Clustering Drives Assortativity and Community Structure in Ensembles of Networks

David V. Foster,1, 2 Jacob G. Foster, Peter Grassberger,1, 3 and Maya Paczuski1
1Complexity Science Group, University of Calgary, Calgary T2N 1N4, Canada
2Department of Sociology, University of Chicago, Chicago 60615, USA
3NIC, Forschungszentrum Jülich, D-52425 Jülich, Germany
(Dated: January 6, 2011)

Clustering, assortativity, and communities are key features of complex networks. We probe dependencies between these attributes and find that ensembles with strong clustering display both high assortativity by degree and prominent community structure, while ensembles with high assortativity are much less biased towards clustering or community structure. Further, clustered networks can amplify small homophilic bias for trait assortativity. This marked asymmetry suggests that transitivity, rather than homophily, drives the standard nonsocial/social network dichotomy.

PACS numbers: 89.75.Hc, 05.10.Ln, 89.75.Fb, 64.60.aq

Ensembles of networks constrained by a homophilic bias to be highly assortative and modular. 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 towards trait (i.e. race, age, education, etc.) assortativity in network ensembles.

High values for the clustering, assortativity, and modularity are often observed in real-world social networks, while nonsocial networks may have low values. Although extensive social science literature posits homophily to be a dominant force in social network formation (since social networks are highly assortative), our results show that a bias for transitive relationships (also called “triadic closure” in sociology literature) is sufficient to obtain this effect in network ensembles. Our work is complementary to that of Newman and Park who produce assortativity and clustering characteristic of social networks by introducing modularity.

FIG. 1: The relationship between the clustering coefficient, $C$, and the assortativity, $r$. Gray points represent social networks, black points represent other types of networks. Social networks: astro phys (scientific collaboration) [15]; condensed matter (scientific collaboration) [15]; MySpace (online social) [16]; dolphins (friendship) [17]; email (communication) [18]; HEP (scientific collaboration) [15]; jazz (musical collaboration) [19]; MySpace (online social) [16]; network science (scientific collaboration) [20]; pussokram (online dating) [21]. Non-social networks: c. elegans (neural) [22]; c. coli (metabolic) [24]; internet (router level) [25]; power (connections between power stations) [7]; TAP (yeast protein-protein binding) [26]; word adjacency (in English text) [20]; Y2H (yeast protein-protein binding) [27].

To begin, we note a distinct empirical correlation between $C$ and $r$ in real networks illustrated in Fig. 1 with social networks (generally) in the high $C$, high $r$ corner, and non-social networks (generally) in the low $C$, low $r$ one. The pattern suggests an interdependence between the two features that transcends a simple nonso-
TABLE I: Important values for the empirical networks

| Name  | N   | L   | r  | C   | Q   | Ref |
|-------|-----|-----|----|-----|-----|-----|
| ER    | 19680 | 41000 | -1.3e-5 | 0.00021 | 0.246 | [31] |
| HEP   | 7610 | 15751 | -29 | 0.33 | 0.40 | [15] |
| NetSci | 1461 | 2742 | -46 | 0.70 | 0.47 | [20] |
| PGP   | 10680 | 24316 | -24 | 0.38 | 0.41 | [22] |

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 $i$.

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

$$H_C = \beta |C - C_t|, \quad H_r = \beta |r - r_t|$$

where $C'$ is the current clustering coefficient and $C_t$ is the target value, and similarly for $r'$. The parameter $\beta$ controls the strength of bias towards 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 [38,39]. One pair of links in the network $G$ is switched to produce a new candidate network $G'$. A valid move is accepted with probability

$$p = e^{H(G) - H(G')} \leq 1$$

and rejected with probability $1 - p$. If $p > 1$ the move is accepted. Initially, the network is rewired $2 \times 10^5$ times at $\beta = 0$ to randomize links and avoid strong hysteresis [37]. Then $\beta$ is increased slowly, rewiring $5 \times 10^4$ 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 trait assortativity, $r_d$, which measures the tendency for nodes to connect to others with the same discrete trait (e.g. race, gender, etc.) [10]. For this we add a homophilic bias $\beta_d$ for links between nodes with the same trait. Defining $r_d \propto \sum_\delta e_{\delta\delta}$, where $e_{\delta\delta}$ is the fraction of links in the network from a node of type $\delta$ to another node of type $\delta$, the Hamiltonian becomes

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

Choosing different values of $C_t$ and $\beta_d$ allows one to explore how transitivity impacts trait assortativity at the ensemble level.

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 (resp. black) symbols show the values for ensembles with constrained $r$ (resp. $C$). 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, but still apparent, 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 neighbours must be of similar degree. Hence increasing clus-
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 average values from 100 samples from an ensemble with specified $r$ or $C$ values. 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 $C$.

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.

is the fraction of all links within community $i$. The modularity of the network given partition $\mathcal{P}$ is defined as:

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

(6)

We use an agglomerative method [43] to approximate the best partition and largest $Q_{\mathcal{P}}$, which we denote $Q$.

