Edge removal balances preferential attachment and triad closing

Hilla Brot,[1] Michal Honig,[3] Lev Muchnik,[2] Jacob Goldenberg,[2] and Yoram Louzoun[1,3,*]
[1]*Gonda Brain Research Center, Bar-Ilan University, Ramat Gan, Israel 52900*
[2]*School of Business Administration, Hebrew University of Jerusalem, Jerusalem, Israel*
[3]*Department of Mathematics, Bar-Ilan University, Ramat Gan, Israel 52900*
(Received 24 January 2013; published 22 October 2013)

Most network formation analysis studies are centered on edge addition. However, edges in real world networks often have a rapid turnover with a large number of edges added and removed between each node addition or removal steps. In such a case, quasiequilibrium is obtained between edge addition and deletion. Edges have been shown to be added to nodes with a high degree and between pairs of nodes with a high number of common neighbors. If not balanced by a degree dependent edge removal, the preference for high degree nodes and node pairs with many common neighbors is expected to increase the average degree of high degree nodes and their clustering coefficient until very large cliques will be formed. Since such large cliques are not formed in real world networks, we conclude that the edge removal probability around high degree nodes and between node pairs with many common neighbors should be higher than around other nodes. We here show the existence of such a balancing mechanism through the relation between the future edge removal probability around nodes and their degree and a similar relation between the edge removal probability and the number of common neighbors of node pairs. In some networks, this preferential detachment process represents an explicit saturation process, and in others, it represents a random deletion process accompanied by a sublinear edge preferential attachment process. A more complex mechanism emerges in directed networks where the preferential detachment can be proportional to the in and out degrees of the nodes involved. In such networks, preferential detachment is stronger for the incoming edges than for the outgoing edges. We hypothesize multiple possible mechanisms that could explain this phenomenon.

DOI: 10.1103/PhysRevE.88.042815 PACS number(s): 89.75.Hc, 05.10.Gg, 05.40.−a, 02.10.Ox

I. INTRODUCTION

The recent extensive usage of social networks, such as Facebook and Twitter has increased our awareness of the effect of network dynamics on our life. In contrast with the initial concept of ever-growing networks, we now realize that most real life networks are dynamic and that edges and nodes are constantly removed and added. Actually, in many networks, the fastest time scale is edge addition and removal, which is much faster than the growth in the number of nodes [1]. In such networks, one can assume that the edges around each node have reached equilibrium in the period between node addition and removal events.

The elements affecting edge addition to networks have been studied experimentally and theoretically [2–4]. In most networks studied, the probability of adding an edge to a node has been shown to increase with the node degree [3,5,6]. This mechanism has been termed "preferential attachment."

We here argue that, in order to balance the preferential attachment of edges to high degree nodes, the edge removal probability cannot be uniform. We sustain this theoretical claim with observations from multiple social networks and show that, even in online social networks, a saturation mechanism exists. Moreover, based on the same argument, we claim that an indirect saturation mechanism must exist even for incoming edges. Such edges are "produced" by other nodes (e.g., hyperlinks pointing to a web page or citations of a paper). This indirect degree dependent edge removal probability can be mediated by a hidden variable, such as the age of the target node.

*Author to whom correspondence should be addressed: louzouy@math.biu.ac.il

Edge deletion properties have received little attention in network research. Most generative network models focus on the edge and node addition processes. Even when edge deletion processes were considered, their goal was to maintain the number of edges in the network by removing edges [7–11] or nodes [2,12,13] randomly. The edge removal process has mainly been studied thoroughly in the special case of network robustness [14–17]. Other models include edge deletion in varying contexts, such as the cost of edges, which limits the amount of edges a single node can have [18–21], the effect of antipreferential attachment (the removal of edges or nodes with a probability inversely proportional to the degree) [22–24], and an optimization model where a degree saturation naturally emerges [25]. The only model that suggested a higher rate of edge removal to high degree nodes [26] was a theoretical model showing the emergence of a power law degree distribution within a certain parametric range.

The effect of edge removal also was studied in sociology in multiple contexts, such as the cost of ties in social networks from an economic point of view [27–29], empirical evidence of the cost and benefits of the maintenance of social ties [30–34], and optimal edge addition and removal strategies in game theory models [21,35–37].

We here show that, in many cases, a degree saturation mechanism must exist based on generic dynamical considerations, even if no explicit saturation mechanism can be proposed.

