Transitivity correlation

Dekker, David; Krackhardt, David; Snijders, Tom A. B.

Published in:
Network Science

DOI:
10.1017/nws.2019.32

Publication date:
2019

Network Science 7 (3): 353–375, 2019
doi:10.1017/nws.2019.32

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

Transitivity correlation: A descriptive measure of network transitivity

David Dekker*[1], David Krackhardt[2] and Tom A. B. Snijders[3,4]

[1]Center for Business Network Analysis, University of Greenwich, London, UK, [2]Heinz College and the Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA (email: krack@cmu.edu) [3]Nuffield College, University of Oxford, Oxford, UK and [4]Department of Sociology, University of Groningen, Groningen, The Netherlands (email: Tom.Snijders@nuffield.ox.ac.uk)
*Corresponding author. Email: david.dekker@gmail.com

Action Editor: Stanley Wasserman

Abstract

This paper proposes that common measures for network transitivity, based on the enumeration of transitive triples, do not reflect the theoretical statements about transitivity they aim to describe. These statements are often formulated as comparative conditional probabilities, but these are not directly reflected by simple functions of enumerations. We think that a better approach is obtained by considering the probability of a tie between two randomly drawn nodes, conditional on selected features of the network. Two measures of transitivity based on correlation coefficients between the existence of a tie and the existence, or the number, of two-paths between the nodes are developed, and called "Transitivity Phi" and "Transitivity Correlation." Some desirable properties for these measures are studied and compared to existing clustering coefficients, in both random (Erdös–Renyi) and in stylized networks (windmills). Furthermore, it is shown that in a directed graph, under the condition of zero Transitivity Correlation, the total number of transitive triples is determined by four underlying features: size, density, reciprocity, and the covariance between in- and outdegrees. Also, it is demonstrated that plotting conditional probability of ties, given the number of two-paths, provides valuable insights into empirical regularities and irregularities of transitivity patterns.

Keywords: network transitivity; transitive triples; transitivity covariance; transitivity correlation; transitivity Phi; clustering coefficient

1. Introduction

The literature contains various measures for the degree of transitivity in a network. The most well known are the transitiviy coefficient, also called the cluster coefficient, introduced by Harary & Kommel (1979), and its close cousin, the local cluster coefficient of Watts & Strogatz (1998). While these measures have direct interpretations, they also are limited in the ability to assess the most critical question—that is, to what extent does the network's two-step paths $i \to j \to k$ induce an increased propensity for a direct tie $i \to j$, to complete the transitive triple?

Transitivity is the qualitative aspect of the transitive triple configuration (e.g., Holland & Leinhardt, 1976) that occurs when there is a tie between an ordered pair of nodes i and j, and there exist at least one node k, such that there are directed ties from i to k, and from k to j. Transitivity measures have been developed to measure the frequency of transitive triples in networks (e.g., Holland & Leinhardt, 1970; Frank, 1980). Newman et al. (2001) writes: "Clustering refers to the *increased propensity* of pairs of people to be acquainted with one another if they have another

acquaintance in common" [p. 026118-12] (*italics added*). Corresponding to this, but using the more traditional term "transitivity" rather than "clustering," we refer to "network transitivity" as the network-level property that captures the increased propensity of pairs of nodes to be directly connected when they are connected through an intermediary node.

The motivation of this paper resides in the fact that although network transitivity has often been quantified to describe theoretical processes, this quantification has been separate from those theories. As noted by Holland & Leinhardt (1976), many theoretical statements in sociology and psychology are framed in terms of transitive triple configurations. In fact, various strands of scientific literature (e.g., biology, sociology, social psychology, and physics) assign an important role to this configuration. Yet, theories explicitly or implicitly using the concept of network transitivity refer to the frequency of transitive triples being higher *relative to* some null situation, for example, bridging in sociology, epidemiology or diffusion in biology, medicine, and marketing science. This means it is not enough to consider the frequency or relative frequency of transitive triples in a network, but a measure needs to capture the increase in frequency compared to a situation without transitivity. Here a set of measures is developed, presented, and analyzed that can capture the increased propensity of transitive triples in networks. To prevent confusion, it should be noted that this set of measures is purely descriptive, and distinctive from the larger body of literature on inferential statistics for structural dependencies in network models (see, e.g., Goodreau et al., 2009). This paper does not deal with inferential statistics. No assumptions of underlying graph models are made in the development of the measures. Probabilities in this article relate solely to those in empirical networks.

It is important to distinguish this set of measures from other related concepts, especially, clustering. In social sciences, micro processes are considered at the basis of many macro social phenomena we can observe (e.g., Coleman, 1990; Faust, 2010; Schelling, 1978). A ubiquitous construct in studying consequences of such processes is the concept of clustering in social networks. Clustering is the extent to which there are regions in the network that have a higher density than there is between these regions. In some of the literature, clustering has been equated to network transitivity. For example, in their pivotal paper, Newman et al. (2001, p. 026118-12) state: "...clustering in social networks [is] also sometimes called network transitivity." However, the two concepts do differ, as is implied by the definitions given above. This paper recognizes the distinction between clustering and transitivity and explores the property of transitivity.

The ideas in this paper also build on the work of authors (e.g., Holland & Leinhardt, 1970; Wasserman, 1975; Feld & Elmore, 1982; Faust, 2007) who have shown that the triad censuses of networks are highly associated with lower-order properties (nodal and dyadic). This implies that descriptive network transitivity measures should control for the lower-order properties. The purpose of this paper is to elaborate on this.

In this paper, first a theoretical development of network transitivity measures is presented. Second, existing and new measures of network transitivity are defined and their properties described. Third, behavior of different measures is compared in examples (stylized, random, and empirically observed networks). Subsequently, findings and further research opportunities are discussed, and conclusions presented.

2. Network transitivity as a comparative quantity

Formalizing and generalizing statements about lower-order network descriptives allow us to make statements about substructures in whole systems. These statements are essential in allowing to make empirical descriptions of theoretical processes on the level of a whole network. The gap between local observations and the global nature of much social theory can be bridged "... by examining local structural properties and determining whether they hold, on the average, across entire social systems" (Holland & Leinhardt, 1976, p. 3). Hence, average occurrence of local structures (or substructures) is considered an important descriptive statistic of whole systems as it

allows to link social structure to global theoretical statements. Transitivity is such a lower-order network descriptive which plays an important role in social theory (e.g., Granovetter, 1973).

Transitivity is a property of ordered labeled 3-subgraphs (Holland & Leinhardt, 1971, 1972) or triples. It thus not only plays a role as conceptual configuration in sociological theory, it is also an attractive statistical concept for network modeling. Its theoretical importance in much of social science stems from a Heiderian view that transitivity occurs in social interactions at a rate that is in excess of what we would expect by chance (Holland & Leinhardt, 1971, p. 108). This has led to statistical modeling of the frequency of transitive triples under a variety of null models (e.g., Holland & Leinhardt, 1971; Frank, 1988; Karlberg, 1999).

Another view, deviating from the approach that focuses on enumerating transitive triples, can be derived from Newman et al. (2001). They define network transitivity as "... the *increased propensity* of pairs of people to be acquainted with one another if they have another acquaintance in common" [p. 026118-12] (*italics added*). Here, the concept of network transitivity is not reflected by a mere average measure of transitive triples, but rather an *average increased propensity* to form transitive triples. This definition suggests measuring an intrinsic comparative transitivity quality of a network.

In the literature, transitivity is measured usually as the ratio of transitive to potentially transitive triples (Harary & Kommel, 1979; Frank, 1980; Karlberg, 1999) or as the average density of personal networks (Watts & Strogatz, 1998; Newman et al., 2001). These measures, based on relative frequencies, do not reveal much about an *increased propensity*, as they do not entail a comparison.

"Network transitivity" quantifies a statement about the *comparative* frequency of transitive triples among relevant triples in the network. It reflects a structural hypothesis that refers to an elevated conditional probability of ties between a pair of nodes, given the existence, or the number, of two-paths connecting them.

To define such a comparison, for a given observed digraph with n nodes, we rely on two simple probability distributions. We stress that these are not probability distributions for the network; that is, they are not defined on the outcome space of all digraphs. Rather, they are empirical distributions, defined by the observed network, focusing on a single—randomly chosen—tie variable. In Section 3.1, we use the probability distribution consisting of the random choice of an ordered triple (i, j, k) of vertices $(i \neq j, i \neq k, j \neq k)$ from the total of n vertices. The probability distribution used in Section 3.2 is the random choice of a pair of vertices.

3. Measurement of transitivity

In this section, we define various measures that express the comparative frequency of transitive triples in a network. We denote the digraph by x, with the variables x_{ij} being the dichotomous (0/1) indicators of the existence of the ties $i \rightarrow j$, for nodes i and j. Per usual, self-ties are excluded ($x_{ii} = 0$ for all i).

3.1 Difference in conditional probability and centered clustering coefficient

For the first empirical probability distribution, we consider, for a randomly chosen ordered triple (i, j, k), the triple of tie variables x_{ij}, x_{ik}, x_{kj}. Formally, this empirical distribution corresponds to a random choice from the outcome space

$$\{(x_{ij},\ x_{ik},\ x_{kj}) \mid 1 \leq i, j, k \leq n; i \neq j \neq k, i \neq k\} \tag{1}$$

where x is the observed network; this outcome space has $n(n-1)(n-2)$ elements. Probabilities under this empirical probability distribution will be denoted by p. The basic comparison is given by the difference between conditional probabilities of a tie, given a two-step path, and a tie given no two-step path,

$$p(x_{ij} = 1 \mid x_{ik}x_{kj} = 1) - p(x_{ij} = 1 \mid x_{ik}x_{kj} = 0) \tag{2}$$

Table 1. Transitivity joint and marginal probabilities

		$x_{ik}x_{kj}$		
		1	0	
x_{ij}	1	p_{11}	p_{10}	p_{1+}
	0	p_{01}	p_{00}	p_{0+}
		p_{+1}	p_{+0}	1

where a positive difference demonstrates an increased propensity toward transitivity. This difference reflects the most relevant alternative to the configuration central to the definition of Newman et al. (2001), namely, the configuration where pairs of people are acquainted with one another if they have *no* other acquaintance in common.

