Clustering drives assortativity and community structure in ensembles of networks

David V. Foster,[1,*] Jacob G. Foster,[2] Peter Grassberger,[1,3] and Maya Paczuski[1]
[1] *Complexity Science Group, University of Calgary, Calgary, Canada T2N 1N4*
[2] *Department of Sociology, University of Chicago, Chicago 60615, USA*
[3] *NIC, Forschungszentrum Jülich, D-52425 Jülich, Germany*

(Received 11 January 2011; revised manuscript received 17 May 2011; published 20 December 2011)

Clustering, assortativity, and communities are key features of complex networks. We probe dependencies between these features and find that ensembles of networks with high clustering display both high assortativity by degree and prominent community structure, while ensembles with high assortativity show much less enhancement of the clustering or community structure. Further, clustering can amplify a small homophilic bias for trait assortativity in network ensembles. This marked asymmetry suggests that transitivity could play a larger role than homophily in determining the structure of many complex networks.

DOI: 10.1103/PhysRevE.84.066117 PACS number(s): 89.75.Hc, 05.10.Ln, 64.60.aq, 89.75.Fb

I. INTRODUCTION

Networks provide convenient representations for diverse phenomena across physical, biological, social, technological, and informational domains [1–4]. Just as it is meaningful to "explain" features of real networks with simple generative mechanisms, it is also instructive to ask what features to expect given no other information about a network save that it has a certain set of properties. Such approaches are based on the principle of maximum entropy [5], which finds wide applicability in many fields of science.

Network properties can be markedly interdependent [6–10]. We focus on three key features of undirected networks: (1) the clustering coefficient, C, which reflects the tendency of the network to form triangles (transitivity) [11,12]; (2) the assortativity, r, which reflects the tendency of similar nodes to connect to one another (homophily) [13]; and (3) the modularity, Q, which reflects the tendency of nodes to form tightly interconnected communities [14]. In order to clarify the interdependancies between these quantities in the simplest possible setting, we study them using a maximum entropy approach.

We show that otherwise unbiased ensembles of networks constrained by a transitive bias to be strongly clustered also become highly assortative by degree (hereafter assortative) and modular. In other words, a transitive bias induces an effective bias toward assortativity and modularity. In contrast, ensembles constrained by a homophilic bias to be highly assortative show only weak clustering or modularity. Hence, at the ensemble level a fundamental asymmetry exists between transitivity and homophily. This asymmetry holds unless the distribution of the number of links attached to each node (the node's degree) is extremely broad. Furthermore, a transitive bias can amplify the effect of a homophilic bias toward trait (i.e., race, age, education, etc.) assortativity [15] in network ensembles.

High values for clustering, assortativity, and modularity are often observed in real-world social networks, while nonsocial networks typically have low values [16]. Although extensive social science literature posits homophily to be a dominant force in social network formation (since social networks are highly assortative) [15,17], our results show that a bias for transitive relationships (also called "triadic closure" in the sociology literature [18]) is sufficient to obtain this homophilic effect in network ensembles. Our work is complementary but distinct from that of Newman and Park, which produces the assortativity and clustering generally considered to be characteristic of social networks by introducing modularity [16]. Our work also complements that of Serrano and Boguñá [8–10], which provides techniques for generating networks with desired degree distributions and degree-dependent clustering coefficients. Like us, Serrano and Boguñá find a relationship between assortativity and clustering, showing that degree-degree correlations set an upper limit to clustering [8,9]. Unlike Serrano and Boguñá, we work in a maximum entropy ensemble where the global clustering coefficient and assortativity can be independently controlled; thus we are able to identify the "typical" level of assortativity (resp. clustering) for a given transitive (resp. homophilic) bias.

