Internet Mathematics Vol. 2, No. 4

Supplemental Material: The $s(g)$-Metric and Assortativity

Lun Li, David Alderson, John C. Doyle, and Walter Willinger

Following the development of Newman [Newman 02], let $P(\{D_i = k\}) = P(k)$ be the node degree distribution over the ensemble of graphs, and define $Q(k) = (k+1)P(k+1)/\sum_{j\in D} jP(j)$ to be the normalized distribution of *remaining degree* (i.e., the number of "additional" connections for each node at either end of the chosen link). Let $\bar{D} = \{d_1-1, d_2-1, \cdots, d_n-1\}$ denote the remaining degree sequence for g. This remaining degree distribution is $Q(k) = \sum_{k'\in\bar{D}} Q(k,k')$, where $Q(k,k')$ is the *joint probability distribution* among remaining nodes, i.e., $Q(k,k') = P(\{D_i = k+1, D_j = k'+1 | (i,j) \in \mathcal{E}\})$. In a network where the remaining degree of any two vertices is independent, i.e., $Q(k,k') = Q(k)Q(k')$, there is no degree-degree correlation, and this defines a network that is neither assortative nor disassortative (i.e., the "center" of this view into the ensemble). In contrast, a network with $Q(k,k') = Q(k)\delta[k-k']$ defines a perfectly assortative network. Thus, graph assortivity r is quantified by the *average* of $Q(k,k')$ over all the links

$$r = \frac{\sum_{k,k'\in\bar{D}} kk'(Q(k,k') - Q(k)Q(k'))}{\sum_{k,k'\in\bar{D}} kk'(Q(k)\delta[k-k'] - Q(k)Q(k'))}, \tag{1}$$

with proper centering and normalization according to the value of perfectly assortative network, which ensures that $-1 \le r \le 1$. Many stochastic graph generation processes can be understood directly in terms of the correlation distributions among these so-called remaining nodes, and this functional form facilitates the direct calculation of their assortativity. In particular, Newman [Newman 02] shows that both Erdös-Renyí random graphs and Barabási-Albert preferential attachment growth processes yield ensembles with zero assortativity.

Newman [Newman 05] also develops the following sample-based definition of

1542-7951/05 $0.50 per page

assortativity

$$r(g) = \frac{\left[\sum_{(i,j)\in\mathcal{E}} d_i d_j/l\right] - \left[\sum_{(i,j)\in\mathcal{E}} \frac{1}{2}(d_i + d_j)/l\right]^2}{\left[\sum_{(i,j)\in\mathcal{E}} \frac{1}{2}(d_i^2 + d_j^2)/l\right] - \left[\sum_{(i,j)\in\mathcal{E}} \frac{1}{2}(d_i + d_j)/l\right]^2},$$

which is equivalent to

$$r(g) = \frac{\left[\sum_{(i,j)\in\mathcal{E}} d_i d_j\right] - \left[\sum_{i\in\mathcal{V}} \frac{1}{2} d_i^2\right]^2 /l}{\left[\sum_{i\in\mathcal{V}} \frac{1}{2} d_i^3\right] - \left[\sum_{i\in\mathcal{V}} \frac{1}{2} d_i^2\right]^2 /l} \tag{2}$$

(see Equation (5.5) in [Li et al. 05]).

While the ensemble-based notion of assortativity in (1) has important differences from the sample-based notion of assortativity in (2), their relationship can be understood by viewing a given graph as a singleton on an ensemble of graphs (i.e., where the graph of interest is chosen with probability 1 from the ensemble). For this graph, if we define the number of nodes with degree k as $N(k)$, we can derive the degree distribution $P(k)$ and the remaining degree distribution $Q(k)$ on the ensemble as

$$P(k) = \frac{N(k)}{n}$$

and

$$Q(k) = \frac{(k+1)P(k+1)}{\sum_{j\in D} jP(j)} = \frac{(k+1)N(k+1)}{\sum_{j\in D} jN(j)}.$$

Also, it is easy to see that

$$\begin{aligned} \sum_{i\in\mathcal{V}} d_i &= \sum_{k\in D} kN(k) = 2l, \\ \sum_{i\in\mathcal{V}} d_i^2 &= \sum_{k\in D} k^2 N(k), \\ &\vdots \\ \sum_{i\in\mathcal{V}} d_i^m &= \sum_{k\in D} k^m N(k), \end{aligned}$$

where m is a positive integer.

