THE DIFFERENCE BETWEEN THE TRANSITIVITY RATIO AND THE CLUSTERING COEFFICIENT

KARL ROHE

Friends of friends are often friends. This process creates triangles (or three cliques) in social networks. The transitivity ratio and the clustering coefficient are the two most popular statistics that measure the number of triangles in a network. Both measures can be expressed as probabilities. Pick a random two star from the graph. Is the random triplet a completely connected triangle? The transitivity ratio and the clustering coefficient differ in how they sample the random two star. As a result, they can produce wildly different results. This paper constructs two sequences of graphs. In the fist sequence, the transitivity ratio converges to zero, while the clustering coefficient converges to one. In the second sequence, the limits are reversed.

0.1. **Notation.**

0.2. **Preliminaries.** A graphs G is represented by a vertex set and an edge set, $G = (V, E)$, where $V = \{1, \ldots, n\}$ contains the actors and

$$E = \{(i, j) : \text{ there is an edge from } i \text{ to } j\}.$$

The standard expressions for the transitivity ratio is

$$trans(G) = \frac{3 \times \text{number of closed triplets in } G}{\text{number of connected triples of vertices in } G}. \tag{1}$$

The three in the numerator accounts for node relabeling; the triplet has 3! different labelings and the two star has 2. Define $N(i) = \{j|(i, j) \in E\}$ as the neighborhood of node i and $deg(i) = |N(i)|$ as the degree of node i. Then, the local clustering coefficient $C(i)$ and the global clustering coefficient $C(G)$ as defined in [**?**] are

$$C(i) = \frac{|\{(j, k) \in E|j, k \in N(i)\}|}{\binom{deg(i)}{2}} \text{ and } C(G) = \frac{1}{n}\sum_i C(i). \tag{2}$$

The following proposition emphasizes the similarity between these two summary statistics.

Proposition 1. *For any fixed graph G, let I, J, K be random variables that are sampled uniformly from V without replacement, then*

$$\begin{aligned} trans(G) &= P\left((J, K) \in E|(J, I), (K, I) \in E\right) \\ C(i) &= P((J, K) \in E|(J, i), (K, i) \in E) \\ C(G) &= \mathbb{E}(C(I)). \end{aligned} \tag{3}$$

A short proof is contained in the appendix.

Importantly, the above probabilities and expectations come from drawing the random variables I, J, K uniformly from V, without replacement. For any fixed graph, these probabilities can be computed explicitly using the formulas in Equations (1) and (2). These nonstandard formulations of for $trans(G)$ and $C(G)$ are meant to emphasize the similarities between the two summary statistics. The key difference is that $C(G)$ stratifies by internal node I, where the internal node is drawn uniformly from the set of all nodes.

Proposition 2. *For any fixed graph G, let $I' \in V$ be a random variable with distribution defined by*

$$P(I' = i) \propto \binom{deg(i)}{2} \quad \textit{for all } i \in V, \tag{4}$$

then

$$trans(G) = \mathbb{E}(C(I')). \tag{5}$$

A short proof is contained in the appendix.

Equation (3) shows that $C(G)$ is the expectation $C(I)$. Proposition 2 shows that the transitivity ratio has a similar formulation; the key difference is that I' comes from a distribution that is much more likely to choose a high degree node.

Section 1 constructs a sequence of graphs B_n with $trans(B_n) \to 0$ and $C(B_n) \to 1$. Section 2 constructs a sequence of graphs F_n with $trans(B_n) \to 1$ and $C(B_n) \to 0$

1. Bike wheel graph

The bike spoke graph contains two hub nodes and n rim nodes. The hub nodes are connected (by the axel, in the bike analogy). Moreover, the hub nodes are connected to every rim node (by a spoke, in the bike analogy). No rim nodes are connected.

Definition 1. *Let B_n denote the bike wheel graph on nodes $V = \{a, b, 1, \ldots, n\}$. The edge set of B_n contains edge (i, j) if $\{i, j\} \cap \{a, b\} \neq \phi$.*

Proposition 3.

$$\lim_{n\to\infty} trans(B_n) = 0 \quad \textit{and} \quad \lim_{n\to\infty} C(B_n) = 1.$$

Proof. Note that $C(a) = C(b) = 2/(n+1)$ and $C(i) = 1$ for $i \in \{1, \ldots, n\}$. Using Proposition 2,

$$\begin{aligned} trans(B_n) &= \frac{\binom{deg(a)}{2}C(a) + \binom{deg(b)}{2}C(b) + n\binom{deg(1)}{2}C(1)}{\binom{deg(a)}{2} + \binom{deg(b)}{2} + n\binom{deg(1)}{2}} \\ &= \frac{3n}{n^2 + 2n} \to 0. \end{aligned}$$