II. EMPIRICAL NETWORKS

To begin, we note the distinct empirical correlation between C and r in real networks illustrated in Fig. 1. Social networks are (generally) located in the high-C, high-r corner, with nonsocial networks (generally) in the low-C, low-r one. Although such correlations do not, by themselves, imply causality, the pattern suggests an interdependence between the two features that is not limited or reducible to the oft-cited dichotomy between nonsocial and social networks [16]. For instance, consider two networks in Fig. 1: TAP is a high-C, high-r protein-protein interaction network, generated by tandem affinity purification experiments [32]; Y2H is a *weakly* clustered, *disassortative* protein-protein interaction network, generated using yeast two hybridization [33]. The difference in clustering can be explained by a key difference in experimental methodology: TAP pulls out bound complexes and assigns links to every pair of proteins in the complex (making the network highly transitive), while Y2H tests each pair of proteins individually for direct binding. Transitivity has a natural origin in the construction of the TAP network, so it is more likely that the observed assortativity is a byproduct

[*] ventres@gmail.com

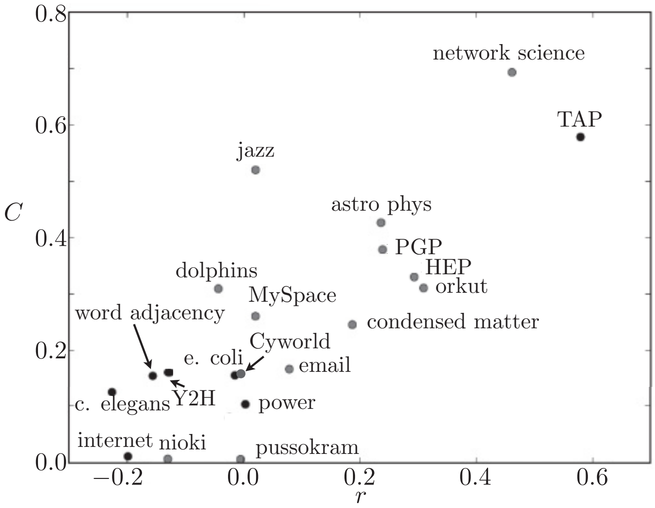

FIG. 1. The relationship between the clustering coefficient, C, and the assortativity, r. The correlation coefficient between r and C for all points is 0.79. Gray points represent social networks, black points represent other types of networks. *Social networks*: astro phys (scientific collaboration) [19]; condensed matter (scientific collaboration) [19]; Cyworld (online social) [20]; dolphins (friendship) [21]; email (communication) [22]; HEP (scientific collaboration) [19]; jazz (musical collaboration) [23]; MySpace (online social) [20]; network science (scientific collaboration) [24]; nioki (online social) [25]; orkut (online social) [20]; PGP (communication network) [26]; pussokram (online dating) [25]. *Nonsocial networks*: *C. elegans* (neural) [27]; *E. coli* (metabolic) [28]; internet (router level) [29]; power (connections between power stations) [11]; TAP (yeast protein-protein binding) [30]; word adjacency (in English text) [24]; Y2H (yeast protein-protein binding) [31].

of an interdependence between transitivity and assortativity rather than an explicit tendency of proteins toward degree homophily.

Since network properties often depend conspicuously on the degree sequence (the number of links attached to each node [34]), we consider ensembles of networks constrained to have the same fixed degree sequence (FDS). Three real world networks are studied in detail: a collaboration network of high-energy physicists (HEP) [19]; a collaboration network of network scientists (NetSci) [24]; and an encrypted communication network (PGP) [26]. We also examine a randomly generated Erdős-Rényi network (ER) [35]. Basic network parameters are given in Table I.

TABLE I. Important values for the networks studied: N is the number of nodes in the network, L is the number of links in the network, r is the assortativity by degree of the network, C is the clustering coefficient of the network, and Q is the modularity of the network.