Equations (1) and (2) can be related term-by-term in the following manner. The first term of the numerator, $Q(k,k')$, represents the joint probability distribution of the (remaining) degrees of the two nodes at either end of a randomly chosen link. For a given graph, let $l(k,k')$ represent the number of links connecting nodes with degree k to nodes with degree k'. Then, we can write $Q(k,k') = l(k,k')/l$, and hence

$$\sum_{k,k'\in\bar{D}} kk'Q(k,k') = \frac{1}{l} \sum_{(i,j)\in\mathcal{E}} d_i d_j.$$

The first term of the denominator of r in Equation (1) can be written as

$$\sum_{k,k' \in \bar{D}} kk' Q(k)\delta[k-k'] = \sum_{k \in \bar{D}} k^2 Q(k) \tag{3}$$

$$= \frac{\sum_{k \in D} (k+1)^3 N(k+1)}{\sum_{i \in D} jN(j)}$$

$$= \frac{\sum_{i \in \mathcal{V}} d_i^3}{2l}, \tag{4}$$

and the centering term (in both the numerator and the denominator) is

$$\sum_{k,k' \in \bar{D}} kk' Q(k)Q(k') = \left(\sum_{k \in \bar{D}} kQ(k) \right)^2 \tag{5}$$

$$= \left(\frac{\sum_{k \in D} (k+1)^2 N(k+1)}{\sum_{i \in D} jN(j)} \right)^2$$

$$= \left(\frac{\sum_{i \in \mathcal{V}} d_i^2}{2l} \right)^2. \tag{6}$$

In both of these cases, the offset of a constant in representing the degree sequence as D versus $\bar{D}$ does not effect the overall calculation. The relationships between the ensemble-based quantities (LHS of (3) and LHS of (5)) and their sample-based (i.e., structural) counterparts (4) and (6) holds (approximately) when the expected degree equals the actual degree.

To see why (6) can be viewed as the center, we consider the following thought experiment: *what is the structure of a deterministic graph with degree sequence D and having zero assortativity?* In principle, a node in such a graph will connect to any other node in proportion to each node's degree. While such a graph may not exist for general D, one can construct a deterministic *pseudograph* $\tilde{g}$ having zero assortativity in the following manner. Let $A = [a_{ij}]$ represent a (directed) node adjacency matrix of nonnegative real values, representing the *link weights* in the pseudograph. That is, links are not constrained to integer values but can exist in fractional form. The zero assortative pseudograph will have symmetric weights given by

$$a_{ij} = \left(\frac{d_j}{\sum_{k \in \mathcal{V}} d_k} \right) \left(\frac{d_i}{2} \right) = \left(\frac{d_i}{\sum_{k \in \mathcal{V}} d_k} \right) \left(\frac{d_j}{2} \right) = a_{ji}.$$

Thus, the weight a_{ij} for each link emanating out of node i is in proportion to the degree of node j, in a manner that is relative to the sum of all node degrees. In general, the graphs of interest to us are undirected, however here it is notationally convenient to consider the construction of directed graphs. Using these weights,

the total weight among all links entering and exiting a particular node i equals

$$\sum_{j\in\mathcal{V}} a_{ij} + \sum_{k\in\mathcal{V}} a_{ki} = d_i/2 + d_i/2 = d_i.$$

Accordingly, the total link weights in the pseudograph are equal to

$$\sum_{i,j\in\mathcal{V}} a_{ij} = \sum_{j\in\mathcal{V}} d_j/2 = l,$$

where l corresponds to the total number of links in a traditional graph. The s-metric for the pseudograph $\tilde{g}_A$ represented by matrix A can be calculated as