The top (resp. bottom) panel in Fig. 3 shows the average $Q$ in ensembles with constrained $C$ (resp. $r$). Transitivity has a more pronounced effect on modularity than does 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 $\beta_d$. 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 ($C = 0$). 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 homophilic bias. We also note that increasing the trait assortativity of an ensemble had no impact on $C$, $r$, or $Q$ (data not shown).

We conjecture that the standard nonsocial/social (dis-
assortative/assortative) dichotomy is driven by transitive relationships in many social networks, such as in scientific collaborations. As shown here, transitivity typically leads to assortativity. This explains the anomalous position of TAP located within social networks, and is consistent with another anomaly in Fig. 1 several online social networks show low clustering and low assortativity [44]. If assortative mixing by degree is the result of homophily by degree in social networks, this anomaly is hard to explain: why should popular people stop seeking each other out simply because the social network moved online? But if assortativity is a side-effect of transitivity, this effect is easier to understand: it is plausible that online social relationships are less transitive, since in the absence of spatially mediated interactions there is a smaller tendency to introduce mutual friends. We have not ruled out the scenario in reference [12]. Indeed, the causal factors driving network evolution are likely to be complex, multifaceted, and idiosyncratic. Our results on the asymmetric dependencies between clustering, assortativity, and modularity provide a 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. Genet. 5, 101 (2004).
[4] M. E. J. Newman, SIAM Review 45, 167 (2003).
[5] S. N. Soffer and A. Vázquez, Phys. Rev. E 71, 057101 (2005).
[6] P. Holme and J. Zhao, Phys. Rev. E 75, 046111 (2007).
[7] D. J. Watts and S. H. Strogatz, Nature 393, 440 (1998).
[8] M. E. J. Newman, Phys. Rev. E 68, 026121 (2003).
[9] M. E. J. Newman, Phys. Rev. Lett. 89, 208701 (2002).
[10] M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004).
[11] G. Kossinets and D. Watts, AJS 115, 405 (2009).
[12] M. E. J. Newman and J. Park, Phys. Rev. E 68, 036122 (2003).
[13] M. McPherson, L. Smith-Lovin, and J. Cook, Annu. Rev. Sociol. 27, 415 (2001).
[14] A. Rapoport, Bull. Math. Biol. 15, 523 (1953).
[15] M. E. J. Newman, PNAS 98, 404 (2001).
[16] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, in Proceedings of the 16th international conference on World Wide Web (ACM, 2007).
[17] D. Lusseau et al., Behav. Ecol. Sociobiol. 54, 396 (2003).
[18] R. Guimerà, L. Danon, A. Daz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E 68, 065103 (2003).
[19] P. M. Gleiser and L. Danon, Adv. Complex Sys. 6, 565573 (2003).
[20] M. E. J. Newman, Phys. Rev. E 74, 036104 (2006).
[21] P. Holme, C. R. Edling, and F. Liljeros, Soc. Networks 26, 155 (2004).
[22] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys. Rev. E 70, 056122 (2004).
[23] J. Duch and A. Arenas, Phys. Rev. E 72, 027104 (2005).
[24] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. Barabási, Nature 407, 651 (2000).
[25] M. E. J. Newman, Network data, http://www-personal.umich.edu/~mejn/netdata/.
[26] A. C. Gavin et al., Nature 415, 141 (2002).
[27] H. Jeong, S. P. Mason, A. Barabási, and Z. N. Oltvai, Nature 411, 41 (2001).
[28] O. Puig et al., Methods 24, 218 (2001).
[29] S. Fields and O. Song, Nature 340, 245 (1989).
[30] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E 64, 026118 (2001).
[31] P. Erdős and A. Rényii, Publ. Math. Debrecen 6, 156 (1959).
[32] S. Maslov and K. Sneppen, Science 296, 565573 (2002).
[33] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski, Phys. Rev. E 76, 066117 (2007).
[34] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science 298, 824 (2002).
[35] J. Berg and M. Lässig, Phys. Rev. Lett. 89, 228701 (2002).
[36] J. Park and M. E. J. Newman, Phys. Rev. E 70, 066117 (2004).
[37] D. V. Foster, J. G. Foster, M. Paczuski, and P. Grassberger, Phys. Rev. E 81, 046115 (2010).
[38] W. K. Hastings, Biometrika 57, 97 (1970).
[39] M. E. J. Newman and G. T. Barkema, Monte Carlo methods in statistical physics (Oxford Univ. Press, 1999).
[40] J. G. Foster, D. V. Foster, M. Paczuski, and P. Grassberger, P. Natl. Acad. Sci. USA 107, 10815 (2010).
[41] C. Porter, J. Computer-Mediated Comm. 10 (2004).
[42] S. Fortunato, Phys. Rep. 486, 75 (2010).
[43] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E 70, 066111 (2004).
[44] H. Hu and X. Wang, Europhys. Lett. 86, 18003 (2009).