Name	N	L	r	C	Q	Ref.
ER	19680	41000	−1.3e-5	0.00021	0.246	[35]
HEP	7610	15751	0.29	0.33	0.40	[19]
NetSci	1461	2742	0.46	0.70	0.47	[24]
PGP	10680	24316	0.24	0.38	0.41	[26]

III. REWIRING PROCEDURE AND NETWORK MEASURES

We use a rewiring procedure [36,37] to sample uniformly from each ensemble. At each step of the procedure two links are chosen at random and their endpoints are exchanged, unless this would create a double link, in which case the step is skipped. This move set preserves the degree of each node but otherwise randomizes connections. To sample maximum entropy ensembles with specific features, we use a network Hamiltonian $H(G)$ [38–41] to define an exponential ensemble by assigning a sampling weight $P(G) \propto e^{-H(G)}$ to each graph G. Here we consider ensembles where $H(G)$ depends on C, r, and/or trait assortativity, defined below. Denoting the number of triangles in G by n_Δ, the degree of node i by k_i, and the number of nodes by N, the clustering coefficient is defined as

$$C = \frac{3n_\Delta}{\frac{1}{2}\sum_{i=1}^{N}(k_i - 1)k_i}. \tag{1}$$

Assortativity by degree is defined as the Pearson correlation coefficient between the degrees of nodes joined by a link [13]:

$$r = \frac{L\sum_{i=1}^{L} j_i k_i - \left[\sum_{i=1}^{L} j_i\right]^2}{L\sum_{i=1}^{L} j_i^2 - \left[\sum_{i=1}^{L} j_i\right]^2}, \tag{2}$$

where L is the number of links in the network and j_i and k_i are the degrees of nodes at each end of link i.

Trait assortativity, r_d, measures the tendency for nodes to connect to others with the same discrete trait (e.g., race, gender, etc.) [13]. Following Newman, we define $r_d \propto \sum_\delta e_{\delta\delta}$, where $e_{\delta\delta}$ is the fraction of links in the network from a node of type δ to another node of type δ.

To get ensembles with specific values of C or r we use the following Hamiltonians:

$$H_{C'} = \beta|C' - C_t|, \quad H_{r'} = \beta|r' - r_t|, \tag{3}$$

where C' is the current clustering coefficient and C_t is the target value, and similarly for r' and r_t. The parameter β controls the strength of bias toward the target. It is a transitive bias in $H_{C'}$ and a homophilic bias in $H_{r'}$.

We employ simulated annealing based on a standard Metropolis-Hastings procedure with a rewiring move set [42,43]. One pair of links in the network G is switched to produce a new candidate network G'. A valid move is accepted with probability

$$p = e^{H(G)-H(G')}, \quad p \leqslant 1, \tag{4}$$

and rejected with probability $1 - p$. If $p > 1$ the move is accepted. Initially, the network is rewired 2×10^5 times at $\beta = 0$ to randomize links and avoid strong hysteresis [41]. Then β is increased slowly, rewiring 5×10^4 times after each increase until C (or r) hits C_t (or r_t). The first network with $C = C_t$ ($r = r_t$) is a single sample from the ensemble of networks with a fixed degree sequence and $C = C_t$ ($r = r_t$). The whole process then repeats, starting with the $\beta = 0$ quench.

We also study the influence of transitivity on the trait assortativity, r_d. For this we add a homophilic bias β_d for links between nodes with the same trait, giving us the Hamiltonian:

$$H_d = \beta|C - C_t| + \beta_d \sum_\delta e_{\delta\delta}. \tag{5}$$

Choosing different values of C_t and β_d allows one to explore how the transitive bias impacts trait assortativity at the ensemble level.

We are also interested in the influence of r and C on modularity. Many methods for extracting community structure exist [44,45]. For simplicity, we use the one proposed by Newman and Girvan [14]: Given a partition of the network, e_{ij} is the fraction of all edges connecting a node in community i to one in community j, and $a_i = \sum_j e_{ij}$ is the fraction of all links within community i. The modularity of the network given partition $\mathcal{P}$ is defined as

$$Q_{\mathcal{P}} = \sum_i \left(e_{ii} - a_i^2\right). \tag{6}$$

We use an agglomerative method [46] to approximate the best partition and largest $Q_{\mathcal{P}}$, which we denote Q. While this method has well-known limitations [47], we only need a rough estimate of the modularity to illustrate our point.