For transitivity as a purely descriptive statistic, a common definition is the ratio of transitive to potentially transitive triples (e.g., Wasserman & Faust, 1994), as proposed by Harary & Kommel (1979):

$$C = \frac{\sum_i \sum_{j\neq i} x_{ij} \sum_{k\neq i,j} x_{ik}x_{kj}}{\sum_i \sum_{j\neq i} \sum_{k\neq i,j} x_{ik}x_{kj}} = \frac{\sum_i \sum_{j\neq i} \sum_{k\neq i,j} x_{ij}x_{ik}x_{kj}}{\sum_i \sum_{j\neq i} \sum_{k\neq i,j} x_{ik}x_{kj}} \tag{3}$$

If the network is nondirected, this is equal to the well-known formula

$$C = \frac{3 \times \text{number of triangles in the graph}}{\text{number of connected triples of vertices}} \tag{4}$$

coined the clustering coefficient by Newman et al. (2001). This is equal to the first term in Equation (2),

$$C = p(x_{ij} = 1 \mid x_{ik}x_{kj} = 1) \tag{5}$$

Comparing Equations (2) and (5) immediately shows that Equation (5) is only a partial expression of theoretical statements about network transitivity, because it lacks a comparative aspect.

Another measure for transitivity is the clustering coefficient defined by Watts & Strogatz (1998) as the mean of local transitivity around the nodes. The version for digraphs is given by

$$LC = \overline{LC_i} = \frac{1}{n}\sum_i \frac{\sum_{j\neq i}\sum_{k\neq i,j} x_{ij}x_{ik}x_{kj}}{OD_i(OD_i - 1)} \tag{6}$$

where $OD_i = \sum_h x_{ih}$ is the outdegree of node i. Just like Equation (3), however, this is not a comparative measure.

To develop a measure that does have a comparative nature, just like Equation (2), we present the two-by-two table for the two random variables x_{ij} and $x_{ik}x_{kj}$ under the empirical probability distribution of randomly drawing a triple (i, j, k). Here x_{ij} indicates the existence of a direct tie between i and j and $x_{ik}x_{kj}$ indicates the existence of a two-path, that is, an indirect connection. The cells in Table 1 contain joint probabilities, while the row and column sums give marginal probabilities, respectively. The joint probability's first index indicates whether $x_{ij} = 1$ or 0, while the second index indicates whether $x_{ik}x_{kj} = 1$ or 0. For example, p_{11} is the joint probability of a tie between the pair (i, j) and a two-step between this pair via k. In the marginal entries, a plus (+) indicates summing over both joint probabilities. For example, p_{1+} is the marginal probability of a tie, which is the sum of the joint probabilities of a tie and a two-path, and a tie and no two-path.

By the definition of conditional probability, Equation (2) is equal to $p_{11}/p_{+1} - p_{10}/(1 - p_{+1})$. It is well known that this difference is the bivariate linear regression coefficient for dichotomous

data (see Falk & Well, 1997, for an excellent exposition). We use this expression to define Equation (2) as *TPB* (Transitivity Phi Beta):

$$TPB = \frac{p_{11}}{p_{+1}} - \frac{p_{10}}{(1 - p_{+1})} \tag{7}$$

A bivariate regression coefficient is equal to the covariance between the two variables divided by the variance of the explanatory variable. This implies that another expression is

$$TPB = \frac{\mathrm{cov}(x_{ij}, x_{ik}x_{kj})}{\mathrm{var}(x_{ik}x_{kj})}, \tag{8}$$

where again the variance and covariance are with respect to the probability distribution of randomly drawing a triple of nodes from the digraph.

This expression emphasizes that centering is the major difference with existing measures. The numerator in Equation (8), which we shall call Transitivity Covariance, is by definition a centered measure for the joint occurrence of ties and two-paths in an observed digraph. The measures in Equations (3) and (6) clearly are not centered. This centering is essential for the comparative nature of our measure for network transitivity.

A major advantage of centering is that it yields the value of 0 if there is no network transitivity in the sense that the existence of a two-path is not associated with the existence of a direct tie. For dichotomous variables, a covariance of 0 is equivalent to independence; therefore, our transitivity measure *TPB* is 0 if and only if, in case a triple (i, j, k) is randomly drawn, the existence of the direct tie $i \rightarrow j$ is independent of the existence of the two-path $i \rightarrow k \rightarrow j$. A direct expression for the Transitivity Covariance is the centered joint probability

$$\mathrm{cov}(x_{ij}, x_{ik}x_{kj}) = \frac{1}{n(n-1)(n-2)} \sum_i \sum_{j \neq i} \sum_{k \neq i,j} x_{ij}x_{ik}x_{kj} - \bar{x} \cdot \overline{xx} \tag{9}$$

where $\bar{x}$ is the proportion of ties, or density in the digraph, and $\overline{xx}$ is the proportion of two-paths among all triples of nodes in the digraph.

Another measure can be obtained as the bivariate correlation coefficient instead of the regression coefficient. For this measure, the Transitivity Covariance is divided not by the variance of the two-path indicator but by the product of the two standard deviations. As both variables are dichotomous, the Pearson product-moment correlation coefficient is also known as the *Phi* coefficient (Falk & Well, 1997). Here, we use the term "Transitivity Phi,"

$$TPhi = \frac{\mathrm{cov}(x_{ij}, x_{ik}x_{kj})}{\sqrt{\mathrm{var}(x_{ij})\mathrm{var}(x_{ik}x_{kj})}} \tag{10}$$

The obvious further advantage of this measure is that it is bounded between −1 and +1.

3.2 Correcting for two-path autocorrelation

The measures proposed in the preceding section do not take into account the multilevel issue that for each pair (i, j) there are $n - 2$ potential vertices k, which play a different role in the triple than i and j. The "clustering" of two-paths through specific k's for a given (i, j), which may be called the autocorrelation between different two-paths connecting the same pair (i, j), is ignored.

Considering the set of all potential "third" vertices k leads to an interest in the relation between the total number of two-paths connecting i and j, and the existence of a direct tie $i \rightarrow j$. Therefore, we now turn to the empirical distribution under consideration which is based on the random draw of an ordered pair (i, j), where the outgoing ties of i and the incoming ties of j from and to other

nodes are also taken into consideration. This distribution is defined as a random choice from the outcome space

$$\left\{\left(x_{ij},\ (x_{ih})_{1\leq h\leq n,\, h\neq i,j},\ (x_{hj})_{1\leq h\leq n,\, h\neq i,j}\right) \mid 1\leq i,j\leq n; i\neq j\right\} \tag{11}$$

where again x is the observed network. This outcome space has $n(n-1)$ elements. To distinguish this from the model of the preceding section, we indicate the other nodes by the letter h, distinguishing them from the single third node k in the preceding section. Accordingly, we define the Transitivity Correlation[1] by

$$TC = \frac{\operatorname{cov}\left(x_{ij}, \sum\limits_{h\neq i,j} x_{ih}x_{hj}\right)}{\sqrt{\operatorname{var}(x_{ij})\operatorname{var}\left(\sum\limits_{h\neq i,j} x_{ih}x_{hj}\right)}} \tag{12}$$

The relation between *TPhi* and *TC* is derived in Appendix (A) and indeed depends on the two-path autocorrelation. Also, note that Equation (12) depends on the information contained in the outcome space defined in Equation (11).

The other measure, similar to *TPB* in Equation (8), replaces variance in two-paths for ordered triples (i,k,j) with the variance of the number of two-paths for ordered node pairs (i,j). This is the bivariate regression coefficient of ties on the number of two-paths between ordered pairs (i,j):

$$TB = \frac{\operatorname{cov}\left(x_{ij}, \sum\limits_{h\neq i,j} x_{ih}x_{hj}\right)}{\operatorname{var}\left(\sum\limits_{h\neq i,j} x_{ih}x_{hj}\right)} \tag{13}$$

This slope gives a linear approximation of the conditional probability of a tie, given the number of two-paths. As such it is more informative about the increased propensity toward transitivity than for example the clustering coefficient C in Equation (3), which gives an mean conditional probability over all two-path counts.

At this point, it should be re-emphasized that the expected values, covariances, etc., referred to in this paper are those of ties between randomly chosen vertices in an observed network, not those of possible underlying random graph processes. A disadvantage of this is that the measures discussed above cannot be used for statistical inference without nontrivial additional assumptions. What is subtracted in centering is not the expected value under a null model for networks. As shown in the next section, a necessary condition for $TC=0$ and $TPhi=0$ is that the number of two-paths is a specified function of n, and the observed density, mutuals, and covariance between in- and outdegrees.

However, there are random graph processes that do generate an expected value of $TC=0$ and $TPhi=0$. For example, in Erdös–Renyi digraphs we have

$$E\left\{X_{ij}\sum_{h\neq i,j} X_{ih}X_{hj}\right\} = E\{X_{ij}\}\,E\left\{\sum_{h\neq i,j} X_{ih}X_{hj}\right\} \tag{14}$$

This does show that the expected value of the numerator in the covariance measures under the Erdös–Renyi digraph model is 0. Since what is subtracted takes account of the indirect connections, this centering is more subtle than the null expected value under the Erdös–Renyi digraph.

Further, we note that *TC* and *TPhi* differ only in the denominator, that is, the standardization. Therefore, one way of studying the differences between these measures is to consider the digraphs for which *TC* or *TPhi* are -1 or $+1$ if such digraphs exist.

Digraphs that are unions of complete subgraphs, to which also isolated points may be added, are completely transitive in the sense that $C = 1$. If all these subgraphs have the same size, then also $TC = TPhi = 1$. However, if the subgraph sizes are different, then TC and $TPhi$ are less than 1.

4. Behavior of transitivity covariance measures

In assessing the utility of these centered measures, we look at the behavior in comparison to existing clustering coefficients. Much of it will depend on the properties of Transitivity Covariance when we know it to be zero.

4.1 Descriptive mathematical properties

Empirical studies find that the frequency of triangles in a network is to a large extent accounted for by lower-order network properties (e.g., Faust, 2007). If it is totally accounted for by lower-order properties, it may be expected that transitivity covariance is close to zero. Then it would be concluded that there is no "increased" (or decreased) propensity toward transitivity. The mean number of transitive triples over all ordered pairs (i, j) is given by

$$TT = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{h=1 \\ h \neq i,j}}^{n} x_{ij}\, x_{ih}\, x_{hj} \tag{15}$$

For any digraph, the condition $TC = 0$ is equivalent to

$$TT = \frac{\left(n\ \mathrm{cov}(OD, ID) + n\, d^2 - M\right) d}{n(n-1)} \tag{16}$$

(for a proof, see Appendix B). Equation (16) expresses the necessary and sufficient condition for a zero correlation between x_{ij} and $x_{ik}x_{kj}$. Therefore, if Equation (16) holds, which is equivalent to $TC = 0$, no elevated or decreased propensity to transitivity may be said to exist in the network. The number of transitive triples then is determined by a function of the number of nodes n and three other statistics: density (proportion of ties), reciprocity (number of mutual ties, M), and covariance between the degree distributions, $\mathrm{cov}(OD, ID)$.