$$\begin{aligned} s(\tilde{g}_A) &= \sum_{j\in\mathcal{V}}\sum_{i\in\mathcal{V}} d_i a_{ij} d_j \\ &= \sum_{j\in\mathcal{V}}\left[\sum_{i\in\mathcal{V}} d_i \left(\frac{d_j}{\sum_{k\in\mathcal{V}} d_k}\right)\left(\frac{d_i}{2}\right)\right] d_j \\ &= \frac{\left(\sum_{j\in\mathcal{V}} d_j^2\right)\left(\sum_{i\in\mathcal{V}} d_i^2\right)}{2\left(\sum_{k\in\mathcal{V}} d_k\right)} \\ &= \frac{\left(\sum_{j\in\mathcal{V}} d_j^2\right)^2}{4l}, \end{aligned}$$

and we have

$$\frac{s(\tilde{g}_A)}{l} = \left(\frac{\sum_{i\in\mathcal{V}} d_i^2}{2l}\right)^2,$$

which is equal to (6).

In principle, one could imagine a deterministic procedure that uses the structural pseudograph $\tilde{g}_A$ to generate the zero-assortativity graph among an unconstrained background set G. That is, graphs resulting from this procedure could have multiple links between any pair of nodes as well as multiple self-loops and would not necessarily be connected. The challenge in developing such a procedure is to ensure that the resulting graph has degree sequence equal to D, although one can imagine that in the limit of large graphs, this becomes less of an issue. By extension, it is not hard to conceive a stochastic process that uses the structural pseudograph $\tilde{g}_A$ to generate a statistical ensemble of graphs having expected assortativity equal to zero. In fact, it is not hard to see why the GRG method [Chung and Lu 03] is very close to such a procedure.

Note that the total weight in the pseudograph between nodes i and j equals $a_{ij} + a_{ji} = d_i d_j/2l$. Recall that the GRG method discussed in [Li et al. 05,

Section 5.1] is based on the choice of a probability $p_{ij} = \rho d_i d_j$ of connecting two nodes i and j and also that in order to ensure that $E(d_i) = d_i$ one needs $\rho = 1/2l$, provided that $\max_{i \neq j \in \mathcal{V}} d_i d_j \leq 2l$. Thus, the GRG method can be viewed as a stochastic procedure that generates real graphs from the pseudograph $\tilde{g}_A$, with the one important difference that the GRG method always results in simple (but not necessarily connected) graphs. Thus, the zero-assortativity pseudograph $\tilde{g}_A$ can be interpreted as the *deterministic outcome* of a GRG-like construction method. Accordingly, one expects that the statistical ensemble of graphs resulting from the stochastic GRG method could have zero assortativity, but this has not been proven.

In summary, graph assortativity captures a fundamental feature of graph structure, one that is closely related to our s-metric. However, the existing notion of assortativity for an individual graph g is implicitly measured against a background set of graphs G that is *not* constrained to be either simple or connected. The connection between the sample-based and ensemble-based definitions makes it possible to calculate the assortativity among graphs of different sizes and having different degree sequences, as well as for different graph evolution procedures. Unfortunately, because this metric is computed relative to an unconstrained background set, in some cases this normalization (against the $s_{\max}$ graph) and centering (against the $\tilde{g}_A$ pseudograph) does a relatively poor job of distinguishing among graphs having the *same* degree sequence (see, for example, [Li et al. 05, Figure 5]).

References

[Chung and Lu 03] F. Chung and L. Lu. "The Average Distance in a Random Graph with Given Expected Degrees." *Internet Math.* 1:1 (2003), 91–113.

[Li et al. 05] Lun Li, D. Alderson, J. C. Doyle, and W. Willinger. "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implication." *Internet Mathematics* 2:4 (2005), 429–521.

[Newman 02] M. E. J. Newman "Assortative Mixing in Networks." *Phys. Rev. Lett.* 89 (2002), 208701.

[Newman 05] M. E. J. Newman. "Power Laws, Pareto Distributions and Zipf's Law." *Contemporary Physics* 46:5 (2005), 323–351.

Lun Li, Engineering & Applied Science Division, California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125 (lun@cds.caltech.edu)

David Alderson, Engineering & Applied Science Division, California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125 (alderd@cds.caltech.edu)

John C. Doyle, Engineering & Applied Science Division, California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125 (doyle@cds.caltech.edu)

Walter Willinger, AT&T Labs—Research, 180 Park Avenue, Florham Park, NJ 07932 (walter@research.att.com)