_l(j)$ denote the stress and betweenness centralities of a node j for all-pair shortest directed paths of *fixed* length l. Then $S_L(j) = \sum_{l=1}^{L} s_l(j)$ and $B_L(j) = \sum_{l=1}^{L} b_l(j)$ represent centralities from paths not longer than L. Similar measures for an edge are defined in the same way. Just as virtually all centrality algorithms, our method calculates these quantities for a node j for shortest directed paths all emanating from a "root" node i, then it sums the obtained values for all $i \in V$ to get the final centralities for j (similarly for edges). While the basic concept of our algorithm is similar to Brandes' [15] and Newman's [16], we derive recursions that simultaneously compute both SC and BC for *all* nodes and edges and for all values $l = 1, \ldots, L$. The algorithm's output thus generates detailed and systematic information about shortest paths in a graph on all length scales, providing a tool for multiscale network analysis.

The algorithm starts from a given root i and builds the L-range subgraph C_L containing all nodes which can be reached in at most L steps from i. Only links which are part of the shortest paths starting from the root are included in C_L. We decompose C_L into shells $G_l(i)$ containing all the nodes at shortest path distance l from the root, and all incoming edges from shell $l - 1$, Fig. 1(a). The root itself is considered to be shell 0 [$G_0(i)$].

Let $s_l^r(i|j) = \sum_{n \in G_l} \sigma_{in}(j)$ denote the number of shortest directed paths of length l from the root through node j in the rth shell $j \in G_r(i)$, and let $s_l^r(i|j, k) = \sum_{n \in G_l} \sigma_{in}(j, k)$ describe the same quantity for an edge (j, k) in the rth

shell, $(j, k) \in G_r(i)$. We define similar quantities for betweenness, as $b_l^r(i|j) = \sum_{n \in G_l} \sigma_{in}(j)/\sigma_{in}$, and $b_l^r(i|j, k) = \sum_{n \in G_l} \sigma_{in}(j, k)/\sigma_{in}$. Then $s_l(j) = \sum_{i \in V} s_l^r(i|j)$ and $b_l(j) = \sum_{i \in V} b_l^r(i|j)$, with similar equations for edges. In these sums r is not an independent variable. Given i and j, it is the radius of shell $G_r(i)$ centered on i and containing j. One can show that the following recursions hold, (see also Fig. 1):

$$s_l^l(i|j) = \sum_k s_{l-1}^{l-1}(i|k), \qquad b_l^l(i|j) = 1, \qquad (1)$$

$$s_l^{r+1}(i|j, k) = s_l^{r+1}(i|k)s_r^r(i|j)/s_{r+1}^{r+1}(i|k), \qquad (2)$$

$$b_l^{r+1}(i|j, k) = b_l^{r+1}(i|k)s_r^r(i|j)/s_{r+1}^{r+1}(i|k), \qquad (3)$$

$$s_l^r(i|j) = \sum_k s_l^{r+1}(i|j, k), \qquad b_l^r(i|j) = \sum_k b_l^{r+1}(i|j, k). \quad (4)$$

The steps below are repeated for $l = 1, \ldots, L$: (1) Build $G_l(i)$, using a breadth-first search; (2) calculate the l-centrality measures ($s_l^l(i|j)$, $b_l^l(i|j)$) of all nodes in $G_l(i)$; (3) moving backwards, through $r = l - 1, \ldots, 1, 0$, calculate the fixed-l centralities of links in $G_{r+1}(i)$ and of nodes in $G_r(i)$, using recursions (1)–(4). Finally, return to step (1) until the last shell $G_L(i)$ is reached. In the end, we obtained the fixed-l betweenness values of all nodes and edges in C_L. This concludes the basic algorithm, which can be modified to compute different variants of BC and SC, such as excluding end points. Similar recursions can also be derived for load and closeness centrality [5,17].