Finally, we note that our general strategy could be easily extended to other network structures. For example, in bipartite networks (e.g., sexual contact networks) the standard definition of clustering as given by Eq. (1) is generally no longer sufficient and cycles of length four must be considered [48]. However, the number of such cycles can be controlled using an appropriate Hamiltonian just like the clustering in Eq. (3).

IV. RESULTS

We examine ensembles constrained to have a particular value of r (resp. C) and measure the value for the other feature C (resp. r) averaged over 100 samples from the ensemble. Results are shown in Fig. 2. The grey symbols show the values for ensembles with constrained r, while the black symbols show the values for ensembles with constrained C. Increasing transitivity to increase C has a strong influence on r in all cases, whereas increasing homophily to increase r has relatively little impact on C. The asymmetry is strongest for narrow degree distributions (e.g., the ER network), and becomes less pronounced as the degree distribution broadens.

The asymmetric relationship between r and C can be understood as follows. For nodes to participate in as many transitive relationships as possible, their neighbors must be of similar degree. Hence increasing clustering also increases r, i.e., a transitive bias *induces* an apparent homophilic bias. By contrast, although increasing r leads to links between nodes of similar degree, these relationships need not be transitive. For narrow degree distributions, one could divide all nodes of degree k into two groups and only permit links between the two groups. Assortativity would be maximized without introducing any clustering. For networks with very broad degree distributions (like PGP), only a few nodes of high degree exist, but they have a large effect on r. Hence for ensembles constrained to have a large r, the highest degree nodes are under strong pressure to link, thus creating transitive relationships between their many neighbors. Note that most social networks do not have broad degree distributions. In such cases homophily has only a weak influence on C at the ensemble level.

Figure 2 also indicates the C and r values for the real-world networks (dashed lines). Ensembles of networks constrained

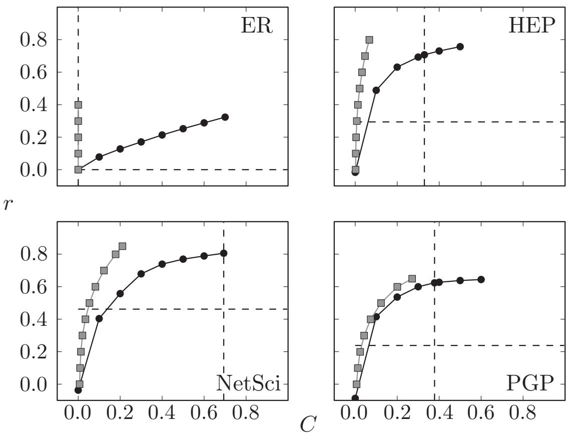

FIG. 2. Controlling assortativity (grey symbols) vs controlling clustering coefficient (black symbols) for various network degree sequences. C is on the x axis, r on the y axis. Each point represents the average value of 100 samples drawn from an ensemble with the specified r or C value. The dashed lines show the values of r and C for the original network. Note the asymmetry between the effect of C on r compared to r on C. Error bars are much smaller than the marker size.

to have the same C as the real network exhibit far greater r. Hence, social networks are actually *disassortative* relative to the ensemble of networks with the same clustering coefficient and degree sequence [49], an insight only possible using a maximum entropy ensemble approach. Indeed, the most likely way to create many triangles is to densely interconnect the higher degree nodes so triangles clump together (as discussed in Ref. [41]; see also the distinction between weak and strong clustering introduced in Refs. [9,10]). Real social networks seem to spread clustering more evenly across the network, thus lowering r. For example, in scientific collaboration networks supervisory relationships may decrease the assortativity by creating links between lower degree students and higher degree professors.