The fact that $TC = 0$ implies a conditioning on $\mathrm{cov}(OD, ID)$ relates to the observation of Feld & Elmore (1982, p. 77), who observe that "... inequality of popularity among individuals implies disproportionate frequencies of particular types of triads, including transitive triad types." They do not make clear how the "increased propensity" of transitive triples depends on degree. Transitivity covariance does control for such popularity-induced transitivity as it incorporates the covariance between in- and outdegrees.

4.2 Telling problem: Don Quichot measures and windmills

An example that illustrates the problems with different measures for an increased propensity of transitive triples in networks is given by structures named windmill graphs (see, e.g., Jackson 2008). A windmill graph has one center node connected to all other nodes, while all other nodes are in "wings," which are even-sized cliques where all nodes are connected within wings, but not to any other node (except the center node). Windmills W_m^r are characterized by two parameters: the size of each wing ($r > 2$) and the number of wings ($m > 1$) (see, e.g., Figure 1).

In such graphs, there are either 1 or $(r - 2)$ two-paths between each pair of nodes, where the latter are always part of a transitive triple, while the former are not. Given this morphological

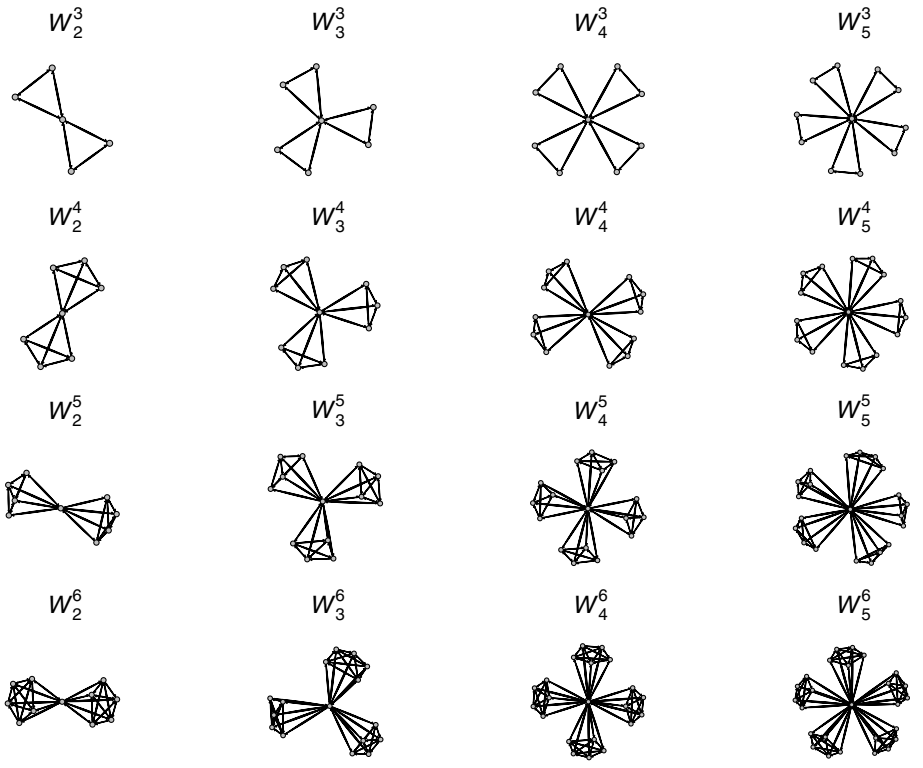


Figure 1. Windmills (W_m^r) with different values for r and m.

constriction, windmills provide an experimental model that allows us to vary the number of non-transitive two-paths and transitive triples as a function of r and m. The number of two-paths in a windmill is given by

$$W_{TS} = m\,r\,(r-1)\,(r-2)\;+\;m\,(m-1)\,(r-1)^2 \tag{17}$$

where the first term on the right-hand side is the number of transitive triples, while the second term gives the number of intransitive triples. The latter increases more strongly in m as it is a quadratic polynomial, while the former is linear in m. The opposite holds for r as the number of transitive two-paths increases cubically, and the intransitive triples quadratically, in r. Hence, this network model allows one to manipulate the total degree of network transitivity.

To assess the behavior of network transitivity measures on the windmill model, we express them in terms of m and r. Table 2 summarizes these expressions, as well as their behavior in the limit when either or both approach infinity. It is important to recall that the clustering coefficients in Equations (3) and (6) give the conditional probability of a transitive triple and the *mean* conditional probability of a transitive triple occurring in a neighborhood, respectively. The contradictory effects of increases in m and r result in an undefined value for the clustering coefficient (C) in the bivariate limit, while it behaves as expected in the univariate limits. As r increases C tends toward 1, while it tends to 0 with increasing m.

Similarly, the local clustering coefficient (LC), in the limit, reflects that, apart from the single central node, all neighborhoods are cliques where all two-paths are transitive, so that it tends toward 1. However, LC firmly contradicts C as $m \to \infty$, which was first noted in Jackson (2008, pp. 36–37).

Table 2. Transitivity measures for windmills.

Measure	$f(m,r)^*$	$\lim_{r\to\infty} f$	$\lim_{m\to\infty} f$	$\lim_{(r,m)\to(\infty,\infty)} f$
Clustering coefficients:				
C	$\frac{r(r-2)}{r(r-2)+(m-1)(r-1)}$	1	0	Undefined
LC	$1-\frac{n-r}{n}\frac{1}{n-2}$	1	1	1
Composites for covariance-based measures:				
$Var(x_{ij})$	$\frac{r(n-r)}{n^2}$	$\frac{m-1}{m^2}$	0	0
$Var(x_{ik}x_{kj})$	$\frac{((n-1)(n-2)-(r-1)(r-2))(n+r(r-3))}{n^2(n-2)^2}$	$\frac{m^2-1}{m^4}$	0	0
$Var(\sum_h x_{ih}x_{hj})$	$\frac{(m-1)r(r-1)(r-3)^2}{n^2}$	∞	0	Undefined
$Cov(x_{ij}, x_{ik}x_{kj})$ **	$\frac{(m-1)r(r-1)(r-3)}{n^2(n-2)}$	$\frac{m-1}{m^3}$	0	0
Covariance-based measures (triadic probability model):				
TPhi	$\frac{\sqrt{(r)(r-3)}}{\sqrt{(nr+r(r-3))(n+r(r-3))}}$	$\frac{1}{\sqrt{m+1}}$	0	0
TPB	$\frac{(n-2)r(r-3)}{(n+r-3)(n+r(r-3))}$	$\frac{m}{m+1}$	0	Undefined
Covariance-based measures (dyadic probability model):				
TC	1	1	1	1
TB	$\frac{1}{r-3}$	0	$\frac{1}{r-3}$	0

*$n = m(r-1)+1$, **Note that $\lim_{(r,m)\to(\infty,\infty)} (n-2)Cov(x_{ij}, x_{ik}x_{kj})$ is undefined.

In windmills, the transitivity covariance-based measures, which are weighted functions of *TPB* in Equation (8), illustrate another important distinction. For increasing wing size r, the difference in conditional probabilities (*TPB*) still depends on the number of wings, m. On the other hand, when m grows, the difference tends toward 0 irrespective of r. The multivariate limit is undefined as it will depend on the asymptotic ratio m/r.

TPhi is restricted to $[-1, 1]$ as it is a correlation coefficient. In the limit in r it becomes a decreasing function of m, and approaches 0 for increasing m. This is the correlation between x_{ij} and $x_{ik}x_{kj}$ for a random triple (i, j, k). For an increasing number of wings m, the conditional probability of the two-path through a random k between a given pair of nodes, that is, $p\{x_{ik}x_{kj} = 1 \| i, j\}$, tends to zero for all pairs (i, j); this implies that the correlation tends to 0. The consideration of a random third node does not bring out the clustering pattern for windmills with many wings, and therefore this pattern yields approximately a zero correlation.

The covariance-based measures that weight on bases of the cumulative number of two-paths, *TC* and *TB*, do signal this autocorrelation. First, *TC* as a bounded measure on $[-1, 1]$ is a constant 1, reflecting the perfect control for the morphological similarity of different size windmills. It indicates the perfect correlation that occurs in these structures, where the presence of a tie implies $(r-2)$ two-paths, while lack of a tie implies 1 two-path—the regularity that defines windmills.

In the limit *TB* in windmills tends to 0. The decline in the ratio of transitivity covariance (based on the number of two-paths) and the variance of the number of two-paths is due to the fact that $Var(\sum_h x_{ih}x_{hj})$ is a factor $(r-3)$ larger than $Cov(x_{ij}, \sum_h x_{ih}x_{hj})$. This shows that for the value of *TC* a direct interpretation is more clear than for *TB*.

4.3 Erdös–Renyi random digraphs

The stylized example on windmills in the previous section shows that a family of morphologically similar networks can produce measures that are undefined in the limit, while they may give

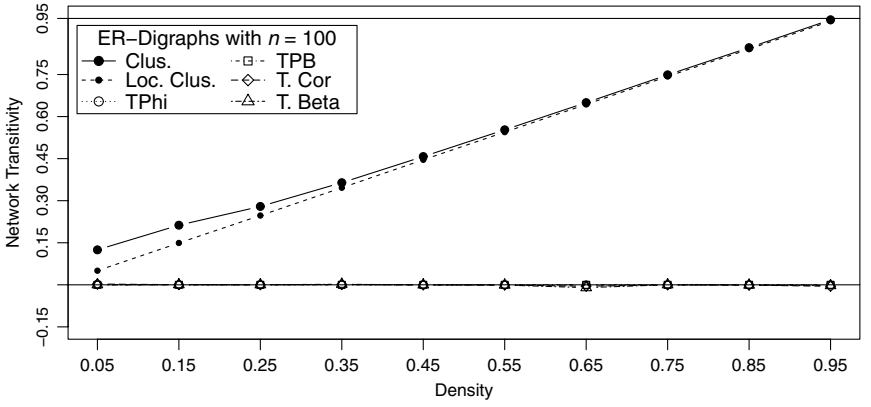


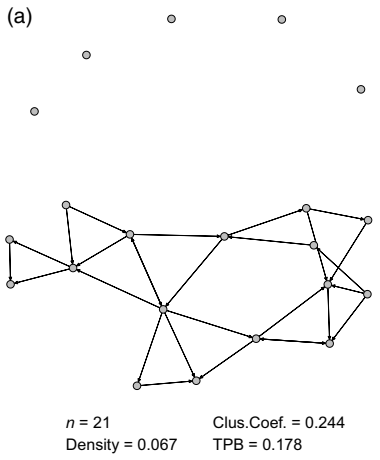
Figure 2. Means of Transitivity Covariance-based measures and means of the clustering coefficients plotted as a function of density in random Erdös–Renyi digraphs ($n = 100$). It becomes immediately clear that the clustering coefficients increase with density, while there certainly is no increased propensity of these random networks to form transitive triples. The transitivity covariance-based measures remain stable around zero as density increases. Each point is a mean value of each of the measures based on 100 random draws of Erdös–Renyi digraphs with given density.

ambiguous readings for small networks. This is not a desirable property. However, in practice other families of networks may be more important to consider.