The L-betweenness values on large networks obey a scaling behavior as function of L. On Fig. 2 we plot the distribution of node betweenness values measured on the Erdős-Rényi random graph (ER) [21], the Barabási-Albert scale-free model (BA) [22], the random geometric graph (RG) [23], and the large social network (SocNet) [20]. Since in large networks B_L grows quickly, it is better to work with the distribution Q_L of the $\ln B_L$ values than with the distribution P_L of B_L values. However, note that $Q_L(\ln B) = BP_L(B)$. As shown on the insets of Fig. 2, the distributions $Q_L(\ln B)$ for different L can be rescaled onto each other by plotting $Q = \sigma_L Q_L$ vs $u = [\ln(B) - \mu_L]/\sigma_L$, where μ_L and σ_L are the mean and the standard deviation for $\ln B_L$. These networks were chosen to represent very different graph classes: the ER, BA, and SocNet have small diameters, while the RG has no shortcuts. The RG is spatially embedded ($d = 2$) unlike ER and BA; the SocNet, however, is influenced by the spatial embedding of people's motility [20]. While BA has a power-law degree distribution $P(k) \sim k^{-3}$, both ER and RG have a Poissonian for $P(k)$, and the SocNet's $P(k)$ resembles a log-normal [24,25]. Both RG and SocNet have high clustering, unlike the others.

Next we show that the scaling behavior observed for range-limited centralities in large graphs is a consequence of the scaling for shell sizes shown to exist, e.g., in random graphs with arbitrary degree distributions [26,27]. Here we present arguments for undirected, uncorrelated graphs and
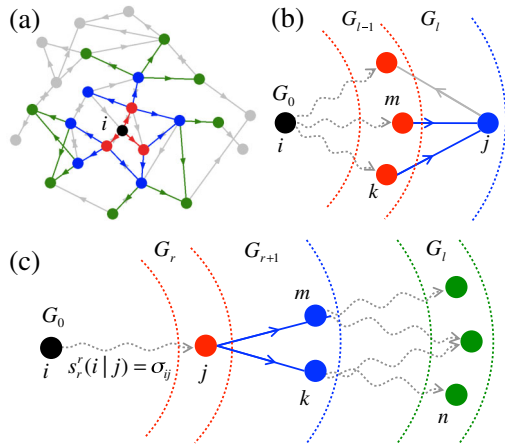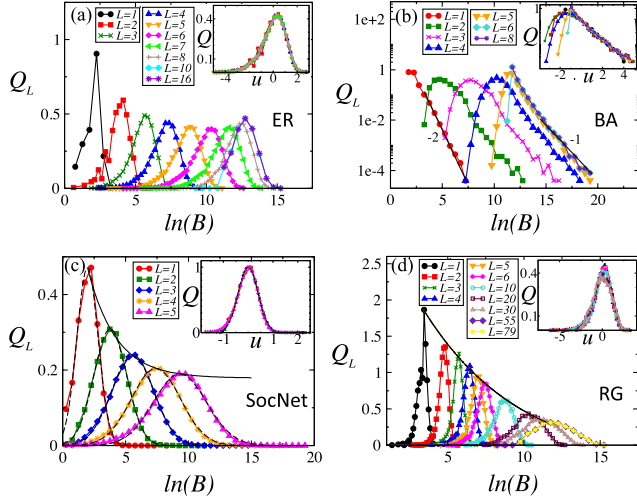

FIG. 1 (color online). (a) Shells of the C_3 subgraph of node i (black) are colored red (medium-light gray), blue (dark gray), green (medium gray). Light gray elements are not part of the subgraph. (b) Eq. (1) calculates SC of a node in G_l [blue (dark gray)] by summing the SC of all its predecessors from $G_{l-1}(i)$ [red (medium-light gray)], e.g., $s_l^l(i|j) = s_{l-1}^{l-1}(i|k) + s_{l-1}^{l-1}(i|m)$. (c) Eqs. (2) and (3) are based on the observations: $\sigma_{in}(j, k) = s_r^r(i|j)\sigma_{kn}$ and $\sigma_{in}(k) = s_{r+1}^{r+1}(i|k)\sigma_{kn}$. Equation (4) calculates the fixed-l centralities for a node [red (medium-light gray)] in $G_r(i)$ by summing the corresponding centralities of its outgoing links [blue (dark gray)] in $G_{r+1}(i)$, e.g., $s_l^r(i|j) = s_l^{r+1}(i|j, k) + s_l^{r+1}(i|j, m)$.