II. RESULTS

We assume a generic edge creation and destruction process with a fixed number of nodes where edges are added and deleted one at a time (i.e., we ignore the possibility of

 ©2013 American Physical Society

simultaneously adding or removing multiple edges). The process will converge to a stable average degree if the edge addition and deletion rates are equal, but this does not imply that the degree distribution has reached equilibrium. Let $p(k)$ be the probability of a node to have a degree k. The master equations describing the rate of change $p(k)$ are as follows:

$$\begin{aligned}\frac{dp(k)}{dt} &= -p(k)[a(k)+b(k)] + p(k+1)b(k+1) \\ &\quad + p(k-1)a(k-1), \\ a(k) &= p(k \to k+1), \\ b(k) &= p(k \to k-1),\end{aligned} \tag{1}$$

where $a(k)$ and $b(k)$ are the edge addition and removal rates, respectively.

Equation (1) can be approximated by the parallel Fokker Planck equation,

$$\frac{\partial p(k)}{\partial t} = -\frac{\partial p(k)[a(k)-b(k)]}{\partial k} + \frac{1}{2}\frac{\partial^2 p(k)[a(k)+b(k)]}{\partial k^2}. \tag{2}$$

Assuming that $p(k)$ is in equilibrium and that the edge addition and removal rates are much faster than the node addition and removal rates (i.e., a quasisteady state assumption on the nodes), one can approximate that, in the period between node addition and deletion events, $p(k)$ obeys

$$-\frac{\partial p(k)[a(k)-b(k)]}{\partial k} + \frac{1}{2}\frac{\partial^2 p(k)[a(k)+b(k)]}{\partial k^2} = 0. \tag{3}$$

In order for a steady state distribution to appear, $a(k) - b(k)$ has to be negative for $k \to \infty$. Thus, for large enough values of k, $a(k)$ should be smaller than $b(k)$. However, since $a(k)$ has been observed to be a rising function of k for high values of k and the averages of $a(k)$ and $b(k)$ must be equal, $b(k)$ must rise (for a high enough value of k) at a faster rate than $a(k)$.

Note that, if $a(k) - b(k) = 0$, the edge dynamics are a pure diffusion process and Eq. (2) will lead to a divergence of the variance (see Ref. [38] for equivalent dynamics in birth death processes leading to a divergence of the variance). In this case, since the number of edges is limited, the variance can approach its maximal value represented by some nodes being in large cliques and all other nodes having few edges. In order to show that the variance is, indeed, expected to rise, we have performed simulations of a network with a constant number of nodes and edges (4000 nodes and 80 000 edges) and the following dynamics:

$$a(k) = b(k) = ck. \tag{4}$$

As expected, the degree variance keeps increasing (Fig. 1).

There is a difference between the ways edge addition and removal are typically computed. The edge addition probability is computed per node (i.e., what is the probability of adding an edge to a node). The edge removal is computed per edge (i.e., what is the probability of removing an edge). The probability to remove an edge from a node is the node degree multiplied by the probability of removing a random edge. Thus, if the probability of adding an edge to a node is proportional to the degree of the node (as is the case in preferential attachment) and the removal probability of edges is uniform (each edge has an equal probability of being removed), $a(k)$ would be equal to $b(k)$, and as mentioned above, the degree variance would diverge.

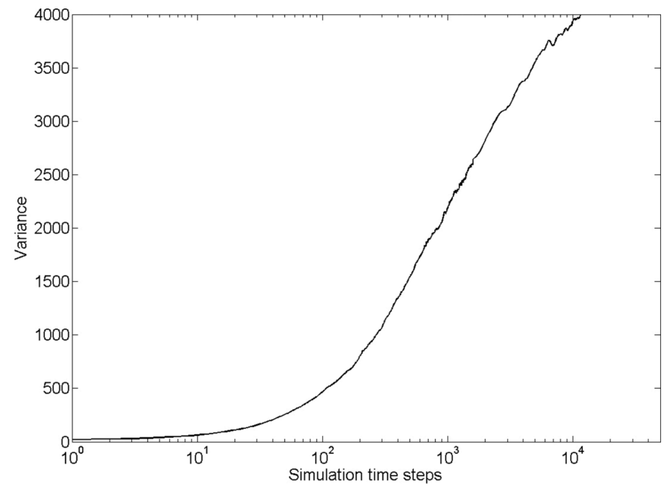

FIG. 1. Degree variance in a preferential attachment model. Variance of degree versus time in a simulation of a network with 4000 nodes, a linear preferential attachment for the edge addition and random edge deletion. The variance of the degree increases with time. This process will eventually lead to a network where most nodes have a low degree and a few nodes produce very large cliques.