For the clustering coefficient, note that $C(G) \leq 1$ for any graph G. Moreover,

$$C(B_n) = \mathbb{E}(C(I)) > \mathbb{E}\big(C(I)\ \mathbf{1}[I \in \{1, \ldots, n\}]\big) = P(I \in \{1, \ldots, n\}) \to 1.$$

□

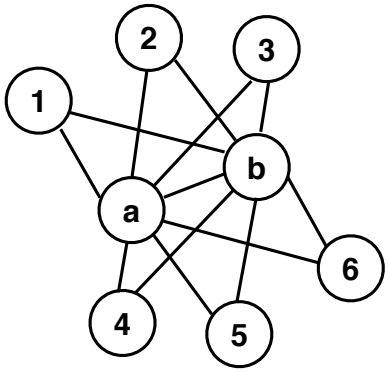

FIGURE 1. This is an image of B_6, the bike wheel graph on 8 nodes. In the bike wheel graph on $n+2$ nodes, nodes a and b are central hub-nodes that are connected to each other and every other node. Nodes $1, \ldots, n$ are rim-nodes that connect to both nodes 1 and 2. There are no connections between rim-nodes.

2. Clique and chain graph

The clique and chain graph on n nodes is the union of two disconnected graphs. First, a clique of $O(n^{2/3})$ nodes. Second, a chain with the remaining nodes. Call this F_n.

The clique and chain graph resembles a single cellular organism with a flagellum (tail) protruding from one side. The key difference between F_n and B_n is that the low degree nodes in F_n (i.e. the nodes in the line graph) do not connect to the high degree nodes (i.e. the nodes in the clique). As a result, the limits of $trans(F_n)$ and $C(F_n)$ are exactly the opposite.

Proposition 4.

$$\lim_{n\to\infty} trans(F_n) = 1 \quad \textit{and} \quad \lim_{n\to\infty} C(F_n) = 0.$$

Proof. The number of two stars in the clique are of the order

$$n^{2/3} \times \binom{n^{2/3}-1}{2} \sim n^{2/3} \times n^{4/3} = n^2.$$

In the chain graph, the number of two stars is order $n \times 1$. So, with overwhelming probability, the sampling technique in $trans$ selects a two star from the clique. These three nodes form a triangle because they all belong to the clique. So, $\lim_{n\to\infty} trans(F_n) = 1$. On the other hand, the clique contains a vanishing fraction of the nodes, $n^{2/3}/n = n^{-1/3}$. So, with overwhelming probability, the sampling technique for C will select a two star from the chain. These three nodes will not form a triangle because they all belong to the chain. □

3. Discussion

Appendix A. Short proofs

Here is a proof of Proposition 1.

Proof.

$$\begin{aligned} P\left((J,K)\in E|(J,I),(K,I)\in E\right) &= \frac{P\left((J,K),(J,I),(K,I)\in E\right)}{P\left((J,I),(K,I)\in E\right)} \\ &= \frac{6\times\text{number of closed triplets in } G/\binom{n}{3}}{2\times\text{number of connected triples of vertices in } G/\binom{n}{3}} \\ &= trans(G) \end{aligned}$$

$$\begin{aligned} P((J,K)\in E|(J,i),(K,i)\in E) &= \frac{P\left((J,K),(J,i),(K,i)\in E\right)}{P\left((J,i),(K,i)\in E\right)} \\ &= \frac{\left|\{(j,k)\in E|j,k\in N(i)\}\right|/\binom{(n-1)}{2}}{\binom{deg(i)}{2}/\binom{(n-1)}{2}} \\ &= C(i) \end{aligned}$$

Equation (3) follows from the definition of $C(G)$. □

Here is a proof of Proposition 2.

Proof.

$$\begin{aligned} \mathbb{E}(C(I')) &= \frac{1}{\sum_i\binom{deg(i)}{2}}\sum_i\binom{deg(i)}{2}C(i) \\ &= \frac{1}{\sum_i\binom{deg(i)}{2}}\sum_i\left|\{(j,k)\in E|j,k\in N(i)\}\right| \\ &= \frac{3\times\text{number of closed triplets in } G}{\text{number of connected triples of vertices in } G}. \end{aligned}$$

□