FIG. 2 (color online). Distribution Q_L of L betweeness for different values of L. (a) ER, $N = 5 \times 10^4$, $\langle k \rangle = 4$, diameter $D = 16$, (b) BA, $N = 5 \times 10^4$, $m = 3$, $D = 8$, (c) SocNet, $N = 5\,568\,785$, $M = 26\,822\,764$ and the distributions are fitted by a log-normal (black dashed curves), (d) RG, $N = 10^4$, $\langle k \rangle = 15$, $D = 79$. The insets show the rescaled distributions, see text.

only deal with BC, extensions to other centralities mentioned above being straightforward. Let us define $\langle \cdots \rangle$ as an average over all root nodes i in the graph. If $z_l(i)$ denotes the number of nodes on shell $G_l(i)$, then we model the growth of shell sizes by a branching-like process $z_{l+1}(i) = z_l(i)\alpha_l[1 + \epsilon_l(i)]$, where $\alpha_l = \langle z_{l+1} \rangle / \langle z_l \rangle$ is the branching factor at an lth shell, and $\epsilon_l(i)$ is a per node, shell occupancy noise term, $|\epsilon_l| \ll 1$, considered to obey $\langle \epsilon_l(i) \rangle = 0$ and $\langle \epsilon_l(i)\epsilon_m(i) \rangle = 2A_l \delta_{l,m}$, with A_l decreasing with l, supported by numerical evidence. For undirected paths we can write $b_{l+1}(j) = (1/2)\sum_{i \in V} b_{l+1}(i|j) = z_{l+1}(j) + (1/2) \times \sum_{m=1}^{l} \sum_{i \in G_m(j)} b_{l+1}^m(i|j) \equiv z_{l+1}(j) + (1/2)u_{l+1}(j)$, where we used the fact that in undirected graphs $i \in G_m(j) \Leftrightarrow j \in G_m(i)$. Note that the number of terms in the inner sum $\sum_{i \in G_m(j)} b_{l+1}^m(i|j)$ is $z_m(j)$, which is rapidly increasing with m, and thus $u_{l+1}(j)$ is expected to have a weak dependence on j. Accordingly, we may approximate $u_{l+1}(j) \simeq \sum_{m=1}^{l} \sum_{i \in G_m(j)} v_{l+1}^m(i)$, where $v_{l+1}^m(i)$ is an average betweenness computed on a shell of radius m, *centered on node i*: $v_{l+1}^m(i) = [\sum_{k \in G_m(i)} b_{l+1}^m(i|k)]/z_m(i)$. Based on the observation that $\sum_{k \in G_m(i)} b_l^m(i|k) = z_l(i)$, we can write that $v_{l+1}^m(i) \simeq z_{l+1}(i)/z_m(i)$. Using the recursion defined above for $z_{l+1}(i)$ as a branching process, and neglecting the small noise term, we obtain that $u_{l+1}(j) \simeq \alpha_l \sum_{m=1}^{l} \sum_{i \in G_m(j)} z_l(i)/z_m(i)$. This allows us to write a recursion for $b_{l+1}(j)$ as $b_{l+1}(j) \simeq \alpha_l[b_l(j) + z_l(j)/2 + z_l(j)\epsilon_l(j)]$, which can be iterated down to $l = 1$, where $b_1(j) = z_1(j) = k_j$ is the degree of j:

$$b_l(j) \simeq \beta_l k_j e^{\xi_l(j)}, \quad (5)$$

with $\beta_l = \frac{l+1}{2} \prod_{m=1}^{l-1} \alpha_m = \frac{l+1}{2} \langle z_l \rangle / \langle k \rangle$, and $\xi_l(j) = \sum_{n=1}^{l-1} \frac{l+1-n}{l+1} \epsilon_n(j)$. Equation (5) allows us to relate the sta-

tistics of fixed-l betweenness to the statistics of shell occupancies. Since the noise term (calculated from *per node* occupancy deviations on a shell) is independent on root degree, the distribution of fixed-l betweenness can be expressed as:

$$\rho_l(b) = \frac{1}{b} \int_1^{N-1} dk P(k) \Phi_l(\ln b - \ln\beta_l - \ln k), \quad (6)$$

