Social Networks 31 (2009) 155–163

Contents lists available at ScienceDirect

Social Networks

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

Clustering in weighted networks

Tore Opsahl*, Pietro Panzarasa

Queen Mary University of London, School of Business and Management, Mile End Road, E1 4NS London, UK

ARTICLE INFO

Keywords:
Clustering
Transitivity
Weighted networks

ABSTRACT

In recent years, researchers have investigated a growing number of weighted networks where ties are differentiated according to their strength or capacity. Yet, most network measures do not take weights into consideration, and thus do not fully capture the richness of the information contained in the data. In this paper, we focus on a measure originally defined for unweighted networks: the global clustering coefficient. We propose a generalization of this coefficient that retains the information encoded in the weights of ties. We then undertake a comparative assessment by applying the standard and generalized coefficients to a number of network datasets.

1. Introduction

While a substantial body of recent research has investigated the topological features of a variety of networks (Barabási et al., 2002; Ingram and Roberts, 2000; Kossinets and Watts, 2006; Uzzi and Spiro, 2005; Watts and Strogatz, 1998), relatively little work has been conducted that moves beyond merely topological measures to take explicitly into account the heterogeneity of ties (or edges) connecting nodes (or vertices) (Barrat et al., 2004). In a number of real-world networks, ties are often associated with weights that differentiate them in terms of their strength, intensity or capacity (Barrat et al., 2004; Wasserman and Faust, 1994). On the one hand, Granovetter (1973) argued that the strength of social relationships in social networks is a function of their duration, emotional intensity, intimacy, and exchange of services. On the other, for non-social networks, weights often refer to the function performed by ties, e.g., the carbon flow ($mg/m^2/day$) between species in food webs (Luczkowich et al., 2003; Nordlund, 2007), the number of synapses and gap junctions in neural networks (Watts and Strogatz, 1998), or the amount of traffic flowing along connections in transportation networks (Barrat et al., 2004). In order to fully capture the richness of the data, it is therefore crucial that the measures used to study a network incorporate the weights of the ties.

A measure that has long received much attention in both theoretical and empirical research is concerned with the degree to which nodes tend to cluster together. Evidence suggests that in most real-world networks, and especially social networks, nodes tend to cluster into densely connected groups (Feld, 1981; Friedkin, 1984; Holland and Leinhardt, 1970; Louch, 2000; Simmel, 1923; Snijders, 2001; Snijders et al., 2006; Watts and Strogatz, 1998). In particular, the problem of network clustering can be investigated from a two-fold perspective. On the one hand, it involves determining whether and to what extent clustering is a property of a network or, alternatively, whether nodes tend to be members of tightly knit groups (Luce and Perry, 1949). On the other, it is concerned with the identification of the groups of nodes into which a network can be partitioned. This can be obtained, for example, by applying algorithms for community detection that assess and compare densities within and between groups (Newman, 2006; Newman and Girvan, 2004; Rosvall and Bergstrom, 2008), or by using the image matrix in blockmodeling for grouping nodes with the same or similar patterns of ties and uncovering connections between groups of nodes (Doreian et al., 2005).

In this paper, we focus our attention only on the problem of determining whether clustering is a property of a network. More specifically, to address this problem one may ask: If there are three nodes in a network, *i*, *j*, and *k*, and *i* is connected to *j* and *k*, how likely is it that *j* and *k* are also connected with each other? In real-world networks, empirical studies have shown that this likelihood tends to be greater than the probability of a tie randomly established between two nodes (Barabási et al., 2002; Davis et al., 2003; Ebel et al., 2002; Holme et al., 2004; Ingram and Roberts, 2000; Newman, 2001; Uzzi and Spiro, 2005; Watts and Strogatz, 1998). For social networks, scholars have investigated the mechanisms that are responsible for the increase in the probability that two people will be connected if they share a common acquaintance (Holland and Leinhardt, 1971; Simmel, 1923; Snijders, 2001; Snijders et al., 2006). The nature of these mechanisms can be cognitive, as in the case of individuals' desire to maintain balance among ties with others (Hallinan, 1974; Heider, 1946), social, as in the case of third-part referral (Davis, 1970), or can be explained in other ways, such as in terms of focus constraints (Feld, 1981; Kossinets and Watts, 2006; Louch, 2000; Monge et al., 1985) or the differing popularity among

* Corresponding author. Tel.: +44 20 7882 6984; fax: +44 20 7882 3615.
E-mail addresses: t.opsahl@qmul.ac.uk (T. Opsahl), p.panzarasa@qmul.ac.uk (P. Panzarasa).

0378-8733/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2009.02.002

individuals (Feld and Elmore, 1982a, b). While clustering is likely to result from a combination of all these mechanisms, network studies have offered no conclusive theoretical explanation of its causes, nor have they concentrated as much on its underpinning processes as on the measures to formally detect its presence in real-world networks (Levine and Kurzban, 2006).

Traditionally, the two main measures developed for testing the tendency of nodes to cluster together into tightly knit groups are the local clustering coefficient (Watts and Strogatz, 1998) and the global clustering coefficient (Feld, 1981; Karlberg, 1997, 1999; Louch, 2000; Newman, 2003). The local clustering coefficient is based on ego's network density or local density (Scott, 2000; Wasserman and Faust, 1994). For node *i*, this is measured as the fraction of the number of ties connecting *i*'s neighbors over the total number of possible ties between *i*'s neighbors. To create an overall local coefficient for the whole network, the individual fractions are averaged across all nodes.

