Combinatorial study of degree assortativity in networks

Ernesto Estrada

Department of Mathematics and Statistics and Department of Physics, SUPA and Institute of Complex Systems,
University of Strathclyde, Glasgow G1 1XQ, United Kingdom

Why are some networks degree-degree correlated (assortative), while most of the real-world ones are anticorrelated (disassortative)? Here, we prove, by combinatorial methods, that the assortativity of a network depends only on three structural factors: transitivity (clustering coefficient), intermodular connectivity, and branching. Then, a network is assortative if the contributions of the first two factors are larger than that of the third. Highly branched networks are likely to be disassortative.

Networks represent the topological skeleton of many complex systems in a variety of scenarios, including physical, technological, socioeconomic, and informational domains [1–4]. The understanding of the structure of these complex networks is vital for comprehending the evolutionary, functional, and dynamical processes taking place in these systems [4–6]. A major role in many of these processes, such as epidemic spreading, synchronization, percolation, social organization, protein architecture, network robustness, among others [7–14], is played by the degree assortativity [7]. A network is *assortative* if high-degree nodes tend to attach to other high-degree nodes, while it is *disassortative* if high-degree nodes tend to attach to low-degree ones. A quantitative measure of assortativity was defined by Newman [7] as the Pearson coefficient of the degree-degree correlation in an undirected network. Mathematically, it is usually written as

$$r = \frac{m^{-1} \sum_e k_i(e) k_j(e) - \left\{ m^{-1} \sum_e \frac{1}{2}[k_i(e) + k_j(e)] \right\}^2}{m^{-1} \sum_e \frac{1}{2}\left[k_i^2(e) + k_j^2(e)\right] - \left\{ m^{-1} \sum_e \frac{1}{2}[k_i(e) + k_j(e)] \right\}^2}, \tag{1}$$

where $k_i(e)$ and $k_j(e)$ are the degrees at both ends of link e and m is the number of links. Obviously, $r > 0$ for assortative, and $r < 0$ for disassortative networks.

Degree assortativity is a sort of rare property in networks. For instance, only 8% of the 11 117 connected networks with eight nodes are assortative. Then, for this property to be manifested, a very specific structural characteristic must be present in a network. More interestingly, despite the lack of abundance of assortative networks in general, there are complete classes of real-world networks, which are assortative. For instance, many social networks, mainly collaboration networks, have been found to be assortative [7,9]. Also, close-packed spatial networks, such as protein residue networks, atomic (molecular) systems, and micellar networks are assortative [13,14]. Transitivity (clustering coefficients) [15,16] and modularity [17] are frequently found in empirical correlations with the assortativity coefficient [7,8,18–20]. However, there are very modular networks with very large clustering coefficients, which are not assortative. Some difficulties are known to exist when the assortativity coefficient is used for studying complex networks. As a consequence, better measures have been proposed in the literature, such as the average degree of the nearest neighbors introduced by Pastor-Satorras *et al.* [21]. Then, if this measure is an increasing function of the degree, nodes with high degrees have a greater probability to be connected among them. On the contrary, a decreasing behavior indicates that high-degree nodes are preferentially attached with low-degree ones. Despite the known fact that

the assortativity coefficient is a global measure that can lead to flawed estimations of the correlations strengths, it continues to be used widely in many complex network applications across the biological, physical, and social sciences. Therefore, it is necessary to understand the role played by different structural factors on the assortativity of a network. The aim of this Brief Report is to unfold such structural factors in an analytic way as well as to provide evidence of the role played by them on the assortativity coefficient of real-world networks.