where $P(k)$ is the degree distribution and $\Phi_l(\xi)$ is the distribution for the noise $\xi_l(j)$, peaked at $\xi = 0$, with fast decaying tails and $\Phi_1(x) = \delta(x)$. From (6) follows that the natural scaling variable for betweenness distribution is $u = \ln b - \ln \beta_l$. An extra l dependence comes from the noise through the width σ_l of Φ_l (for $l > 1$), which can be easily accounted for by the rescaling $u \mapsto u/\sigma_l$, $\rho_l \mapsto \rho_l \sigma_l$, collapsing the distributions for different l values onto the same functional form. As Φ_l is sharply peaked around 0, the most significant contribution to the integral (6) for a given b comes from degrees $k \simeq b/\beta_l$. Since $k \geq 1$, we have a rapid decay of $\rho_l(b)$ in the range $b < \beta_l$, a maximum at $\bar{b} = \beta_l \bar{k}$ where $\bar{k}$ is the degree at which $P(k)$ is maximum, and a sharp decay for $b > (N-1)\beta_l$. In many networks, shell size grows exponentially (ER, AB, and also in the SocNet), that is $\alpha_l \simeq \alpha = \langle z_2 \rangle / \langle k \rangle$, until l reaches the average shortest path distance. This implies that $\beta_l \sim \alpha^l$ and b_l grows exponentially with l [Fig. 3(a)]. In this case, since b_l is rapidly increasing with l, the cumulative $B_L(j) = \sum_l b_l(j)$ will be dominated by the largest l values
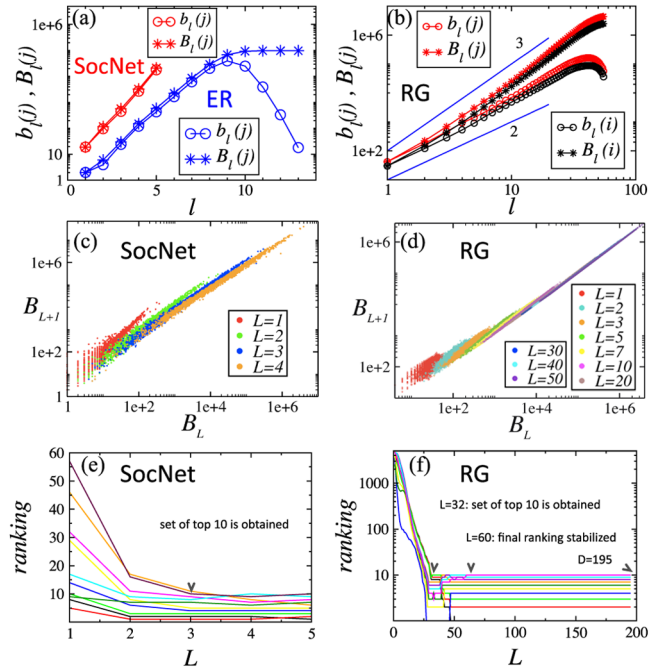

FIG. 3 (color online). (a) b_l (circles) and B_l (stars) vs l for some node j in SocNet [red (medium gray)] and ER [blue (dark gray)]. (b) same as (a) for RG for two arbitrary nodes i and j. B_{L+1} vs B_L for (c) SocNet and (d) RG. Each dot corresponds to a node. Ranking by BC vs L for the top 10 nodes in (e) SocNet and (f) RG (from Fig. 4).
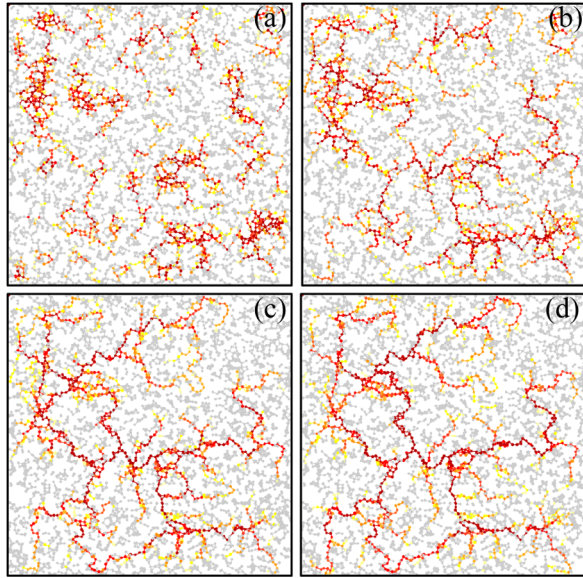
FIG. 4 (color online).   Vulnerability backbone in a RG ($N = 5 \times 10^3$, $\langle k \rangle = 5$) for (a) $L = 5$, (b) $L = 15$, (c) $L = 45$, (d) $L = D = 195$. Darker red indicates nodes with higher B_L. In agreement with Fig. 3(f), the VB is already well approximated at $L = 45$, (c).

and thus, B_L obeys a similar scaling supporting the observations in Fig. 2. For pure scale-free networks $P(k) = ck^{-\gamma}$, and $\rho_l(b) \propto (b/\beta_l)^{1-\gamma}$ for $l > 1$. In networks where the shell size grows as a power law (spatially embedded networks without shortcuts), such as RG, roadways, etc., $\beta_l \sim l^d$, where d is the embedding dimension, $b_l(j) \sim l^d$ and $B_L \sim l^{d+1}$ [Fig. 3(b)].