If there is known to be independence between ties and two-paths in a network, there would not be expected any elevation or increased propensity in transitive triples, or network transitivity. Here we compare the behavior of different measures for Erdös–Renyi digraphs (Erdös & Renyi, 1959). In these networks, all ties are independent, and the probability for a tie is constant, determining the expected density. Hence, within this family of networks on average we do not expect to find any increased propensity for transitive triples to occur. Consequently, on average a transitivity measure should be independent of the density, in other words, control for the density. Through simulations, we first analyze the dependence of different measures on density. Figure 2 shows the results of these analyses for Erdös–Renyi digraphs. It shows that the covariance based measures are, as expected, independent of density; while *C* and *LC* are, respectively, linear and nonlinear functions of density (for the latter result, see also Newman, 2003).

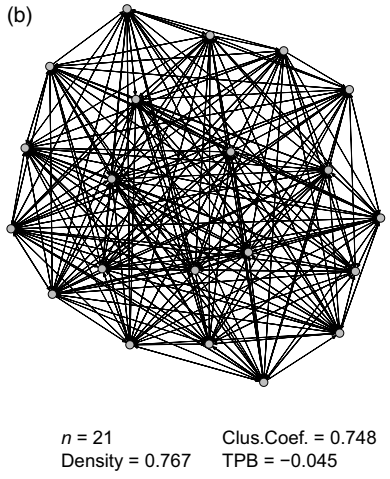
To illustrate the interpretive difference in the clustering coefficient measures and the class of measures proposed in this paper, consider the two digraphs represented in Figure 3. Both are composed of 21 nodes. Each has an abundance of ties and of two-paths, some of which are completed, forming a transitive triple. However, the differences revealed by the two measures is profound. In the first network in Figure 3(a), the clustering coefficient is a modest 0.244, signifying that 24% of the directed two-paths are completed with a directed tie to make the triple transitive. The Transitivity Phi Beta (*TPB*) coefficient also indicates a moderate positive association between the number of two-paths and the probability of a completing directed tie, compared to the probability of a tie when no directed two-path exists. These two statistics are consistent in the story they tell: Network Transitivity is evident.

In looking at the second network in Figure 3, however, a demonstrably different picture emerges from the two measures. The clustering coefficient of 0.748, indicating that better than 74% of the two-paths are completed with a direct tie that makes the triple transitive, suggests a substantially higher rate of transitivity is exhibited in this network (Figure 3(b)) than in the previous network (Figure 3(a)). While this sounds like a victory for the forces of transitivity, the Transitivity Phi Beta however sports a small negative value, −0.045. The reason for this switch is because the apparent high rate of completed two-paths is in fact lower than the probability of a tie given that no two-path around the directed tie exists. That is, having a two-path from *i* to *j* makes



		$x_{ik}x_{kj}$ = 1	$x_{ik}x_{kj}$ = 0	
x_{ij}	1	10	522	532
	0	31	7417	7448
		41	7939	7980

Difference between conditional probabilities equals *TPB*, where the first conditional probability equals the clustering coefficient: $(p(x_{ij} = 1|x_{ik}x_kj = 1) - p(x_{ij} = 1|x_{ik}x_kj = 0) = 0.244 - 0.066 = 0.178$



		$x_{ik}x_{kj}$ = 1	$x_{ik}x_{kj}$ = 0	
x_{ij}	1	3518	2600	6118
	0	1184	678	1862
		4702	3278	7980

Difference between conditional probabilities equals *TPB*, where the first conditional probability equals the clustering coefficient: $(p(x_{ij} = 1|x_{ik}x_kj = 1) - p(x_{ij} = 1|x_{ik}x_kj = 0) = 0.748 - 0.793 = -0.045$

Figure 3. Examples of positive and negative values for network transitivity ($n = 21$).

it less likely that i will be directly connected to j in this network. The clustering coefficient hides this critical interpretive fact.

To further delineate how this happens, we compare the actual count of triples in each graph. The 2×2 tables aside each network in Figure 3 demonstrate these relationships. There are 7,980 ordered triples in each graph. These triples are distributed among the four cells in the 2×2 tables: those triples where an $i \to k \to j$ exists and the triple is completed with an $i \to j$ tie to make it transitive; those triples where an $i \to k \to j$ exists, but no direct $i \to j$ exists (making the triple clearly intransitive); those triples where an $i \to k \to j$ does not exist, but a direct tie $i \to j$ exists anyway; and those triples where neither an $i \to k \to j$ nor an $i \to j$ tie exists.

In Figure 3(a), the proportion of two-paths completed with direct ties is $10/41 = 0.244$, the clustering coefficient. This is a modest degree of transitivity. But, we note that the conditional probability of a tie existing given that no two-path exists is $522/7,939 = 0.066$. If we subtract out this baseline value from the proportion of two-paths, we see that the advantage for having the two-path $i \to k \to j$ in the triple, is that it increases the probability that a direct tie $i \to j$ will exist by 0.178 (0.244 − 0.066), which is the value of the Transitivity Phi Beta. Thus, a straightforward interpretation of the *TPB* is that it exactly captures the added propensity for a directed tie given that a two-path exists compared to the case where no two-path exists.

The equivalent 2×2 table for the network in Figure 3(b) shows a considerably different pattern. The network is much denser, and commensurately the total number of two-paths is much larger (4, 702). The proportion of two-paths that are completed into transitive triples is $3,518/4,702 = 0.748$, the clustering coefficient for this network. However, the proportion of $i \rightarrow k \rightarrow j$ triples where there is no two-path, but nonetheless have a direct tie $i \rightarrow j$ is $2,600/3,278 = 0.793$. That is, the conditional probability that a tie exists given that a two-path exists is lower than if a two-path does not exist, yielding a negative *TPB*, $0.748 - 0.793 = -0.045$. Again, the useful and direct interpretation of this negative *TPB* is that it describes the precise reduction in proportion of ties that exist due to the existence of a two-path.

As appealing and as intuitive as this interpretation of the *TPB* is, it hides one other factor that is useful to researchers as they explore the tendencies toward induced transitivity in the structures they are studying. In the *TPB*, each triple is treated as an independent case. Left open is the question of whether the conditional probability of a directed pair being tied is a function not only of whether a two-path exists, but also the number of two-paths that exist between *i* and *j*.

The graphs in Figure 4 show how *TC* helps to capture this effect (see Appendix A on the relation between *TC* end *TPhi*). Again, we repeat the same two networks as in Figure 3, but this time we plot the proportion of ordered dyads that are tied as a function of the number two-paths surrounding each dyad. In the case of the low-density graph (Figure 4(a)), we find that each ordered pair has either 0, 1, or 2 two-paths indirectly connecting them. The size of the points in the plot is proportional to the log of the number of instances of pairs of nodes that have that many two-paths. So, for example, we see in this plot that the majority of ordered pairs has no two-paths; a smaller number has 1 two-path; and only a tiny fraction has 2 two-paths. As can be seen in the plot, almost none of the ordered pairs with 0 two-path are connected; around 20% of the ordered pairs with 1 two-path are connected; and all the pairs (although there is a small number of them) which are surrounded by 2 two-paths are connected directly. This relationship is summarized by the TC of 0.228, again, a modest but not insubstantial correlation. This leads to a similar conclusion the Clustering Coefficient would suggest.

In comparison, the plot in Figure 4(b), however, is especially informative. First, the density in the network results in every pair of points having at least 8 two-paths. A few have as many as 15 two-paths. And many two-paths have directed ties associated with them. But, what is most striking is the steep descending relationship they have with the proportion of directed ties. Those with 8 indirect two-paths are almost all connected directly; those with 9 two-paths slightly less so; those with 10 two-paths even less so; and so on until those with 15 two-paths are less than 20% likely to be directly tied. The negative relationship ($TC = -0.33$) between the number of two-paths and the conditional probability of a tie is readily apparent through this graph. This relationship is captured by the measures proposed in this paper, not the clustering coefficients.

4.4 Observed networks

The covariance-based measures of transitivity can be interpreted as linear approximations of the relationship between direct ties and two-path ties. In particular, *TB* is the linear regression coefficient of the tie indicator on the number of two-paths between the node pair. Graphical inspection of this relationship may provide insight about the appropriateness of the linearity assumption. Due to combinatorial restrictions, the relationship may be highly nonlinear, which can be directly assessed from a plot. Example data sets were obtained via public websites[2,3].

In Figure 5(a)–(c), 12 network data sets from different fields are analyzed. Each figure contains a diagram, an associated graph plot, and relevant summary statistics. The diagram shows the conditional probability for a tie given the number of two-paths on the vertical axis, and the number of two-paths on the horizontal axis. Information in the diagram is based on the depicted network although for clarity isolate nodes have been excluded.