Let us start our analysis by rewriting the assortative coefficient in matrix form as follows:

$$r = \frac{\langle \mathbf{k}|\mathbf{A}|\mathbf{k}\rangle - \frac{1}{2m}(\langle \mathbf{1}|\mathbf{E}|\mathbf{1}\rangle)^2}{\langle \mathbf{1}|\mathbf{E}^2|\mathbf{1}\rangle - \langle \mathbf{k}|\mathbf{A}|\mathbf{k}\rangle - \frac{1}{2m}(\langle \mathbf{1}|\mathbf{E}|\mathbf{1}\rangle)^2}, \tag{2}$$

where $|\mathbf{k}\rangle = \langle \mathbf{1}|\mathbf{A}$ is a vector of node degrees, with $\mathbf{A}$ being the adjacency matrix of the network ($\mathbf{A}_{ij} = 1$ if nodes i and j are linked, zero otherwise) and $|\mathbf{1}\rangle$ is an all-ones vector. The matrix $\mathbf{E} = \nabla^T \nabla$ where ∇ stands for the $n \times m$ incidence matrix of the network (n is the number of nodes), whose entry ∇_{ue} is 1 if the node u is incident with the link e or is zero otherwise. The first term in the numerator of Eq. (2) easily corresponds to $\langle \mathbf{k}|\mathbf{A}|\mathbf{k}\rangle = 2 \sum_e k_i(e) k_j(e)$. The nondiagonal entries of the matrix $\mathbf{E} = \nabla^T \nabla$ represent the adjacency between links in the network. Two links are adjacent if they are incident on the same node, i.e., they have a common node. The diagonal entries of this matrix are equal to 2 as every edge is only incident to two nodes. Then, it is easy to see that the degree

of a link, which is the sum of the nondiagonal entries of a row or column of $\mathbf{E}$, is given by $\delta(e) = k_i(e) + k_j(e) - 2$. Then, $\langle \mathbf{1}|\mathbf{E}\rangle_e = \delta(e) + 2 = [k_i(e) + k_j(e)]$, and we have that $\langle \mathbf{1}|\mathbf{E}|\mathbf{1}\rangle^2 = \{\sum_e [k_i(e) + k_j(e)]\}^2$. Then, the quadratic forms in Eq. (2) can be expressed as

$$\langle \mathbf{k}|\mathbf{A}|\mathbf{k}\rangle = 2|P_1| + 4|P_2| + 2|P_3| + 6|C_3|, \qquad (3)$$

$$\frac{1}{2m}(\langle \mathbf{1}|\mathbf{E}|\mathbf{1}\rangle)^2 = \frac{2|P_2|^2}{|P_1|} + 2|P_1| + 4|P_2|, \qquad (4)$$

$$\langle \mathbf{1}|\mathbf{E}^2|\mathbf{1}\rangle = 2|P_1| + 10|P_2| + 2|P_3| + 6|S_{1,3}| + 6|C_3|, \qquad (5)$$

where the symbol $|\ldots|$ stands for the number of fragments. The fragments (subgraphs) involved are paths and cycles of i nodes P_i and C_i, respectively, and $S_{1,3}$ represents a star graph of four nodes. The number of subgraphs that appear in Eqs. (3)–(5) can be calculated analytically as follows: $|P_1| = \frac{1}{2}\sum_{i=1}^n k_i, |P_2| = \sum_{i=1}^n \binom{k_i}{2}, |C_3| = \frac{1}{6}\text{tr}(\mathbf{A}^3), |P_3| = \sum_{i,j\in E}(k_i - 1)(k_j - 1) - 3|C_3|$, and $|S_{1,3}| = \sum_{i=1}^n \binom{k_i}{3}$. Let us introduce the following ratios: $|P_{r/s}| = |P_r|/|P_s|$ and $C = 3|C_3|/|P_2|$. Notice that C is exactly the clustering coefficient or transitivity introduced by Newman *et al.* [22]. Then, the assortativity coefficient can be written combinatorially as follows:

$$r = \frac{|P_2|\left(|P_{3/2}| + \frac{3|C_3|}{|P_2|} - |P_{2/1}|\right)}{3|S_{1,3}| - |P_2|(|P_{2/1}| - 1)}$$
$$= \frac{|P_2|(|P_{3/2}| + C - |P_{2/1}|)}{3|S_{1,3}| - |P_2|(|P_{2/1}| - 1)}. \qquad (6)$$