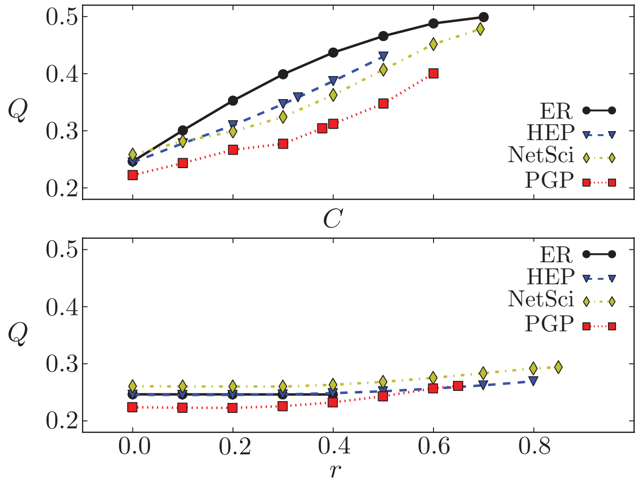

FIG. 3. (Color online) Modularity Q for various ensembles of networks with different target values for C (top row) or r (bottom row). Clustering has a much larger impact on modularity than assortativity does. Error bars are much smaller than the marker size.

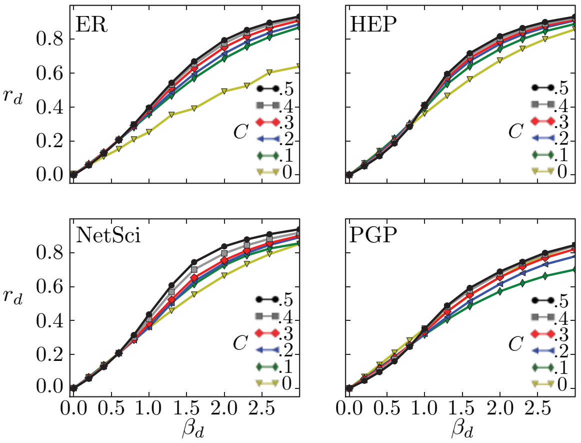

FIG. 4. (Color online) Trait assortativity r_d (y axis) for ensembles of networks with varying C (indicated in the legend) and homophilic bias β_d (x axis). For narrow degree distributions, clustering amplifies the response of trait assortativity to a homophilic bias. For broad degree distribution the opposite occurs for small β_d. Error bars are much smaller than the marker size.

We also measured the influence of r and C on modularity. The top (resp. bottom) panel in Fig. 3 shows the average Q in ensembles with constrained C (resp. r). Clustering (and hence transitivity) has a more pronounced effect on modularity than does assortativity (and hence homophily). The modularity achieved for the highly clustered ensembles approximates the actual modularity for the real networks (HEP, NetSci, and PGP; see Table I), unlike assortative ensembles without a transitive bias.

Finally, we consider the effect of transitivity on trait assortativity, r_d. For each of the degree sequences, we create ensembles of networks with different target C values and varying homophilic biases β_d. Since the actual data sets do not contain trait values, we assign each node one of three possible traits at random with equal probability. For ER, HEP, and NetSci we observe that ensembles with larger C enhance r_d relative to ensembles with the same homophilic bias but no clustering ($C = 0$); see Fig. 4. This is especially clear for the narrowest (ER) degree sequence. For the PGP network, which has a broad degree distribution, clustering appears to compete with the homophilic bias (e.g., the curves cross), leading to a more complicated scenario. The interdependence between clustering and trait assortativity thus appears to depend on the degree sequence, but for narrow degree sequences the positive relationship holds and transitivity enhances the effect of a homophilic bias. We also note that increasing the trait assortativity of an otherwise unconstrained ensemble had no impact on C, r, or Q (data not shown).

V. CONCLUSIONS

On the basis of these results for (maximum entropy) ensembles of networks, we conjecture that the widely discussed dichotomy between assortative social networks and disassortative *nonsocial* networks could be a result of a deeper dichotomy between networks with and without transitive relationships. As shown here, transitivity typically leads to assortativity at an ensemble level; hence networks with transitive relationships (like many social networks) will also tend to be assortative. This proposal can explain the high assortativity of TAP, and is consistent with another observation in Fig. 1: several online social networks show low clustering and low assortativity [50]. If assortative mixing by degree in social networks is the result of homophily, this anomaly is hard to explain: Why should popular people stop seeking each other out simply because the network is online? But if assortativity is driven by transitivity, the “anomaly” disappears: in the absence of spatially mediated interactions online, a smaller tendency may exist to introduce mutual friends.