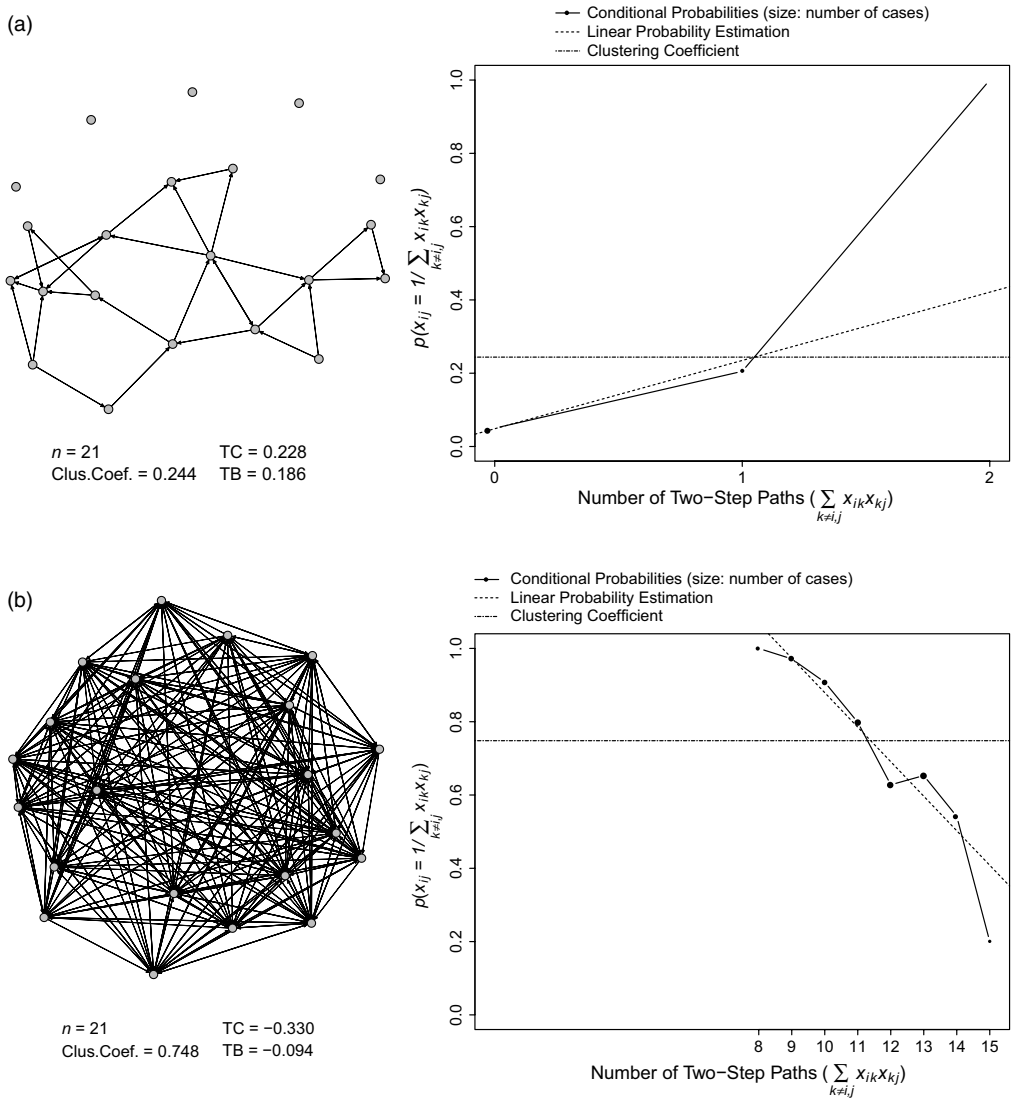
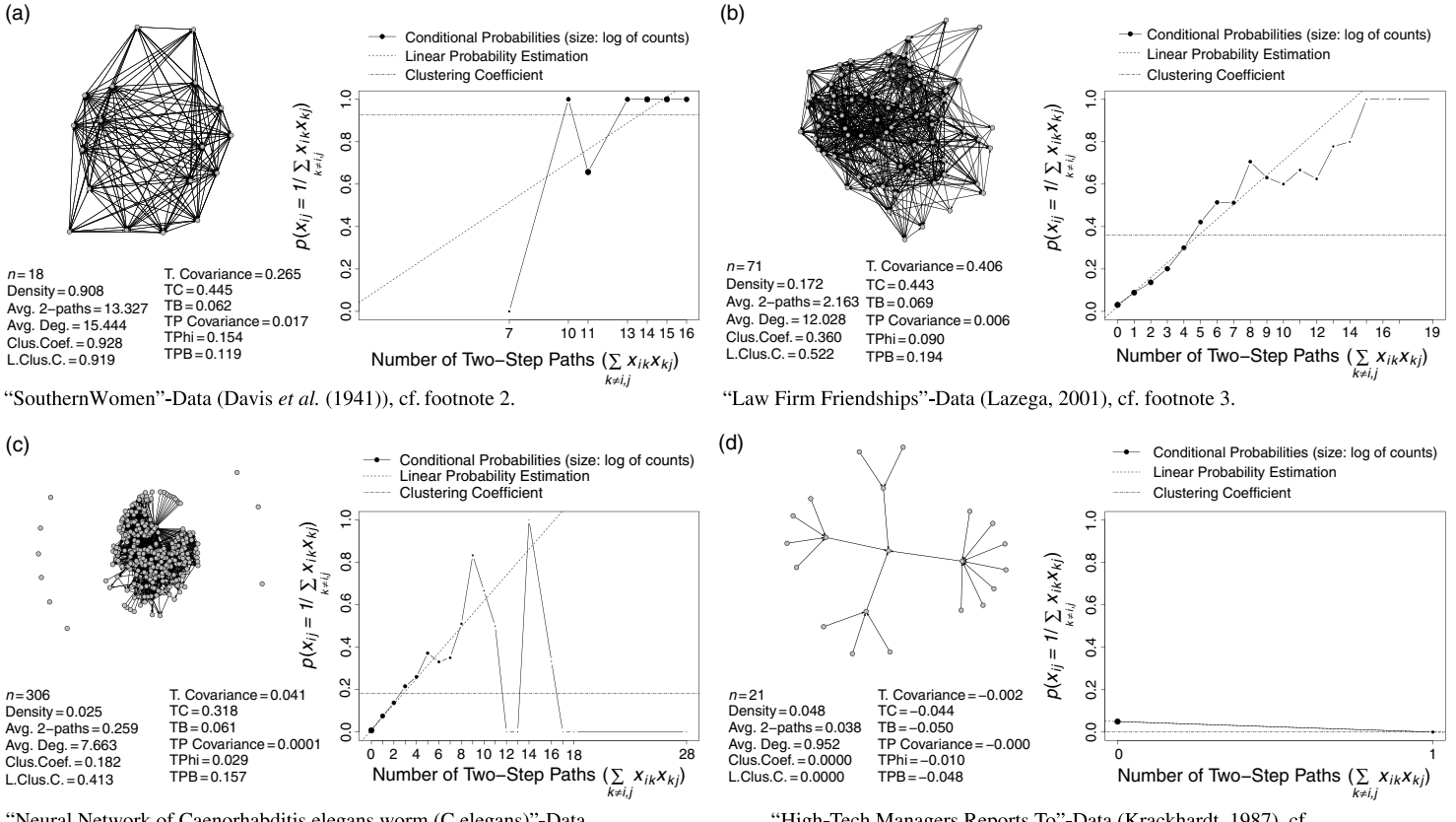


Figure 4. A comparison of two networks and their revealed transitivity structures.

Number of observations (ordered pairs of nodes) in each category of two-path counts is indicated by the size of dots. Each dot is connected with a straight line to emphasize the differences and direction in change of conditional probabilities between categories.

The horizontal dotted line indicates the clustering coefficient (C) for that network. This can be interpreted as the "mean conditional probability" over all categories of two-path counts. By definition, this measure discards all information about the differences between categories of two-path counts.

The dashed linear regression line between ties and number of ordered two-paths gives a linear approximation for these differences. The slope of this line is given by *TB*, which hence allows for a network level indication of an increased (decreased) propensity toward transitivity. A downside is that *TB* does not allow for comparison between networks or a direct interpretation. However, *TC* is a linear transformation of *TB*, which serves these purposes.



"SouthernWomen"-Data (Davis *et al.* (1941)), cf. footnote 2.

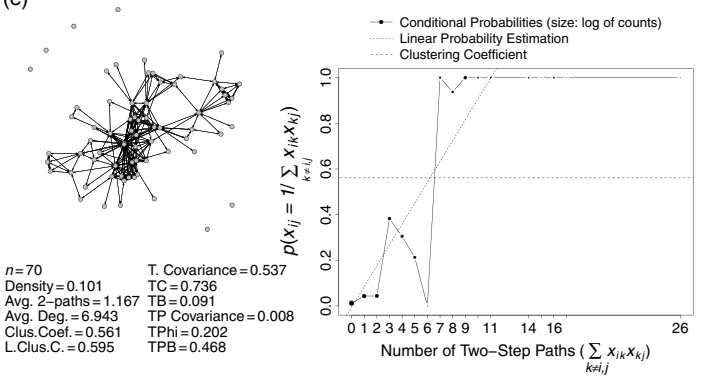
"Law Firm Friendships"-Data (Lazega, 2001), cf. footnote 3.

"Neural Network of Caenorhabditis elegans worm (C.elegans)"-Data (Watts & Strogatz, 1998), Dichotomized, cf. footnote 2.

"High-Tech Managers Reports To"-Data (Krackhardt, 1987), cf. footnote 3.

Figure 5. A. Illustrating 10 descriptive network statistics on 12 data sets. All reported graphs and descriptives are based on $x_{ij} \in \{0, 1\}$. In case data are valued, they are dichotomized by the rule: $x_{ij} = 1 \,\forall\, y_{ij} > 0$ else $x_{ij} = 0$, unless differently indicated. On the vertical axis conditional probabilities for ties are shown, and dot sizes indicate the number of ordered pairs for each count of two-paths (horizontal axis). The horizontal line indicates the clustering coefficient (*Clus.Coef.*) for that network. It can be interpreted as the weighted mean conditional probability over all groups of two-path counts. Third, the linear regression line between ties and number of ordered two-paths is shown. The slope of this line is given by *TB*.

(e)

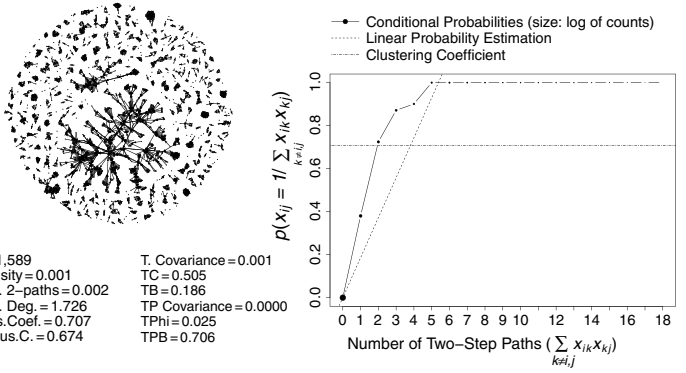


$n = 70$
Density = 0.101
Avg. 2–paths = 1.167
Avg. Deg. = 6.943
Clus.Coef. = 0.561
L.Clus.C. = 0.595

T. Covariance = 0.537
TC = 0.736
TB = 0.091
TP Covariance = 0.008
TPhi = 0.202
TPB = 0.468

"Rodriguez Madrid Train Bombing"-Data (Hayes, 2006), Dichotomized, cf. footnote 3.