It is easy to see that the denominator of Eq. (6) is non-negative such that the sign of the Pearson correlation coefficient depends only on the sign of the numerator of Eq. (6). Consequently, the fact that a network is assortative or disassortative depends only upon the sign of the term $|P_{3/2}| + C - |P_{2/1}|$ in the numerator of Eq. (6). Then, our main result here is that we can write the conditions for degree assortativity as follows:

(1) assortative ($r > 0$): if and only if $|P_{3/2}| + C > |P_{2/1}|$,

(2) neutral ($r = 0$): if and only if $|P_{3/2}| + C = |P_{2/1}|$ and $3|S_{1,3}| - |P_2|(|P_{2/1}| - 1) \neq 0$,

(3) disassortative ($r < 0$): if and only if $|P_{3/2}| + C < |P_{2/1}|$.

Then, there are only three structural factors that determine the assortativity of a network. The transitivity coefficient C measures the proportion of two paths that form triangles. We recall that the transitivity is bounded as $0 \leqslant C \leqslant 1$. Then, the role played by the transitivity in the assortativity of a network depends on the relative values of the other two parameters $|P_{3/2}|$ and $|P_{2/1}|$. In fact, if $|P_{3/2}| > |P_{2/1}|$, the network will be assortative irrespective of its transitivity. Otherwise, if $|P_{3/2}| < |P_{2/1}| + 1$, the network is disassortative no matter how large the clustering coefficient is. The term $|P_{2/1}|$ measures how relatively *branched* a network is by accounting for the ratio of two paths to the number of links. A linear chain (path) is the network with the least relative branching and smallest $|P_{2/1}|$. This ratio increases with an increase in the number of pendant nodes so that, among trees,
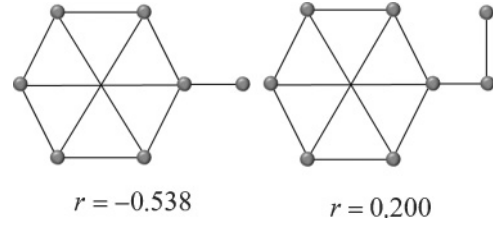

FIG. 1. (Left) Network of type G' obtained by linking a k-regular graph and a single node. This network is disassortative as indicated by its Pearson coefficient. (Right) Network of type G'' in which a k-regular graph is linked to another regular graph, e.g., a path of length 2. The network is assortative as indicated by its Pearson correlation coefficient.

the star has the largest value of $|P_{2/1}|$. Among all networks with n nodes, the complete graph has the largest value of $|P_{2/1}|$. The third term determining the assortativity of a network $|P_{3/2}|$ is a little bit trickier to interpret. In a regular network G of degree k, it is easy to see that it only depends on the degree and transitivity of the network $P_{3/2}(G) = (k-1) - C$. Now, let us consider network G', which is created by linking together network G and a network formed by only one node (see Fig. 1, left). It is easy to see that $P_{3/2}(G') = P_{3/2}(G) = (k-1) - C$. We remark that network G' is more branched than G as it has $P_{2/1}(G') \approx k - (n-3)/(n-1)$ for $n(G') \gg 2$ instead of $P_{2/1}(G) = k - 1$. However, $|P_{3/2}|$ has not been affected by the increase in branching. Now, let us create network G'' by linking together network G and a regular network consisting of two nodes linked by a link (P_2) (see Fig. 1, right). In this case, the ratio $|P_{3/2}|$ changes to $P_{3/2}(G) \approx (k-1) + 2k/[(n-2)(k-1)] - C$ for $n \gg k$. The generalization of this result to the linking of any two regular graphs will be considered elsewhere. The important message, at this point, is that, by linking together G and P_2, we have created network G'' consisting of two modules, and the value of $|P_{3/2}|$ reflects the intermodular connectivity in this network [23]. This assumes that the modules are well connected internally as is the general case for the definition of modular networks. In closing, the assortativity of a network is determined by three structural factors: transitivity, intermodular connectivity, and relative branching. We will analyze these factors in detail now.