Thus, three models are possible:

(a) The edge addition is sublinear, and no saturation mechanism exists (the edge removal probability does not increase with the degree of the nodes surrounding the edge).

(b) The edge addition probability is linear or supralinear, and a saturation mechanism exists.

(c) The edges are not in steady state, and the node addition and removal should be taken into account when studying the edge dynamics.

This last possibility is not the case in the networks examined here since the ratio between the edges and the nodes turnovers is approximately 100 (data not shown).

We here study multiple networks and show that the edge dynamics in networks do obey possibilities (a) or (b). We sampled multiple snapshots in different networks and calculated the degree distributions of pairs of connected nodes that will be disconnected in the next step, normalized by the degree distribution of all existing edges (edges existing in the current step) and compared this distribution with the degree distribution of nodes where an edge will be added, normalized by the node degree distribution [Eq. (5)],

$$a(k) = \frac{\text{Number of edge addition events to nodes with degree } k}{\text{Number of nodes with degree } k}, \tag{5a}$$

$$b(k) = \frac{\text{Number of edge removal events to nodes with degree } k}{\text{Number of nodes with degree } k}. \tag{5b}$$

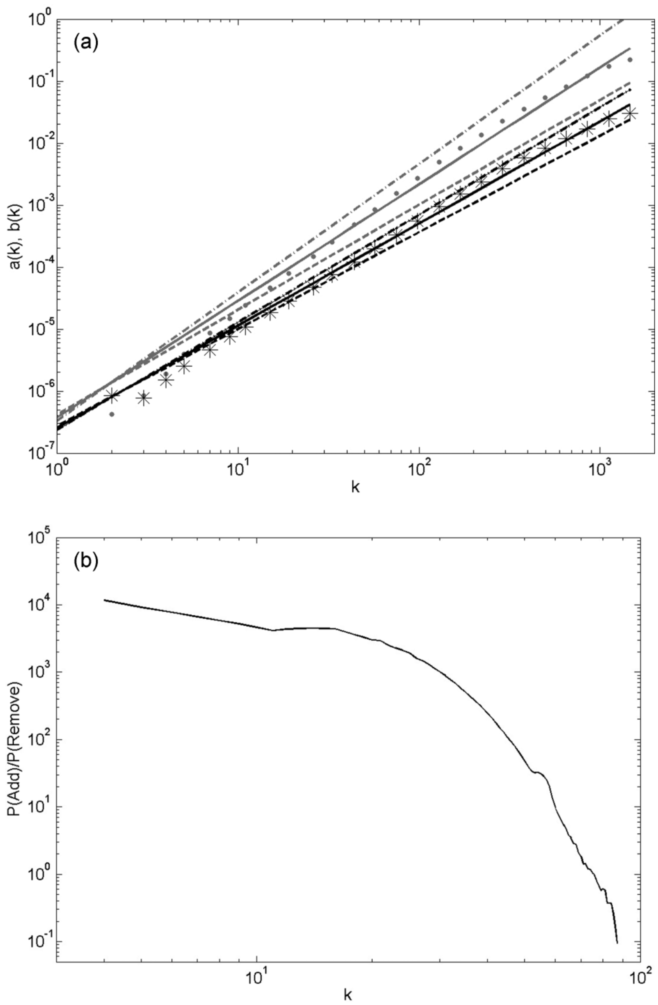

FIG. 2. (a) LJ snapshot 1. Fit of $a(k)$ and $b(k)$ to a scale free distribution. The dots [$b(k)$ removal events] and stars [$a(k)$ addition events] are measurements, and the lines are a scale free regression with the appropriate power given for each fit. The slope of $a(k)$ is 1.65, and the slope of $b(k)$ is 1.88. The dashed lines are confidence intervals. (b) $P(\text{Add})/P(\text{Remove})$ versus k. Ratio between addition and removal probabilities as a function of the degree when the power of the two is similar. For low degrees (k), the addition probability is higher than the removal probability. For high degrees, the addition probability is lower than the removal probability.