As the contributions of the noise terms $\epsilon_l(j)$ to $\xi_l(j)$ coming from larger shells are decreasing with increasing l (their weight decreases as $(l + 1)^{-1}$ in addition to the decreasing of their magnitude $|\epsilon_l(j)|$) the $\xi_l(j)$ quantities rapidly converge to a constant. From (5), for a pair of nodes i, j: $\ln[b_l(i)/b_l(j)] = \ln(k_i/k_j) + \xi_l(i) - \xi_l(j)$ showing that their relative ranking by l betweenness freezes with increasing l. Consequently, B_L and B_{L+1} become more correlated with increasing L [Figs. 3(c) and 3(d)] and the ranking of the nodes by their BC also freezes [Figs. 3(e) and 3(f)], allowing early prediction of top betweenness nodes. Spatially embedded networks (RG) without shortcuts represent the worst case, but relative to their diameter the convergence of ranking is still fast [Fig. 3(f)]. An important application of top betweenness predictability is determining the "vulnerability backbone" (VB) of a graph (crucial for network defense purposes [12,14]) which is made by the smallest fraction of highest betweenness nodes forming a percolating cluster through the network. Figure 4 for RG (worst case) shows that the VB (red subgraph) can accurately be predicted already from $L = 45$ betweenness values [Fig. 4(c)] compared to the diameter ($D = 195$) based full betweennesses [Fig. 4(d)].

Finally, we note that the scaling behavior can be used to provide a lower bound L^* of the diameter, from observing

that finite size effects appear when the sum of average shell sizes hits N: $\sum_{l=1}^{L^*} \langle z_l \rangle = \sum_{l=1}^{L^*} \frac{2}{l+1} \beta_l \langle k \rangle \simeq N$. This allows us to find L^* from the scaling behavior of β_l. In particular, for the SocNet $L^* = 10$.

In summary, we have shown that the contributions to centrality measures coming from different length scales of the geodesics exhibit characteristic scaling in large graphs. Exploiting this universal property with the methods presented here makes it possible to predict betweenness values, distributions, and ranking with relatively low computational costs.

*mercseyr@nd.edu
†toro@nd.edu

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge Univ. Press, New York, 1994).
[2] N. E. Friedkin, Am. J. Sociology **96**, 1478 (1991).
[3] S. P. Borgatti and M. G. Everett, Soc. Netw. **28**, 466 (2006).
[4] L. C. Freeman, Sociometry **40**, 35 (1977).
[5] S. P. Borgatti, Soc. Netw. **27**, 55 (2005).
[6] J. M. Anthonisse, Stichting Math. Centr., Amsterdam Report No. BN 9/71, 1971.
[7] S. Sreenivasan *et al.*, Phys. Rev. E **75**, 036105 (2007).
[8] L. Dall'Asta *et al.*, Phys. Rev. E **71**, 036135 (2005).
[9] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).
[10] B. Danila *et al.*, Phys. Rev. E **74**, 046114 (2006).
[11] R. Guimerà *et al.*, Phys. Rev. Lett. **89**, 248701 (2002).
[12] P. Holme *et al.*, Phys. Rev. E **65**, 056109 (2002).
[13] A. E. Motter, Phys. Rev. Lett. **93**, 098701 (2004).
[14] L. Dall'Asta *et al.*, J. Stat. Mech. (2006) P04006.
[15] U. Brandes, J. of Math. Sociology **25**, 163 (2001).
[16] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[17] U. Brandes, Soc. Netw. **30**, 136 (2008).
[18] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).
[19] R. Geisberger *et al.*, in *ALENEX* (SIAM, San Francisco, 2008), p. 90.
[20] M. C. González *et al.*, Nature (London) **453**, 779 (2008).
[21] P. Erdős and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).
[22] A. L. Barabási and R. Albert, Science **286**, 509 (1999).
[23] J. Dall and M. Christensen, Phys. Rev. E **66**, 016121 (2002).
[24] J. P. Onnela *et al.*, Proc. Natl. Acad. Sci. U.S.A. **104**, 7332 (2007).
[25] M. Seshadri *et al.*, in *Proc. of ACM SIGKDD* (ACM, New York, 2008).
[26] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[27] J. Shao *et al.*, Phys. Rev. E **80**, 036105 (2009).