Let us start our analysis by considering network G', and let us assume, for the sake of simplicity, that $C \equiv 0$. It is straightforward to realize that $(k-1) < k - (n-3)/(n-1)$, which means that $P_{3/2}(G') + C(G') < P_{2/1}(G')$ and network G' is disassortative (see Fig. 1). That is, the addition of one pendant node to a regular network has increased the branching but has not increased the intermodular connectivity at all, resulting in disassortativity. In the case of network G'', the relative branching is given by $P_{2/1}(G) = (k-1) - [2(k-1) - 4]/(nk - 2k + 4)$, and because $(k-1) + 2k/[(n-2)(k-1)] > (k-1) - [2(k-1) - 4]/(nk - 2k + 4)$, we have that $P_{3/2}(G'') + C(G'') > P_{2/1}(G'')$ for all $k \geqslant 3$, which means that any network of type G'' is assortative for $k \geqslant 3$ (see Fig. 1). In other words, network G'' displays larger intermodular connectivity than branching, which is translated into assortativity. To summarize, for a network being assortative, it is enough that the intermodular connectivity be larger than its relative branching. If both terms are almost

TABLE I. Relative branching ($|P_{2/1}|$), transitivity (C), intermodular connectivity ($|P_{3/2}|$) , and assortativity coefficient for real-world networks.

Network	$|P_{2/1}|$	$|P_{3/2}|$	C	r
Prison	4.25	4.09	0.288	0.103
Protein residue	4.41	4.45	0.417	0.412
St. Marks	10.54	10.46	0.291	0.118
Geom	17.42	22.09	0.224	0.168
Corporate	19.42	20.60	0.498	0.268
Roget	9.55	10.08	0.134	0.174
Jazz	127.30	144.84	0.771	0.412
Zachary	6.77	4.49	0.256	−0.476
Drugs	14.58	12.84	0.368	−0.118
Transcription	12.51	3.01	0.016	−0.410
Bridge Brook	22.42	17.31	0.191	−0.664
USAir97	43.36	36.97	0.396	−0.208
Internet	91.00	11.53	0.015	−0.229

identical, then the transitivity will make the difference between assortativity and disassortativity. We remark, without proof, that, for increasing intermodular connectivity over branching, the two modules linked together must be of different sizes.

There are some empirical rules that can guide understanding assortativity in complex networks. For instance, the existence of cliques or quasicliques increases both branching and clustering. Then, a first rule is that, if these cliques (quasicliques) are separated at relatively large distances, the increase in branching is much larger than that of the intermodular connectivity, and the network is likely to be disassortative. However, if many cliques (quasicliques) are linked together, the network is likely to be assortative. On the other hand, starlike structures tend to increase branching more than intermodular connectivity. Therefore, a second empirical rule is that the networks with many such structures are likely to be disassortative. A third rule is that, in an almost regular network $|P_{3/2}| \approx |P_{2/1}|$, the network can be assortative if its transitivityis large enough. In order to analyze this empirical evidence in light of the results obtained in this Brief Report, we propose studying several real-world networks.