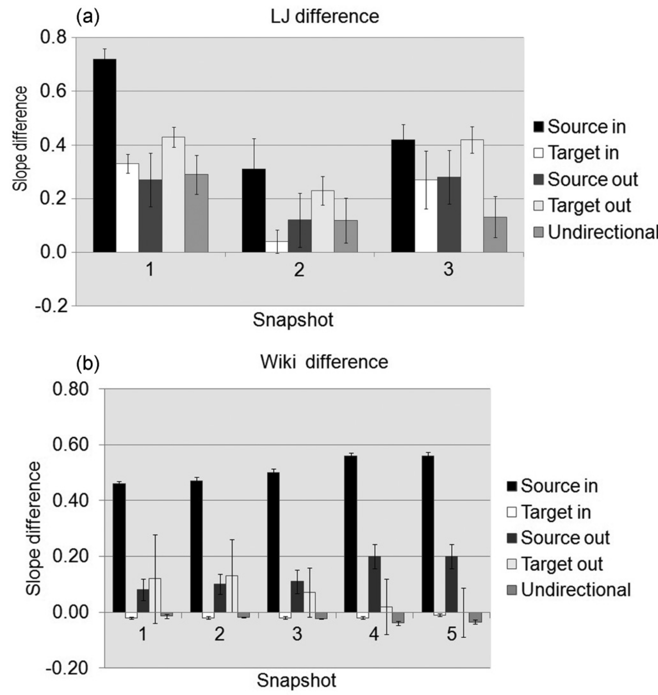

FIG. 3. (a) LJ difference. (b) Wikipedia difference. Difference between the slopes of the power law fit of the edge removal and addition processes for the LiveJournal and Wikipedia networks for the source in, source out, target in, target out, and unidirectional networks. The unidirectional networks are the same snapshots where the direction of the edges is ignored. The x axis is the snapshot number. Most cases show a positive difference between the removal and the addition process powers. Some cases show no difference, but when the addition and removal probabilities are explicitly computed, the removal probability is higher than the addition probability for high values of k.

In order to ensure the generality of the results, we have analyzed the LiveJournal (LJ) blog system and Wikipedia—two sets of evolving networks differing in both the information the networks contain and the way this information is shared among the nodes. These networks differ in their generation and maintenance. While each LJ user can add and remove hyperlinks originating from his or her own profile, the Wikipedia hyperlink network results from a collaborative and distributed effort. We collected four LJ periodic snapshots [39–43] with a 45 day interval and five Wikipedia snapshots [44] in Hebrew (90 000 articles), Russian (450 000 articles), and French (700 000 articles) with 6 month intervals.

In order to quantify this process, for each pair of snapshots, we estimate $a(k)$ as the fraction of nodes with degree k where a new edge has been added and $b(k)$ as the fraction of nodes with degree k where an edge has been removed [see, for example, Fig. 2(a)]. In most cases, a good fit to a scale free function was obtained. One can, thus, compare the slopes by comparing the powers of the distribution. If $a(k)$ rises more slowly than $b(k)$, then the slope of $a(k)$, as defined by the power of the power law, should be lower than the one of $b(k)$. Indeed, for all snapshots of the LJ and Wikipedia networks, the power in the scale free distribution of $b(k)$ was higher than or equal to the one of $a(k)$ (Fig. 3). When the powers are very close, the explicit ratio between $a(k)$ and $b(k)$ can be used, and indeed, the ratio is smaller than 1 for high enough values of k [Fig. 2(b)].

In the case of directed networks, the description is more complex since the in and out degrees can have different effects. Moreover, the in and out degrees can be related. Thus, the edge addition probability should be separated into four components—the in and out degrees of the source and target nodes of each edge (source in, source out, target in, target out). Following the logic mentioned above, we expect that, for all four degree definitions, the edge removal probability would increase at a faster rate with the degree than the edge addition probability. We have computed the slopes of $a(k)$ and $b(k)$ for each such definition, and again, the slope of $b(k)$ is higher than the one of $a(k)$ (see Fig. 3 for three time points in the evolution of the LJ network and five time points in the Hebrew Wikipedia network) with an exception for the out degree of the target node in the Wikipedia network where they are similar.