(f)

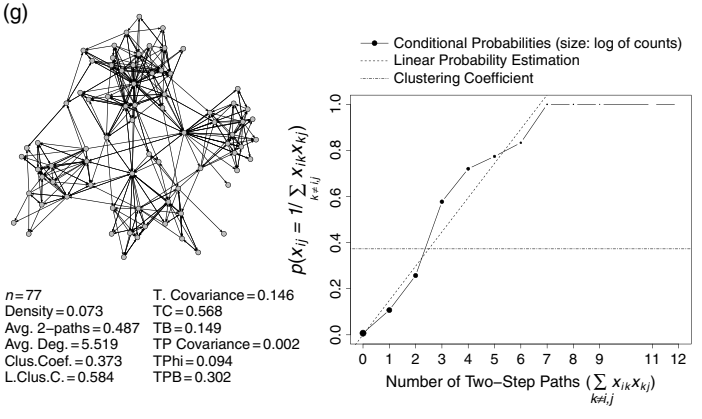


$n = 1{,}589$
Density = 0.001
Avg. 2–paths = 0.002
Avg. Deg. = 1.726
Clus.Coef. = 0.707
L.Clus.C. = 0.674

T. Covariance = 0.001
TC = 0.505
TB = 0.186
TP Covariance = 0.0000
TPhi = 0.025
TPB = 0.706

"Coauthors' Network"-Data (Newman, 2006), Dichotomized, cf. footnote 3.

(g)

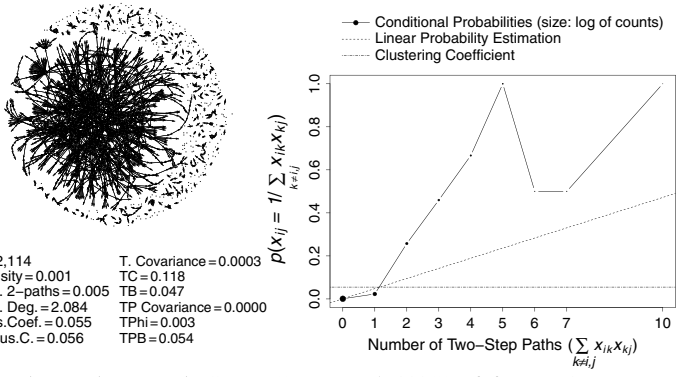


$n = 77$
Density = 0.073
Avg. 2–paths = 0.487
Avg. Deg. = 5.519
Clus.Coef. = 0.373
L.Clus.C. = 0.584

T. Covariance = 0.146
TC = 0.568
TB = 0.149
TP Covariance = 0.002
TPhi = 0.094
TPB = 0.302

"R&D Team in Manufacturing Company"-Data (Cross & Parker, 2004), Dichotomized $x_{ij} = 1 \forall\ y_{ij} > 4$, cf. footnote 2.

(h)



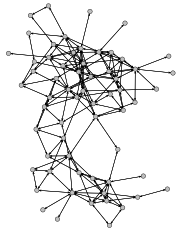
$n = 2{,}114$
Density = 0.001
Avg. 2–paths = 0.005
Avg. Deg. = 2.084
Clus.Coef. = 0.055
L.Clus.C. = 0.056

T. Covariance = 0.0003
TC = 0.118
TB = 0.047
TP Covariance = 0.0000
TPhi = 0.003
TPB = 0.054

"Protein-Protein Interaction"-Data ((Jeong *et al.*, 2001)), cf. footnote 3.

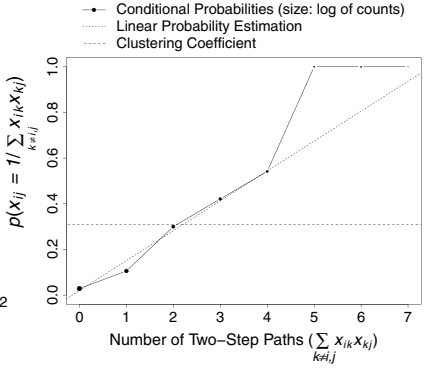
Figure 5. B

(i)



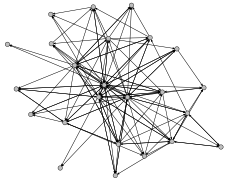
$n = 62$
Density = 0.084
Avg. 2–paths = 0.488
Avg. Deg. = 5.129
Clus.Coef. = 0.309
L.Clus.C. = 0.347

T. Covariance = 0.110
TC = 0.432
TB = 0.131
TP Covariance = 0.002
TPhi = 0.073
TPB = 0.227



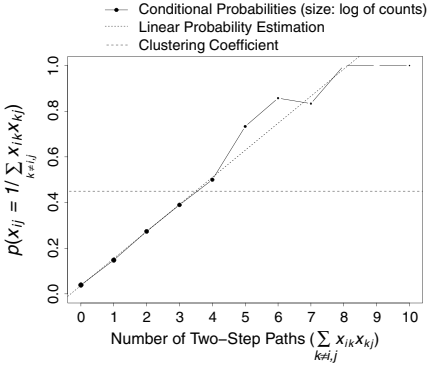
"Dolpins Frequently Associating"-Data (Lusseau, 2003), cf. footnote 3.

(j)



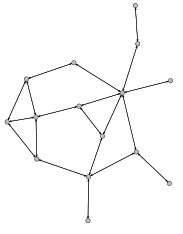
$n = 24$
Density = 0.245
Avg. 2–paths = 1.755
Avg. Deg. = 5.625
Clus.Coef. = 0.450
L.Clus.C. = 0.676

T. Covariance = 0.361
TC = 0.481
TB = 0.118
TP Covariance = 0.016
TPhi = 0.141
TPB = 0.223



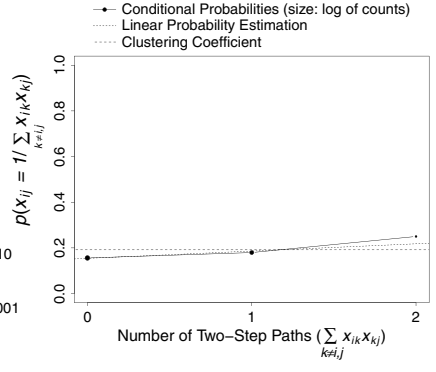
"24 Country Minerals and Fuels Trade"-Data (Smith & White, 1992; Wasserman & Faust, 1994), cf. footnote 3.

(k)



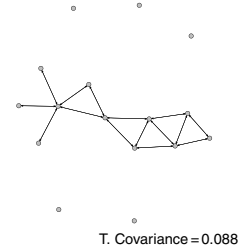
$n = 16$
Density = 0.167
Avg. 2–paths = 0.392
Avg. Deg. = 2.500
Clus.Coef. = 0.191
L.Clus.C. = 0.178

T. Covariance = 0.010
TC = 0.047
TB = 0.032
TP Covariance = 0.001
TPhi = 0.011
TPB = 0.026



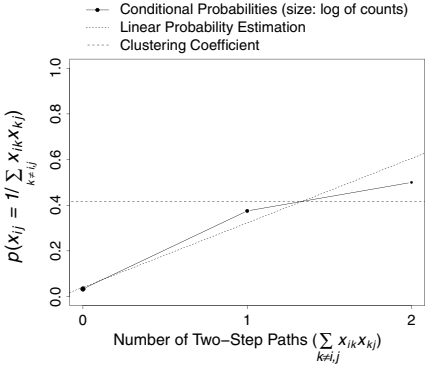
"Florentine Marriage"-Data (Padgett & Ansell, 1993), cf. footnote 3.

(l)



$n = 16$
Density = 0.125
Avg. 2–paths = 0.300
Avg. Deg. = 1.875
Clus.Coef. = 0.417
L.Clus.C. = 0.245

T. Covariance = 0.088
TC = 0.475
TB = 0.282
TP Covariance = 0.006
TPhi = 0.131
TPB = 0.298



"Florentine Business"-Data (Padgett & Ansell, 1993), cf. footnote 3.

Figure 5. C

The relevant summary statistics here are n, the number of nodes in the network, density, and the mean number of two-step paths between the n ($n-1$) node pairs, average degree (Avg. Deg.), Clustering Coefficient, C (Clus. Coef.), Local Clustering Coefficient, LC (L. Clus. C.), transitivity covariance (T. Covariance), transitivity correlation, TC, transitivity beta, TB, transitivity Phi covariance (TP Covariance), transitivity Phi, $TPhi$, and transitivity Phi beta, TPB.

The example networks are from a variety of fields, and differ in size ($n = 16$ to $n = 2,114$) and structure ($d = 0.001$ to $d = 0.908$). In most examples, there is a positive TB, implying that in all these networks there is network transitivity. The exception is the formal organizational "reports to" relationship among high-tech managers (Figure 5(d)). The negative value for network transitivity here is induced by the design of formal organizational networks, which are usually set up as trees. Although in some examples, a low clustering coefficient ($C < 0.2$), such as for *C. elegans* (Figure 5(c)), protein interactions (Figure 5(h)), and Mediaeval Florentine Family Weddings (Figure 5(k)), could be interpreted as no tendency toward transitivity in the network, this would be a mistake. The positive regression coefficient TB indicates an elevated propensity toward transitive triples occurring on average throughout these networks as the number of two-paths between pairs increases.

It must be emphasized again that no inferential claims can be made about the statistical significance of these descriptive statistics. This would require further nontrivial assumptions about underlying digraph distributions. What could be done is to make a case by case comparison. For example, in the Florentine families data (Figure 5(i) and (k)), it would be a valid statement to say that network transitivity is higher in the observed business network compared to the marriage network.

Further, this is not restricted to comparing networks on the same group of nodes, but holds for comparison between any type of network if we would compare TC. For example, comparing the Southern women club with friendships in a law firm, the latter has (slightly) lower TC (0.443 vs. 0.445), and hence lower network transitivity. In this case, the clustering coefficient would have led to the same conclusion. But, this is not always so.

Comparing the intercountry trade of minerals and fuel data (Figure 5(j)) with frequent, and, very frequent information exchange (Figure 5(g)) shows very similar diagrams. However, the clustering coefficients ($C = 0.417$ and 0.344, respectively) would suggest a different conclusion than when comparison is done on Transitivity Correlation ($TC = 0.481$ and 0.584, respectively). This is due to differences in density of the two networks. The conditional probability of a transitive triple is higher in observed mineral and fuel trade network compared to information exchange, due to a higher density. The increased propensity toward transitive triples is *more increased* in the information exchange network, and in this sense it shows more network transitivity.

