Social Networks xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

A new look at clustering coefficients with generalization to weighted and multi-faction networks

Kenneth S. Berenhaut*, Rebecca C. Kotsonis, Hongyi Jiang

Wake Forest University, 1834 Wake Forest Rd, Winston Salem, NC, USA

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Clustering coefficient
Multiple random walks
Two-mode networks
Geodesic distance
Weighted networks
Multi-faction networks

ABSTRACT

In this paper we propose a new method for studying local and global clustering in networks employing random walk pairs. The method is intuitive and directly generalizes standard local and global clustering coefficients to weighted networks and networks containing nodes of multiple types. In the case of two-mode networks the values obtained for commonly considered social networks are in sharp contrast to those obtained, for instance, by the method of Opsahl (2013), and provide a different viewpoint for clustering. The approach is also applicable in questions related to the general study of segregation and homophily. Applications to existent data sets are considered.

1. Introduction

In this paper we propose a new method for studying local and global clustering in networks employing random walk pairs. The method is intuitive and directly generalizes standard local and global clustering coefficients to two-mode networks (as well as weighted networks and networks containing nodes of multiple types).

One often considered local property of social networks is that of triadic closure (within the ego network of an individual). In particular, suppose an individual v has k_v neighbors, and hence $k_v(k_v-1)/2$ distinct pairs of neighbors. A standard measure is the local clustering coefficient, C_v, which is the proportion of these pairs who are themselves connected (see Watts and Strogatz, 1998), i.e.

$$C_v = \frac{number\ \ of\ \ pairs\ \ of\ \ neighbors\ \ of\ \ v\ \ that\ \ are\ \ connected}{number\ \ of\ \ pairs\ \ of\ \ neighbors\ \ of\ \ v}. \tag{1}$$

Consider a network represented as an undirected graph, $G=(V, E)$, with a set of n vertices or nodes, $V=\{v_1, v_2, \ldots, v_n\}$, and a set of connections or edges, E. Averaging C_v over all nodes v then leads to a measure, C_G, of global clustering (see Watts and Strogatz, 1998)

$$C_G = \frac{1}{n}\sum_{v\in V} C_v. \tag{2}$$

Akin to (1), an alternative measure of global clustering (Newman et al., 2001) is given by

$$C_G^* = \frac{number\ \ of\ \ paths\ \ of\ \ length\ \ 2\ \ in\ \ G\ \ that\ \ are\ \ closed}{number\ \ of\ \ paths\ \ of\ \ length\ \ 2\ \ in\ \ G}, \tag{3}$$

where a path of length two is a triple $(u, v, w) \in V^3$ satisfying $\{(u, v), (v, w)\} \subseteq E$, and such a path is closed when, in addition, $(u, w) \in E$. Barrat et al. (2004) and Opsahl and Panzarasa (2009) extended C_v and C_G^* to weighted graphs by incorporating weights of triangles in (1) and (3), respectively (see also Saramäki et al., 2007 and the references therein).

The study of closure in the neighborhood of an individual is motivated by considerations of tension and cohesiveness from the perspective of the individual. Triplets of nodes (triads; see Simmel and Wolff, 1950) – and sentiments, connections, and interactions between members – have been a topic of interest for several decades. For discussion of social capital derived for members based on existent strong or weak connections (or lack thereof) see for instance Granovetter (1973), Burt (1992) and Coleman (1988). Consonance in triads, and implications for the network as a whole, has been considered through aspects of cognitive and structural balance (see Heider, 1946; Cartwright and Harary, 1956; Holland and Leinhardt, 1971). For discussion of the influence of social contexts on triadic closure, see for instance Feld (1981) and Kossinets and Watts (2009), and the references therein.

Rather than simply considering connections between neighbors of a node $v \in V$, one might, more generally, be interested in the

* Corresponding author at: Department of Mathematics and Statistics, Wake Forest University, 1834 Wake Forest Rd, Winston-Salem, NC 27109, USA.
E-mail address: berenhks@wfu.edu (K.S. Berenhaut).

http://dx.doi.org/10.1016/j.socnet.2017.08.003
0378-8733/

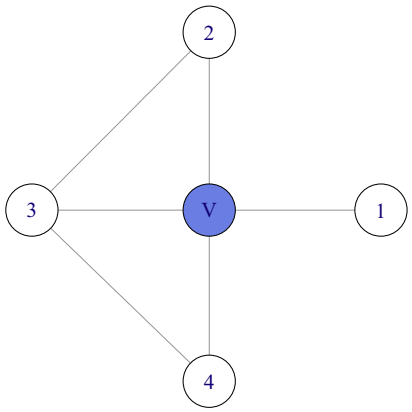

Fig. 1. A simple five-node network.

proximity of neighbors (of an individual) to each other in the graph. Instead of a binary perspective of connection, one might naturally consider how far apart two randomly chosen neighbors of a particular node are.

Specifically, consider a particular node $v \in V$ and uniformly and independently select two neighbors of v, say W_1 and W_2 (note that W_1 and W_2 may be equal). A quantity of interest is then the expected value of $d_G(W_1, W_2)$, where for two nodes x and y, $d_G(x, y)$ denotes the shortest path distance between x and y. Note that, here $d_G(W_1, W_2)$ is either 0 (if $W_1 = W_2$), 1 (if W_1 and W_2 are neighbors), or 2 (since v is a common neighbor).