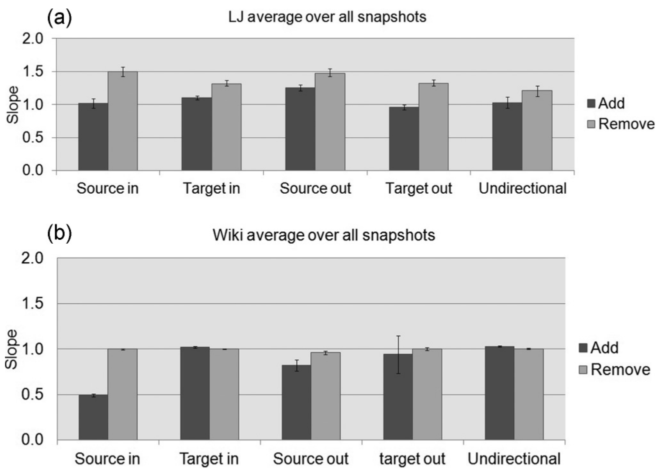

FIG. 4. (a) LJ average over all snapshots. (b) Wikipedia average over all snapshots. Absolute value of slopes of removal and addition probabilities as a function of the degree. Each bar pair is a degree type. (a) is a LJ snapshot, and (b) is the Wikipedia snapshot. All the p values in these graphs are less than 0.01, except from the target in Wikipedia.

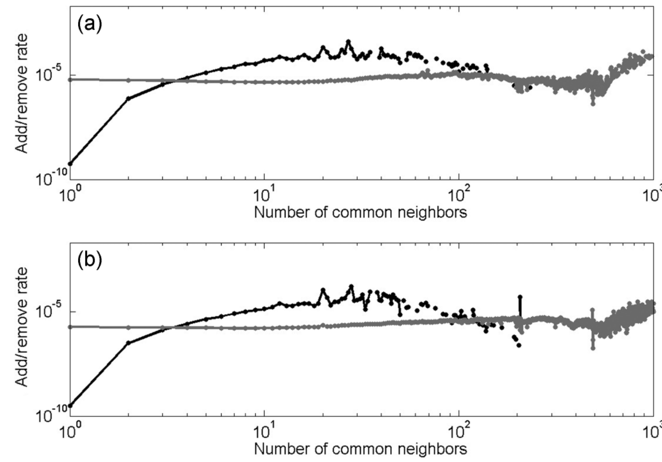

FIG. 5. (a) LJ snapshot 1. (b) LJ snapshot 3. Histogram of the frequency of the number of common friends for pairs of nodes where an edge was created, normalized by the frequency of the number of common friends of pairs in nonconnected pairs and the number of common neighbor distributions in removed edges divided by the number of common neighbor distributions in existing edges. One can see again that, for a high number of common neighbors, the edge removal probability is higher.

However, as mentioned above, when explicitly computing the ratio between $a(k)$ and $b(k)$, the ratio is smaller than 1 for large k values and higher than 1 for low k values.

When the absolute slopes of $b(k)$ and $a(k)$ are computed, a difference emerges between the LJ and the Wikipedia networks. In the Wikipedia network, the edge addition probability for all degree definitions is sublinear, and the edge deletion probability is linear [i.e., no saturation mechanism is needed or observed—Fig. 4(b)]. In LJ [Fig. 4(a)], the edge addition is typically linear or supralinear, and the edge removal is supralinear (an explicit saturation mechanism must exist). The source of the difference is probably the more dynamical aspect of LJ and the fact that each blog is typically written by a single author, in contrast to Wikipedia articles that can be written by multiple authors.

From a mechanistic point of view, there is a fundamental difference between the in degree and the out degree. The out degree is determined by the node, and a saturation mechanism defined on the same node can explain its saturation. The in degree is determined by a large number of other nodes pointing to the node. The saturation mechanism of the in degree must, thus, be a systemic property of the network. Similarly, the observed saturation on the target node must be an indirect mechanism. Indeed, on average, in the LJ (where saturation is observed), the saturation is stronger for the source node, and the saturation for the out degree is stronger than for the in degree (Fig. 4). Still, saturation takes place even for the target node and for the in degree of the source node, implying that the edge removal probability is not only affected by the activity of the person updating the node. Nodes with many incoming edges have a content that increases the probability of removing incoming edges. Such nodes can be, for example, more active nodes [45].

Beyond the degree distributions dynamics, which are characterized by a relation between the degree and the edge addition and removal probability, networks also are characterized by triad closing [46]. A large fraction of the newly added edges are added between nodes with a common neighbor [47–49], leading to the commonly observed high clustering coefficient [3,42,50,51]. However, a cumulative triad closing mechanism [52] will result in the growing edge density within clusters and will culminate eventually with the formation of cliques, unless balanced by the destruction of triangles. Following the logic of the degree dynamics, we expect the edge deletion probability to increase with the number of common neighbors to balance the triad closing.