Further, a remarkable finding that illustrates the value of these plots is that in three cases (Figure 5(c), (e), and (h)) with positive TB, the probability of a tie does not show a monotonic increase with increasing two-path counts. Most clearly, this is shown in the neural network of *C. elegans*, where beyond 9 two-paths between 2 nodes, the probability for a tie strongly diminishes (except at 14 two-paths). Reasons for this could be myriad, but it is important to consider that it could be indicative of missing, incomplete, or biased data. The example in Figure 5(h) has been shown to be an incomplete data set, which limited conclusions of the study on this data set (see for critiques Coulomb et al., 2005; Han et al., 2005; Stumpf et al., 2005). Or, due to ill-defined relationships, for example, interactions could traverse through different media not considered (e.g., complementary use of e-mail and phone), so that not all relevant interactions may have been observed. Similarly, the network in Figure 5(e) displays a drop in tie probability at 3 and 6 two-paths, while a sharp increase occurs at 7. As this data set is a covert network constructed from secondary sources, it could be indicative of a missing source, or a bias because some sources are irrelevant or receive too much emphasis. At least, non-monotonicity in the plots deserves a further theoretical explanation when no data-related reasons can be found.

5. Discussion

This paper proposed new measures for transitivity based on covariances and correlations between ties and two-paths, and described some of their numerical properties. These new measures all are defined as correlations, covariances, or regression coefficients in empirical distributions defined by the network, expressing comparisons between the probability of a tie between two nodes depending on the occurrence of two-paths connecting the nodes, such as in Equation (2).

5.1 Statistical inference

The measures developed in this paper are proposed as descriptives, and not primarily for use in statistical inference (for an overview of issues in statistical modeling for social network analysis, see Snijders, 2011). Statistical inference about transitivity in networks can be directed at testing the null hypothesis of no transitivity, for which the proposed measures can be used as test statistics.

The latter topic is treated by Karlberg (1999). This author defines two transitivity indices as potential test statistics, and uses as a null distribution the $U \mid (OD, ID)$ specification, that is, the uniform distribution conditional on given in- and out-degree vectors. His first test statistic is Equation (3). His second test statistic is an average of local transitivity indices, where the local transitivity is defined as the density of the out-neighborhood of the node, divided by the maximal density given the indegree, outdegree, and number of mutual ties of the node. We suggest that our proposed statistic *TC* could also be a suitable statistic for testing this null. A suitable null distribution could be $U \mid (OD, ID, M)$, the uniform distribution conditional on given in- and out-degree vectors and a given number M of reciprocated ties. Although generating random networks from these distributions is not discussed here, it should be noted that generating samples from the $U \mid (OD, ID)$ as well as from the $U \mid (OD, ID, M)$ distribution faces serious combinatorial restrictions. A computer program that can simulate samples from these two distributions is ZO (Snijders, 2017), based on Snijders (1991), and obtainable from `http://www.stats.ox.ac.uk/~snijders/socnet.htm`. More recently another method for doing this was proposed by Tao (2016). Further literature about the generation of networks with given in- and outdegrees is Rao et al. (1996), Roberts (2000), Verhelst (2008), and Chatterjee et al. (2011).

5.2 Absence and presence of transitivity

One of our conclusions is that condition (16), depending on outdegrees, indegrees, and number of mutual ties, expresses absence of transitivity. This echoes and refines Feld and Elmore's (1982) observation, extended later by Faust (2007), that interpretations of the number of transitive triples in a network should take into account the degree distributions. It is also related to the statement, made by Snijders et al. (2006) and Lusher et al. (2012, p. 70), that the number of independent two-paths (also called dyadwise shared partners) should be included in specifications of Exponential Random Graph Models as a "prerequisite," or lower-order configuration, for testing the transitive closure expressed by k-triangles (also called edgewise shared partners).

In the observed network examples in Section 4.4, we have mainly found positive values for Transitivity Covariance. This unambiguously shows that there is an increased propensity toward transitive triples in these networks, in line with the predominance of transitive triples found in a much larger set of networks already by Davis (1970). However, in some cases the diagrams that depict how the probability of tie depends on the number of two-step connections between the nodes also show that the observed probabilities for ties may become highly variable for high values of the number of two-paths. This in itself is thought-provoking theoretically, and might inspire other measures that express deviations from a linear relation. However, other explanations are also possible, such as randomness, lack of data quality, or existence of covert ties.

5.3 Extensions

Next to transitivity we may consider balance (Heider, 1958). When balance is treated for graphs or digraphs without considering edge signs, it is usual to treat absent edges as negative ties. Instead of Transitivity Covariance, the "Balance Covariance" would then be based on the association between x_{ij} and $x_{ik}\,x_{kj} + x^c_{ik}\,x^c_{kj}$, where x^c is the complement of digraph x, with tie variables $x^c_{ij} = 1 - x_{ij}$. As the values of $x_{ik}\,x_{kj} + x^c_{ik}\,x^c_{kj}$ still are in $\{0, 1\}$, the analyses will remain similar. The measure can straightforwardly be further adjusted to accommodate other statements about triads.

Further refinements could be made regarding, for example, the implicit assumptions about homogeneity of nodes. In case nodes are explicitly organized in groups, a distinction between different subsets of nodes, or different blocks of ties, may refine conclusions about increased, or decreased, levels of a tendency toward transitivity. Adjusted covariance-based measures could be derived in this way, controlling for grouping of nodes.

Further developments could also be made for networks with valued ties. A generalized form of network transitivity for valued ties was proposed by Opsahl & Panzarasa (2009). It is still unknown in which way this would lead to different conclusions and interpretations than those presented here.

6. Conclusion

We defined two new measures for transitivity: Transitivity Phi *TPhi*, defined as the observed correlation, in a randomly drawn triple, between the tie variable between two nodes and the two-path connection between them; and the Transitivity Correlation *TC*, defined as the observed correlation, for a randomly drawn pair of nodes, between the tie variable and the number of two-paths between the two nodes. The foremost advantage of these measures is that they offer a quantitative expression for the "*increased* propensity" of transitive triples which is the definition of transitivity as formulated, for example, by Newman et al. (2001). By contrast, the clustering coefficient *C*, one of the basic measures for transitivity, reflects the observed conditional probability of a tie, given a two-path, not a comparative quantity. Under the Erdös–Renyi model, the clustering coefficient can have any expected value in $(0, 1)$ depending on the density; under this model, the expected value of *TPhi* and *TC* is 0. Because of their comparative nature, these correlation measures allow for comparison between networks, even networks of unequal size or density, and from different contexts.

The two measures are both based on considering the tie variable for a random pair (i, j) of nodes; the difference is that *TPhi* considers one randomly selected third node, whereas *TC* considers all other nodes as potential intermediates. Both are functions of the ego-networks of all nodes in the digraph, where the ego-network is defined as the digraph induced by the node and all nodes in its direct out-neighborhood. Clearly, *TC* takes into account much more of the structure of the ego-networks than *TPhi*, specifically, the dependence between the different two-paths connecting any two nodes.

The results found in the comparison of measures for windmill graphs led to the conclusion that the difference between these two measures can imply large differences in conclusions about transitivity. For windmills with many wings, the consideration of the two-path dependence by *TC* leads to a value tending to 1, contrasting with the value for *TPhi* tending to 0. We interpret windmill graphs as being highly transitive, and find this a strong argument in favor of *TC* over *TPhi*.

Correlations between binary variables are known to have a restricted range. Only for graphs that are unions of disconnected complete subgraphs of equal sizes, both *TC* and *TPhi* assume the maximum of 1. This shows that there may be room for developing other measures for transitivity that assume their maximum value for all totally transitive graphs, without the restriction of equal-size components.

A finding that we believe to be new is that the condition that *TC* is zero is equivalent to a condition on the covariance between in- and outdegrees, the number of mutual ties, the density, and the number of nodes. This leads to interest in the uniform distribution for digraphs conditional on these four quantities. This distribution presumably is very difficult to handle; the distribution of digraphs, for a given number of nodes, conditional on the vectors of in- and outdegrees and the number of mutual ties may be presumed to be easier to handle, although this distribution already poses huge problems (Tao, 2016; Snijders, 2017).

Acknowledgments. Special thanks go to the participants of the seminar series at the Centre for Business Network Analyses (University of Greenwich)and at the CASOS seminar (Carnegie Mellon University), and to the editor and two anonymous reviewers for comments and suggestions.

Conflict of interest. Authors have nothing to disclose.

Notes

1 This measure has been implemented in function *gtrans* in the R-package 'sna' (Butts, 2016, p. 112).
2 Data via Opsahl (2017).
3 Data via Freeman (2017).

References

Butts, C. T. (2016). *Package "sna"*. Available at: https://cran.r-project.org/web/packages/sna/sna.pdf
Chatterjee, S., Diaconis, P., & Sly, A. (2011). Random graphs with a given degree sequence. *The Annals of Applied Probability*, 1400–1435.
Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.
Coulomb, S., Bauer, M., Bernard, D., & Marsolier-Kergoat, M.-C. (2005). Gene essentiality and the topology of protein interaction networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1573), 1721–1725.
Cross, R. L., & Parker, A. (2004). *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*. Boston, MA: Harvard Business Press.
Davis, A., Gardner, B. B., Gardner, M. R., & Warner, W. L. (1941). *Deep South: A Sociological Anthropological Study of Caste and Class*. Chicago: University of Chicago Press.
Davis, J. A. (1970). Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices. *American Sociological Review*, *35*, 843–852.
Erdös, P., & Renyi, A. (1959). On random graphs. *Publ Math Debrecen*, *6*(290–297), 290–297.
Falk, R., & Well, A. D. (1997). Many faces of the correlation coefficient. *Journal of Statistics Education*, *5*(3), 1–18.
Faust, K. (2007). Very local structure in social networks. *Sociological methodology*, *5*(2), 148–256.
Faust, K. (2010). A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks*, *32*(3), 221–233.
Feld, S. L., & Elmore, R. (1982). Patterns of sociometric choices: Transitivity reconsidered. *Social Psychology Quarterly*, *45*(2), 77–85.
Frank, O. (1980). Sampling and inference in a population graph. *International Statistical Review Revue Internationale de Statistique*, *48*(1), 33.
Frank, O. (1988). Random sampling and social networks: A survey of various approaches. *Mathematiques Informatique et Sciences Humaines*, *26*, 19–33.
Freeman, L. C. (2017). Datasets. moreno.ss.uci.edu/data.html, Accessed February 1, 2017.
Goodreau, S. M., Kitts, J. A., & Morris, M. (2009). Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, *46*(1), 103–125.
Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380.
Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., & Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, *23*(7), 839–844.
Harary, F., & Kommel, H. J. (1979). Matrix measures for transitivity and balance. *Journal of Mathematical Sociology*, *6*(2), 199–210.
Hayes, B. (2006). Connecting the dots can the tools of graph theory and social-network studies unravel the next big plot? *American Scientist*, *94*(5), 400–404.
Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: John Wiley & Sons.
Holland, P. W., & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology*, *76*(3), 492–513.