Despite its ability to capture the degree of social embeddedness that characterizes the nodes of a network, nonetheless the local clustering coefficient suffers from a number of limitations. First, in its original formulation, it does not take into consideration the weights of the ties in the network. As a result, the same value of the coefficient might be attributed to networks that share the same topology but differ in terms of how weights are distributed across ties and, as a result, may be characterized by different likelihoods to befriend the friends of one's friends. Second, the local clustering coefficient does not take into consideration the directionality of the ties connecting a node to its neighbors (Wasserman and Faust, 1994).[1] Recently, there have been a number of attempts to extend the local clustering coefficient to the case of weighted networks (Barrat et al., 2004; Lopez-Fernandez et al., 2004; Onnela et al., 2005; Zhang and Horvath, 2005). However, the issue of directionality still remains mainly unresolved (Caldarelli, 2007), thus making the coefficient suitable primarily for undirected networks.

Moreover, the local clustering coefficient, even in its weighted version, is biased by correlations with nodes' degrees: a node with more neighbors is likely to be embedded in relatively fewer closed triplets, and therefore to have a smaller local clustering than a node connected to fewer neighbors (Ravasz and Barabási, 2003; Ravasz et al., 2002). An additional bias might stem from degree–degree correlations. When nodes preferentially connect to others with similar degree, local clustering is positively correlated with nodes' degree (Ravasz and Barabási, 2003; Ravasz et al., 2002; Soffer and Vàzquez, 2005). Lack of comparability between values of clustering of nodes with different degrees thus makes the average value of local clustering sensitive with respect to how degrees are distributed across the whole network.

Unlike the local clustering coefficient, the global clustering coefficient is based on transitivity, which is a measure used to detect the fraction of triplets that are closed in directed networks (Wasserman and Faust, 1994, p. 243). It is not an average of individual fractions calculated for each node, and, as a result, it does not suffer from the same type of correlations with nodes' degrees as the local coefficient. Despite its merits, however, in its original formulation, the global coefficient applies only to networks where ties are unweighted. To address this limitation, and make the coefficient suitable also to networks where ties are weighted, researchers have typically introduced an arbitrary cut-off level of the weight, and then dichotomized the network by removing ties with weights that are below the cut-off, and then setting the weights of the remaining ties equal to one (Doreian, 1969; Wasserman and Faust, 1994). The outcome of this procedure is a binary network consisting of ties that are either present (i.e., equal to 1) or absent (i.e., equal to 0) (Scott, 2000; Wasserman and Faust, 1994). For example, Doreian (1969) studied clustering in a weighted network by creating a series of binary networks from the original weighted network using different cut-offs. To address potential problems arising from the subjectivity inherent in the choice of the cut-off, a sensitivity analysis was conducted to assess the degree to which the value of clustering varies depending on the cut-off (Doreian, 1969). However, this analysis tells us little about the original weighted network, apart from the fact that the value of clustering changes at different levels of the cut-off.

In this paper, we focus on the global clustering coefficient, and propose a generalization that explicitly takes weights of ties into consideration and, for this reason, does not depend on a cut-off to dichotomize weighted networks. In what follows, we start by discussing the existing literature on the global clustering coefficient in undirected and unweighted networks. In Section 3, we propose our generalized measure of clustering. We then turn our attention to directed networks, and discuss the current literature on clustering in those networks. We extend our generalized measure of clustering to cover weighted and directed networks. In Section 5, we empirically test our proposed measure, and compare it with the standard one, by using a number of weighted network datasets. Finally, in Section 6 we summarize and discuss the main results.

2. Clustering coefficient

The global clustering coefficient is concerned with the density of triplets of nodes in a network. A triplet can be defined as three nodes that are connected by either two (open triplet) or three (closed triplet) ties. A triangle consists of three closed triplets, each centered on one node. The global clustering coefficient is defined as the number of closed triplets (or 3× triangles) over the total number of triplets (both open and closed). The first attempt to measure the coefficient was made by Luce and Perry (1949). For an undirected network, they showed that the total number of triplets could be found by summing the non-diagonal cells of a squared binary matrix. The number of closed triplets could be found by summing the diagonal of a cubed matrix. For clarity, we will refer to the global clustering coefficient as the *standard* clustering coefficient *C*:

$$C = \frac{3 \times \text{number of triangles}}{\text{number of triples}} = \frac{\sum \tau_{\Delta}}{\sum \tau} \quad (1)$$

where $\sum \tau$ is the total number of triplets and $\sum \tau_{\Delta}$ is the subset of these triplets that are closed as a result of the addition of a third tie. The coefficient takes values between 0 and 1. In a completely connected network, $C = 1$ as all triplets are closed, whereas in a classical random network $C \to 0$ as the network size grows. More specifically, in a classical random network, the probabilities that pairs of nodes have of being connected are, by definition, independent (Erdős and Rényi, 1959; Solomonoff and Rapoport, 1951). Therefore, *C* is equal to the probability of a tie in these networks (Newman, 2003).

A major limitation of the clustering coefficient is that it cannot be applied to weighted networks. As a result, the same outcome might be attributed to networks that differ in terms of distribution of weights and that, for this reason, might be characterized by different likelihoods of one's neighbors being connected with each other. This limitation could therefore bias the analysis of the network structure. In order to overcome this shortcoming, in the following section we will propose a generalization of the clustering coefficient that explicitly captures the richness of the weights attached to ties, while at the same time it produces the same results as the standard clustering coefficient when ties are unweighted.

[1] Node *i*'s neighbor might be: (1) a node that has directed a tie toward *i*; (2) a node to which *i* has directed a tie; or (3) a node that has directed a tie toward *i* and, at the same time, has received a tie from *i*.