In order to test for such a saturation mechanism, we compared the probability of adding and removing edges between node pairs with k common neighbors. In contrast with the degree distribution, the common neighbor based addition and removal probability functions have no clear functional shape (Fig. 5). However, following an initial rise, the edge addition probability decreases with the number of common neighbors, whereas, the edge removal probability increases. The increased edge removal probability in node pairs with multiple common neighbors may be partially due to the relation between the degree and the total number of common neighbors. However, it is not purely induced by it since the functional forms are completely different. These measurements were not performed for Wikipedia since very few edges had multiple common neighbors.

III. DISCUSSION

These findings suggest that some empirical evidence from real life social networks for saturation [30–34] may actually represent the general case. They demonstrate that the edge removal process is not a purely random process. Instead, it is a function of the properties of the nodes neighboring the edge. Moreover, in some cases, both the source and the target nodes of an edge exhibit a saturation mechanism with the degree. In bi-directional social networks, based on mutual friendship, this process is obvious since the relationship is symmetric. The situation is different in the investigated networks where the processes of edge deletion and creation depend exclusively on the source node. A few hypotheses may explain this phenomenon: The edge maintenance process may demand a minimum involvement and effort from the target node, and in its absence, the source is likely to delete the edge, or more plausibly, highly active nodes have high edge addition and deletion rates and a high degree. Similarly, node pairs in dense regions may be updated faster, leading to a higher edge removal rate.

These results have important applications beyond the domain of networks. The model studied and the parallel empirical measurements are examples of closed systems with upper bounds on the studied variables (i.e., in this case, the degree and the number of common neighbors).

In systems with no such upper bound, a stable scale free distribution can be obtained from a multiplicative random walk for the relative value $k/\langle k\rangle$. However, such a distribution is a result of a constant rescaling with $\langle k\rangle$ increasing over time [53] (e.g., the relative wealth distribution in a growing economic system). This is similar to the preferential attachment mechanisms proposed in ever-growing networks. However, we now know that networks (and economic systems) can stabilize at an approximately constant size (capital). In such cases, the pure multiplicative random walk will push the distribution to the upper and lower bounds of the variable range, leading to the formation of large cliques and nonconnected nodes, unless explicit saturation mechanisms emerge that either limit the increase rate to be sublinear or make the decrease rate supralinear.