Thus, for $v \in V$ define γ_v via

$$\gamma_v = \gamma_v(G) \stackrel{def}{=} \mathbb{E}(d_G(W_1, W_2)), \tag{4}$$

where $\mathbb{E}$ represents expected value. The value of γ_v is bounded between zero and two and is readily interpreted as giving meaningful information regarding clustering near the node of interest. For fixed degree, the exact value is a function of the number of connections among neighbors of v (as is C_v in (1)). In particular, setting $Y = d_G(W_1, W_2)$, $k = k_v$ and letting $e = e_v$ be the number of pairs of connected neighbors of v, i.e.

$$e = e_v = |\{(u_1, u_2) \in V \times V : \{(u_1, u_2), (v, u_1), (v, u_2)\} \subseteq E\}|, \tag{5}$$

we have the probabilities $\mathbb{P}(Y = 0) = 1/k$, $\mathbb{P}(Y = 1) = e/k^2$ and $\mathbb{P}(Y = 2) = 1 - (e + k)/k^2$. Hence

$$\begin{aligned} \gamma_v = \mathbb{E}(Y) &= \frac{e}{k^2} + \frac{2(k^2 - (e + k))}{k^2} \\ &= \frac{2k(k-1) - e}{k^2}. \end{aligned} \tag{6}$$

Consider the following simple example.

Example 1. Fig. 1 gives a simple network with five nodes. For the central node v, we have $\mathbb{P}(Y = 0) = 1/4$, $\mathbb{P}(Y = 1) = 4/16$, $\mathbb{P}(Y = 2) = 1/2$, and the expected distance between two uniformly and independently selected neighbors W_1 and W_2 of v is given by $\gamma_v = 1.25$. Corresponding values for the other nodes are $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0, 1/2, 8/9, 1/2)$.

Referring to Eq. (6), γ_v is monotone in e for fixed k. Hence, inserting the extreme values of zero and $k(k-1)$ for e in (6), leads to the following simple result.

Lemma 1. *Suppose node v has degree k_v in the graph G, and for any positive integer x, define $m(x) = (x-1)/x$ and $M(x) = 2(x-1)/x$. Then*

$$m(k_v) \le \gamma_v \le M(k_v), \tag{7}$$

and the lower and upper bounds in (7) are best possible.

Considering Lemma 1, one potential normalization of the quantity γ_v is the "min–max scaling" of γ_v (see for instance Jain et al., 2005; Aksoy and Haralick, 2001), i.e.

$$\gamma_v^* \stackrel{def}{=} \frac{\gamma_v - m(k_v)}{M(k_v) - m(k_v)}. \tag{8}$$

It follows easily that $0 \le \gamma_v^* \le 1$ for $v \in V$. There is a, perhaps surprising, simple equivalence between the normalized distance in (8) and the standard clustering coefficient, as in (1).

Theorem 1. *Suppose $v \in V$. Then*

$$C_v = 1 - \gamma_v^*. \tag{9}$$

Proof. Writing $k = k_v$, and inserting the values for $m(k)$ and $M(k)$ in (7) gives

$$\begin{aligned} \gamma_v^* &= \frac{2k(k-1) - e - k(k-1)}{k(k-1)} \\ &= 1 - \frac{e}{k(k-1)} = 1 - C_v. \end{aligned} \tag{10}$$

□

Akin to (2), we can define the global value γ_G via

$$\gamma_G \stackrel{def}{=} \frac{1}{n} \sum_{v \in V} \gamma_v. \tag{11}$$

The value of γ_G can then be interpreted as the expected distance between two randomly chosen neighbors of a randomly chosen node from V.

For discussion of some further concepts related to the clustering coefficients C_v, such as redundancy, efficiency, and effective size, see for instance Latora et al. (2013) and the references therein.

The intuitive sense of γ_v in (4) and Theorem 1 suggest generalization to other scenarios. The main benefits of the approach taken here center on coverage of clustering in a wide variety of contexts (binary one-mode, weighted one-mode, two-mode, and more generally any undirected network wherein a particular subset of nodes is of interest), and the natural and inherent emphasis on "stronger" ties and network exploration, resulting from the employment of random walks (see, in particular discussion surrounding Fig. 4 in Section 2 and Figs. 8 and 9 in Section 3). To our knowledge this is the first approach directly applicable in all the above scenarios.

The remainder of the paper proceeds as follows. In Sections 2 and 3, we consider local clustering in weighted networks and networks with varying node attributes (including two-mode networks), respectively. Section 4 contains some discussion regarding computational aspects and Section 5 concludes with applications to existent data sets. An appendix is included, which deals with some technicalities from Section 3 regarding two-mode networks.

2. Weighted networks

Generalization of clustering coefficients to graphs endowed with a weight function on edges has been considered by several authors (see for instance Saramäki et al., 2007; Phan et al., 2013). The process leading to the definition of γ_v in (4) above extends naturally to such graphs, with similar connections to existing methods (see Theorem 2).

Assume that $G = (V, \omega)$ is an undirected weighted graph, with a set of vertices V, and a symmetric weight function ω from $V \times V$ to the non-negative reals, $\mathbb{R}^+$. Such graphs arise in many ecological, social, physical and economic studies where the weights can represent varying tie strength, intensity or capacity (see for instance

Please cite this article in press as: Berenhaut, K.S., et al., A new look at clustering coefficients with generalization to weighted and multi-faction networks. Soc. Netw. (2017), http://dx.doi.org/10.1016/j.socnet.2017.08.003