Here, we study 13 undirected real-world networks, including social networks representing the corporate elite in the United States (Corporate), a scientific collaboration network in computational geometry (Geom), inmates in prison (Prison), injectable drug users (Drugs), the Zachary karate club (Zachary), and the collaboration between jazz musicians (Jazz); four biological networks representing the transcription interaction networks in yeast (Transcription), a protein residue

network (Protein residue), and the food webs of Bridge Brook and St. Marks; a network based on Roget's Thesaurus of English (Roget), two technological networks representing the airport transportation network in the United States in 1997 (USAir97), and the Internet at the autonomous systems (AS) level as from April 1997. Details and references can be found in Ref. [24]. In Table I, we illustrate the results for these networks. The first three networks display *clustering-driven assortativity* as they have $|P_{3/2}| \approx |P_{2/1}|$ but display large enough clustering as to perform $|P_{3/2}| + C > |P_{2/1}|$. However, the structural causes for their assortativity are different. The first two networks (Prison and Protein residue) are characterized by a few cliques (quasicliques), which are not connected directly to each other, but somehow spread across the networks (first empirical rule). The third network (St. Marks) is an *almost-regular* network with large clustering (third empirical rule). The next four networks in Table I display *modularity-driven assortativity* as they have $|P_{3/2}| > |P_{2/1}|$. That is, these networks display larger intermodularity than their relative branching. In general, they can be seen as networks formed by many relatively small cliques (quasicliques), which are relatively close to each other (first empirical rule). For instance, in the collaboration network of jazz musicians, there are two cliques of 171 and 110 musicians, respectively, to which almost all others of the more than 1600 small cliques are linked to by a few ties. These cliques increase branching very much (see $|P_{2/1}|$ in Table I), but the way in which they are connected makes the intermodule connectivity very large and $|P_{3/2}| > |P_{2/1}|$. Finally, all disassortative networks shown in Table I are branched very much as to make $|P_{2/1}| > |P_{3/2}|$ (second empirical rule). In particular, the network of Internet at AS and the transcription network of yeast display a large unbalance between branching and intermodularity. In the Internet network, there are many starlike subgraphs having up to 169 pendant nodes connected to the same hub. This, of course, dramatically increases $|P_{2/1}|$ with very little impact on $|P_{3/2}|$.

To conclude, in this Brief Report, we have found a combinatorial expression for the assortativity coefficient in terms of network structural parameters. The assortativity of a network depends on the balance between three structural factors: transitivity (clustering), intermodular connectivity, and relative branching. The first two perform a positive contribution to the assortativity of a network, while branching is more likely associated with disassortative networks. We have found analytical and empirical evidence about the different topological organizations of networks that contribute to each of these factors. These results can be useful in understanding the relationship between assortativity and other network structural (dynamical) parameters.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[4] S. H. Strogatz, Nature (London) **410**, 268 (2001).

[5] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, Cambridge, UK, 2008).
[6] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, Adv. Phys. **56**, 167 (2007).

[7] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).

[8] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003); A. Capocci, G. Caldarelli, and P. De Los Rios, *ibid.* **68**, 047101 (2003).

[9] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).

[10] V. M. Eguíluz and K. Klemm, Phys. Rev. Lett. **89**, 108701 (2002).

[11] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. Lett. **90**, 028701 (2003).

[12] M. Di Bernardo, F. Garofalo, and F. Sorrentino, Int. J. Bifurcation Chaos **17**, 3499 (2006).

[13] G. Bagler and S. Sinha, Bioinformatics **23**, 1760 (2007).

[14] D. Turgut, A. R. Atilgan, and C. Atilgan, PLoS ONE **5**, e15551 (2010).

[15] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[16] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).

[17] X.-K. Xu, J. Zhang, and M. Small, Phys. Rev. E **82**, 046117 (2010).

[18] M. A. Serrano and M. Boguñá, Phys. Rev. Lett. **97**, 088701 (2006).

[19] P. Holme and J. Zhao, Phys. Rev. E **75**, 046111 (2007).

[20] D. V. Foster, J. G. Foster, P. Grassberger, and M. Paczuski, e-print arXiv:1012.2384.

[21] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).

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

[23] Note that, if the two modules are cliques, P_3 is the longest distance in the network.

[24] E. Estrada, Eur. Phys. J. B **52**, 563 (2006).