Holland, P. W., & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Small Group Research*, *2*(2), 107–124.
Holland, P. W., & Leinhardt, S. (1972). Holland and Leinhardt reply: Some evidence on the transitivity of positive interpersonal sentiment. *American Journal of Sociology*, *77*(6), 1205–1209.
Holland, P. W, & Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology*, *7*, 1–45.
Jackson, M. O. (2008). *Social and Economic Networks*. Vol. 3. Princeton: Princeton University Press.
Jeong, H., Mason, S. P., Barabási, A.-L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411*(6833), 41–42.
Karlberg, M. (1999). Testing transitivity in digraphs. *Sociological Methodology*, *29*(1), 225–251.
Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, *9*, 109–134.
Lazega, E. (2001). *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press on Demand.
Lusher, D., Koskinen, J., & Robins, G. (2012). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. New York: Cambridge University Press.
Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(Suppl 2), S186–S188.
Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, *64*(2), 19.
Newman, M. E. J. (2003). The structure and function of complex networks. *Siam Review*, *45*(2), 167–256.
Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, *74*(3), 036104.
Opsahl, T. (2017). Datasets. https://toreopsahl.com/datasets/, Accessed February 1, 2017.
Opsahl, T., & Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks*, *31*(2), 155–163.
Padgett, J. F., & Ansell, C. K. (1993). Robust action and the rise of the medici, 1400-1434. *American Journal of Sociology*, *98*(6), 1259–1319.
Rao, A. R., Jana, R., & Bandyopadhyay, S. (1996). A markov chain monte carlo method for generating random (0, 1)-matrices with given marginals. *Sankhyā: The Indian Journal of Statistics, Series A*, 225–242.
Roberts, J. M. (2000). Simple methods for simulating sociomatrices with given marginal totals. *Social Networks*, *22*(3), 273–283.
Schelling, T. C. (1978). *Micromotives and Macrobehavior*. New York [etc.]: Norton.
Smith, D. A., & White, D. R. (1992). Structure and dynamics of the global economy: Network analysis of international trade 1965–1980. *Social Forces*, 857–893.
Snijders, T. A. B. (1991). Enumeration and simulation methods for 0–1 matrices with given marginals. *Psychometrika*, *56*(3), 397–417.
Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, *36*, 99–153.
Snijders, T. A. B. (2011). Statistical models for social networks. *Annual Review of Sociology*, *37*, 131–153.
Snijders, T. A. B. (2017). *Manual for ZO version 2.3*. Department of Sociology, University of Groningen, Groningen. http://www.stats.ox.ac.uk/~snijders/socnet.htm.
Stumpf, M. P. H., Wiuf, C., & May, R. M. (2005). Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(12), 4221–4224.
Tao, T. (2016). An improved MCMC algorithm for generating random graphs from constrained distributions. *Network Science*, *4*, 117–139.
Verhelst, N. D. (2008). An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika*, *73*(4), 705–728.
Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
Wasserman, S. S. (1975). Random directed graph distributions in the triad census in social networks. *Journal of Mathematical Sociology*, *5*(1), 61–86.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.

Appendix A: Relation *TPhi* and *TC*

The covariance between directed ties and the number of two-paths is

$$\operatorname{cov}\Big(x_{ij}, \sum_{h\neq i,j} x_{ih}x_{hj}\Big) = \frac{1}{n\,(n-1)} \sum_i \sum_{j\neq i} x_{ij} \Big(\sum_{h\neq i,j} x_{ih}\, x_{hj}\Big) - \bar{x}\ \overline{xx} \tag{A1}$$

The difference between *TC* and *TPhi* is in scaling. Consider that covariance in Equation (A1) is a weighted measure of the numerator in Equation (9):

$$\operatorname{cov}\Big(x_{ij}, \sum_{h\neq i,j} x_{ih}x_{hj}\Big) = (n-2)\operatorname{cov}(x_{ij},\ x_{ik}x_{kj}) \tag{A2}$$

Further, the denominator in Equation (12) differs from that of *TPhi* only in *s.e.*($\sum_h x_{ih}x_{hj}$) [see Equation (10)]. The variance of the number of two-paths between any ordered pair can be rewritten as

$$\operatorname{var}\Big(\sum_h x_{ih}x_{hj}\Big) = (n-2)\left[\operatorname{var}(x_{ik}x_{kj}) + (n-3)\operatorname{cov}(x_{ik}x_{kj}, x_{i\ell}x_{\ell j})\right] \tag{A3}$$

Under conditions where $\operatorname{cov}(x_{ik}x_{kj}, x_{i\ell}x_{\ell j}) = \operatorname{var}(x_{ik}x_{kj})$, *TC* reduces to *TPhi* as

$$\operatorname{var}\Big(\sum_h x_{ih}x_{hj}\Big) = (n-2)^2\operatorname{var}(x_{ik}x_{kj}) \tag{A4}$$

But, more generally, we can state

$$TC = \alpha \times TPhi \tag{A5}$$

where α is

$$\begin{aligned}\alpha &= \frac{TC}{TPhi}\\ &= \frac{(n-2)\sqrt{\operatorname{var}(x_{ik}x_{kj})}}{\sqrt{((n-2)[\operatorname{var}(x_{ik}x_{kj}) + (n-3)\operatorname{cov}(x_{ik}x_{kj}, x_{i\ell}x_{\ell j})])}}\end{aligned} \tag{A6}$$

for $\ell \neq k$. Note that (A6) can be rewritten as

$$\alpha = \frac{(n-2)}{\sqrt{((n-2)\,[1+(n-3)\rho])}} \tag{A7}$$

where

$$\rho = \frac{\operatorname{cov}(x_{ik}x_{kj},\ x_{i\ell}x_{\ell j})}{\operatorname{var}(x_{ik}x_{kj})} \tag{A8}$$

is the autocorrelation between two-paths in a digraph. Now ρ is a correlation so that $\rho \leq 1$; further, rewriting Equation (A3),

$$\begin{aligned}0 \leq \operatorname{var}\Big(\sum_{k\neq i,j} x_{ik}x_{kj}\Big) &= (n-2)\operatorname{var}(x_{ik}x_{kj}) + (n-2)(n-3)\operatorname{cov}(x_{ik}x_{kj},\ x_{i\ell}x_{\ell j})\\ &= (n-2)\big(1+(n-3)\,\rho\big)\operatorname{var}(x_{ik}x_{kj})\end{aligned} \tag{A9}$$

which implies that $\rho \geq -1/(n-3)$. With Equation (A7) this implies that $\alpha \geq 1$ except for $TPhi = 0$, where $\rho = -1/(n-3)$ and *TPhi* is undefined. It can be concluded that $TC \geq TPhi$. The distinction between *TC* and *TPhi* is about size not direction. Although the relation between α and ρ is nonlinear, α is monotonically decreasing as ρ increases. The ratio of the two transitivity correlations is a function of n and the two-path autocorrelation in the digraph. The autocorrelation between two-paths is itself a Phi coefficient, expressing the difference in conditional probabilities of a two-path via a node k given a two-path via another node ℓ exists and of a two-path via k given that no other two-path exists. As such, it can be interpreted as a measure of network centrality, where a smaller ρ indicates an elevated uniqueness of nodes as intermediate in two-paths.

Appendix B: When is transitivity covariance equal to zero

We derive an condition equivalent to the property that the Transitivity Correlation *TC*, or equivalently the Transitivity Covariance, is zero.

The Transitivity Covariance is defined as the covariance, for a randomly chosen pair (i, j), between the direct tie x_{ij} and the number $\sum_{h\neq i,j} x_{ih}x_{hj}$ of directed two-paths between these nodes as defined in Equation (A2). Network density in (di)graphs is the mean tie-indicator variable

$$d = \bar{x} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} x_{ij} \tag{A10}$$

Over all ordered pairs (i, j), the mean number of two-paths is

$$\overline{xx} = TSP = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{h=1 \\ h \neq i,j}}^{n} x_{ih}\, x_{hj} \tag{A11}$$

and the mean number of transitive triples is given by

$$TT = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{h=1 \\ h \neq i,j}}^{n} x_{ij}\, x_{ih}\, x_{hj} \tag{A12}$$

Let

$$M = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} x_{ij}\, x_{ji} \tag{A13}$$

be the sum of reciprocal or mutual ties (note that reciprocal ties are two-paths that do not contribute to a transitive triple). The mean number of two-paths in Equation (A11) can be rewritten as

$$TSP = \frac{1}{n\,(n-1)} (OD \cdot ID - M) \tag{A14}$$

where ID and OD are the vectors of in- and outdegrees, and $OD \cdot ID$ is their inner product.

Substitution in Equation (A2) gives

$$\operatorname{cov}(x_{ij}, \sum_{h \neq i,j} x_{ih}\, x_{hj}) = (n-2)\left(TT - \frac{(OD \cdot ID - M)\, d}{n(n-1)} \right) \tag{A15}$$

The inner product of two vectors can be expressed in terms of covariance, in this case the covariance between in- and outdegrees for the probability distribution that a node is randomly chosen. This gives

$$\operatorname{cov}(OD, ID) = \frac{1}{n} (OD \cdot ID) - d^2 \tag{A16}$$

Substitution gives

$$\operatorname{cov}(x_{ij}, \sum_{h \neq i,j}^{n} x_{ih}\, x_{hj}) = (n-2)\left(TT - \frac{\left(n\ \operatorname{cov}(OD, ID) + n\, d^2 - M\right) d}{n(n-1)} \right) \tag{A17}$$

which simplifies to Equation (16) under the condition of no network (in)transitivity.

Cite this article: Dekker D., Krackhardt D., and Snijders T. A. B. (2019). Transitivity correlation: A descriptive measure of network transitivity. *Network Science* **7**, 353–375. https://doi.org/10.1017/nws.2019.32