[1] Y. Louzoun, L. Muchnik, and S. Solomon, Bioinformatics **22**, 581 (2006).
[2] J. Davidsen, H. Ebel, and S. Bornholdt, Phys. Rev. Lett. **88**, 128701 (2002).
[3] R. Albert and A. L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[4] H. Jeong *et al.*, Nature (London) **411**, 41 (2001).
[5] H. Jeong, Z. Néda, and A. L. Barabási, Europhys. Lett. **61**, 567 (2003).
[6] Q. Zhou *et al.*, Proc. Natl. Acad. Sci. USA **104**, 16438 (2007).
[7] H. Brot *et al.*, Physica A **391**, 6645 (2012).
[8] J. Cheng, A. W.-c. Fu, and J. Liu, *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, Indianapolis, 2010* (ACM, New York, 2010), p. 459.
[9] K. Park, Y.-C. Lai, and N. Ye, Phys. Rev. E **72**, 026131 (2005).
[10] B. Rudolf, M. Markošová, M. Čajági, and P. Tiňo, Phys. Rev. E **85**, 026114 (2012).
[11] S. Gollapudi, K. Kenthapadi, and R. Panigrahy, *17th International World Wide Web Conference (WWW2008), Workshop on Social Web Search and Mining (swsM2008)*, Citeseer (2008).
[12] M. Cavaliere, S. Sedwards, C. E. Tarnita, M. A. Nowak, and A. Csikász-Nagy, J. Theor. Biol. **299**, 126 (2012).
[13] W. Miura, H. Takayasu, and M. Takayasu, Phys. Rev. Lett. **108**, 168701 (2012).
[14] S. He, S. Li, and H. Ma, Physica A **388**, 2243 (2009).
[15] S. Martin, R. Carr, and J. L. Faulon, Physica A **371**, 870 (2006).
[16] Y. Wang *et al.*, *Proceedings of the 2008 International Conference on Advanced Infocomm Technology* (ACM, New York, 2008).
[17] S. Saavedra, F. Reed-Tsochas, and B. Uzzi, Proc. Natl. Acad. Sci. USA **105**, 16466 (2008).
[18] L. A. N. Amaral *et al.*, PNAS **97**, 11149 (2000).
[19] A. Das Sarma and A. Trehan, *IEEE INFOCOM 2012 31st Annual International Conference on Computer Communications, Orlando, FL, 2012* (IEEE, Piscataway, NJ, 2012).
[20] P. Holme and G. Ghoshal, Phys. Rev. Lett. **96**, 098701 (2006).
[21] J. M. Pacheco, A. Traulsen, and M. A. Nowak, J. Theor. Biol. **243**, 437 (2006).
[22] Q. Chen and D. Shi, Physica A **335**, 240 (2004).
[23] J. S. Kong and V. P. Roychowdhury, Physica A **387**, 3335 (2008).
[24] D. Shi *et al.*, Europhys. Lett. **76**, 731 (2006).
[25] R. M. D'Souza *et al.*, Proc. Natl. Acad. Sci. USA **104**, 6112 (2007).
[26] M. Deijfen and M. Lindholm, Physica A **388**, 4297 (2009).
[27] S. A. Boorman, Bell J. Econom. **6-1**, 216 (1975).
[28] M. O. Jackson and A. Wolinsky, J. Econ. Theory **71**, 44 (1996).
[29] W. W. Sharkey, *Handbooks in Operations Research and Management Science* (Elsevier, New York, 1995), Vol. 8, Chap. 9, p. 713.
[30] J. S. House, D. Umberson, and K. R. Landis, Annu. Rev. Sociol. **14**, 293 (1988).
[31] R. C. Kessler and J. D. McLeod, Am. Sociol. Rev. **49**, 620 (1984).
[32] A. Fliaster and F. Schloderer, Human Relations **63**, 1513 (2010).
[33] M. T. Hansen, J. M. Podolny, and J. Pfeffer, in *Social Capital of Organizations (Research in the Sociology of Organizations)*, edited by S. M. Gabbay and R. T. A. J. Leenders, Vol. 18 (Emerald Group Publishing Limited, Bingley, UK, 2001), pp. 21–57.
[34] R. Cross, A. Parker, L. Prusack, and S. P. Borgatti, Organisational Dyn. **30**, 100 (2001).
[35] J. M. Pacheco, A. Traulsen, and M. A. Nowak, Phys. Rev. Lett. **97**, 258103 (2006).
[36] F. C. Santos, J. M. Pacheco, and T. Lenaerts, PLoS Comput. Biol. **2**, e140 (2006).
[37] K. P. Spiekermann, Synthese **168**, 273 (2009).
[38] L. Pechenik and H. Levine, Phys. Rev. E **59**, 3893 (1999).
[39] S. C. Herring *et al.*, in *40th Annual Hawaii International Conference on System Sciences, 2007, HICSS 2007, Big Island, Hawaii, 2007*, edited by R. H. Sprague, Jr. (IEEE, Piscataway, NJ, 2007), p. 79.
[40] M. Kurucz, A. A. Benczúr, and A. Pereszlényi (unpublished).
[41] I. MacKinnon and R. Warren, *Proceedings ICLM'06 Proceedings of the 2006 Conference on Statistical Network Analysis* (Springer, Berlin, 2007), p. 176.
[42] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[43] P. Zakharov, arXiv:physics/0602063 [physics.soc-ph] (2006).
[44] http://download.wikimedia.org/wikipedia.
[45] L. Muchnik *et al.*, Bull. Am. Phys. Soc. http://meetings.aps.org/link/BAPS.2013.MAR.F28.4BAPS0003–0503.
[46] A. Rapoport, Bull. Math. Biol. **15**(4), 523 (1953).
[47] K. Klemm and V. M. Eguíluz, Phys. Rev. E **65**, 057102 (2002).
[48] E. Ravasz *et al.*, Science **297**, 1551 (2002).
[49] S. N. Soffer and A. Vázquez, Phys. Rev. E **71**, 057101 (2005).
[50] M. E. J. Newman, Phys. Rev. E **64**, 025102 (2001).
[51] D. Watts and S. Strogatz, Nature (London) **393**, 440 (1998).
[52] R. Itzhack, Y. Mogilevski, and Y. Louzoun, Physica A **381**, 482 (2007).
[53] M. Levy and S. Solomon, Int. J. Mod. Phys. C **7**, 595 (1996).