However, we cannot make strong claims about causality, nor have we ruled out the scenario in Ref. [16]. Indeed, the causal factors driving network evolution could be complex, multifaceted, and idiosyncratic. Nevertheless, our results on the asymmetric dependencies between clustering, assortativity, and modularity at the ensemble level provide an additional warning about inferring causality from naive observations of network structure.

[1] A. Broder *et al.*, Comp. Netw. **33**, 309 (2000).
[2] S. Boccaletti *et al.*, Phys. Rep. **424**, 175 (2006).
[3] A. Barabási and Z. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[4] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[5] E. T. Jaynes, Phys. Rev. **106** (1957).
[6] S. N. Soffer and A. Vazquez, Phys. Rev. E **71**, 057101 (2005).
[7] P. Holme and J. Zhao, Phys. Rev. E **75**, 046111 (2007).
[8] M. A. Serrano and M. Boguñá, Phys. Rev. E **72**, 036133 (2005).
[9] M. A. Serrano and M. Boguñá, Phys. Rev. E **74**, 056114 (2006).
[10] M. A. Serrano and M. Boguñá, Phys. Rev. E **74**, 056115 (2006).
[11] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[12] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
[13] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[14] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[15] G. Kossinets and D. Watts, Am. J. Sociol. **115**, 405 (2009).
[16] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).
[17] M. McPherson, L. Smith-Lovin, and J. Cook, Annu. Rev. Sociol. **27**, 415 (2001).
[18] A. Rapoport, Bull. Math. Biophys. **15**, 523 (1953).
[19] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).
[20] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, *Proceedings of the 16th international conference on World Wide Web* (Organization ACM, 2007).
[21] D. Lusseau *et al.*, Behav. Ecol. Sociobiol. **54**, 396 (2003).
[22] R. Guimerá, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103 (2003).
[23] P. M. Gleiser and L. Danon, Adv. Complex Syst. **6**, 565 (2003).
[24] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[25] P. Holme, C. R. Edling, and F. Liljeros, Soc. Networks **26**, 155 (2004).

[26] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys. Rev. E **70**, 056122 (2004).

[27] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).

[28] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. Barabási, Nature **407**, 651 (2000).

[29] M. E. J. Newman, "Network data", http://www-personal.umich.edu/~mejn/netdata/.

[30] A. C. Gavin *et al.*, Nature **415**, 141 (2002).

[31] H. Jeong, S. P. Mason, A. Barabási, and Z. N. Oltvai, Nature **411**, 41 (2001).

[32] O. Puig *et al.*, Methods **24**, 218 (2001).

[33] S. Fields and O. Song, Nature **340**, 245 (1989).

[34] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).

[35] P. Erdős and A. Rényii, Publ. Math. (Debrecen) **6**, 290 (1959).

[36] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[37] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski, Phys. Rev. E **76**, 046112 (2007).

[38] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).

[39] J. Berg and M. Lässig, Phys. Rev. Lett. **89**, 228701 (2002).

[40] J. Park and M. E. J. Newman, Phys. Rev. E **70**, 066117 (2004).

[41] D. V. Foster, J. G. Foster, M. Paczuski, and P. Grassberger, Phys. Rev. E **81**, 046115 (2010).

[42] W. K. Hastings, Biometrika **57**, 97 (1970).

[43] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, Oxford, United Kingdom, 1999).

[44] M. A. Porter, J.-P. Onnela, and P. J. Mucha, Notices of the AMS **56**, 1082 (2009).

[45] S. Fortunato, Phys. Rep. **486**, 75 (2010).

[46] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).

[47] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).

[48] P. G. Lind, M. C. González, and H. J. Herrmann, Phys. Rev. E **72**, 056127 (2005).

[49] J. G. Foster, D. V. Foster, M. Paczuski, and P. Grassberger, Proc. Natl. Acad. Sci. USA **107**, 10815 (2010).

[50] H. Hu and X. Wang, Europhys. Lett. **86**, 18003 (2009).