# Weighted Higher-order Clustering

Alex Libman (asl237), Peter Mocarski (pmm248), Neil Sun (ns664)

## 1 Introduction

Clustering is an important component of understanding network structure. We develop a metric for analyzing the clustering of weighted graphs. We base our work on two closely related papers in the area of network clustering. Both papers seek to generalize the idea of the clustering coefficient, which is the proportion of length-2 wedges that close a triangle in the graph. The first paper, "Higher-order clustering in networks" (Yin, Benson, and Leskovec 2018) [5], generalizes the clustering coefficient in graphs to higher-order clustering coefficients. The second paper, "Clustering in weighted networks" (Opsahl & Panzarasa 2009) [4] generalizes the clustering coefficient in graphs to weighted networks. Each paper takes a different approach to generalizing the clustering coefficient, and our goal was to combine the two ideas to define weighted higher-order clustering in graphs, to see if this type of analysis could reveal new insights in network and clustering analysis.

## 2 Background

### 2.1 Higher-order Clustering in Networks

Yin et al. (2018) explores the tendency for edges to cluster in a graph, specified by the graph's clustering coefficient. Higher-order clustering coefficients can be viewed as clique expansion probabilities. More explicitly, "the $l$th-order clustering coefficient $C_l$ measures the probability that an $l$-clique and an adjacent edge, i.e., an $l$-wedge, is closed, meaning that the $l-1$ possible edges between the $l$-clique and the outside node in the adjacent edge exist to form an $(l+1)$-clique." With this in mind, the paper writes the global $l$th order clustering coefficient $C_l$ as follows

$$C_l = \frac{(l^2+l)|K_{l+1}|}{W_l}$$

where $K_{l+1}$ is the set of $(l+1)$-cliques, and $W_l$ is the set of $l$-wedges. Higher-order local clustering coefficients are also defined as follows

$$C_l(u) = \frac{l|K_{l+1}(u)|}{|W_l(u)|}$$

where $K_{l+1}(u)$ is the set of $(l+1)$-cliques containing node $u$ and $W_l(u)$ is the set of $l$-wedges with center $u$.

With this in mind, the paper finds that clustering coefficients can be bound by extremal bounds, and in the case of the $G_{n,p}$ model, the expected value of the clustering coefficient $C_l$, in both the local and global case, approach $p^{l-1}$ for large graphs.

Finally, the higher-order clustering coefficients of various real-world graphs are calculated and analyzed, and when compared against null models, provide new and useful insights. For example, high-order clustering in the friendship network of an early variant of Facebook shows consistent clustering beyond just triadic closure, implying higher-order structure.

### 2.2 Clustering in Weighted Networks

Opsahl et al. (2009) seeks to define a generalization of the global clustering coefficient for weighted graphs. They define a triplet as three nodes connected by either two or three edges. They propose four different metrics to define the value of a triplet: the arithmetic mean of the weights, the

geometric mean of the weights, the maximum value of the weights, and the minimum value of the weights. From here, they define the weighted clustering coefficient as

$$C_\omega = \frac{\text{total value of closed triplets}}{\text{total value of triplets}}$$

Some advantages of this formulation of the weighted clustering coefficient include that it generalizes to the unweighted case, and it produces values between 0 and 1. The paper tests this definition on several weighted network datasets by comparing the value of $C_\omega$ on the weighted data to the value of $C$ after converting the weighted data to unweighted data using an arbitrary cutoff. For example, $C_{GT0}$ is found by converting all edges with weights greater than 0 to unweighted edges, and finding the usual global clustering coefficient. Among other results, they find that using different cutoffs for translating weighted graphs into unweighted graphs is an unreliable metric for analyzing the clustering of weighted graphs. They also observe that, since the value of $C_\omega$ is greater than the value of $C_{GT0}$ in all of their datasets, strongly-weighted triplets are more likely to be closed than weakly-weighted triplets.

# 3 Our Work

While the generalization of higher-order clustering coefficients proposed by Yin et al. (2018) provides valuable insights into a multitude of real-world examples, it is limited to graphs which are undirected and unweighted. Many real-world networks do not conform to these constraints, however. Consider the simple example of a road map. One-way roads represent connections in the graph which are inherently directed. Additionally, much of the richness of a road map comes not from its connections, but from its weighted distances. For this project, we have chosen to focus solely on incorporating weights into the model of higher-order clustering coefficients, combining the works of Opsahl and Panzarasa (2009) with Yin et al. (2018).

There are multiple ways of interpreting weights in the higher-order clustering framework, and exploring the characteristics, benefits, and downsides of each of these different methods can be an interesting foray. Analyzing the application of different weighting schemes on differently structured datasets could also provide valuable insight.

# 4 Extension and Analysis

## 4.1 Global Clustering Coefficient

We first begin our extension by focusing on the $G_{n,p}$ model introduced by Erdős and Rényi. We extend the model by further requiring that all edges that end up present in the random graph are assigned a weight uniformly and independently at random from the interval $[0, a]$.

With the new weightings on the edges, we need a way to be able to weigh different components of the graph. Thus, we define the weighting of an $\ell$-wedge *wedge* composed of weighted edges as $\omega(wedge)$, where $\omega$ is a function over the edges composing the wedge. Namely, we will consider $\omega(wedge)$ as the average weight and the minimum weight $\left(\omega_{avg}(wedge) = \frac{1}{|wedge|}\sum_{e\in wedge} w_e \text{ and } \omega_{min}(wedge) = \min_{e\in wedge} w_e \text{ respectively}\right)$.

We modify the definition of the higher-order global clustering coefficient given in Yin et. al (2018) to take into account the weightings of the wedges.

$$C_\ell = \frac{\sum_{w\in\widetilde{W}_\ell}\omega(w)}{\sum_{w\in W_\ell}\omega(w)}$$

where $W_\ell$ is the set of all $\ell$-wedges, and $\widetilde{W}_\ell$ is the set of all $\ell$-wedges that form an $(\ell+1)$-clique when edges are induced amongst the nodes (i.e. all closed $\ell$-wedges).

The computation of the expected global clustering coefficient of a random graph $G$ drawn from the modified $G_{n,p}$ model is similar to that in Yin et al.

$$\begin{aligned}
E_G[C_\ell] &= E_G[E_{W_\ell}[C_\ell | W_\ell]] \\
&= E_G\left[E_{W_\ell}\left[\frac{1}{\sum_{x \in W_\ell} \omega(x)} \sum_{w \in W_\ell} \omega(w) P[w \text{ is closed}]\right]\right] \\
&= E_G\left[E_{W_\ell}\left[\frac{1}{\sum_{x \in W_\ell} \omega(x)} \sum_{w \in W_\ell} \omega(w) p^{\ell-1}\right]\right] \\
&= E_G[E_{W_\ell}[p^{\ell-1}]] \\
&= p^{\ell-1}
\end{aligned}$$

Interestingly enough, this is the same result as that for unweighted graphs. This implies that in the $G_{n,p}$ null model with independently assigned weights, no matter which weighting $\omega$ is used, the expected ratio of closed wedges to total wedges still determines the global clustering coefficient. This is because the model does not lend itself to generating a graph with a great degree of inherent structure, and the weights assigned end up averaged out due to their independence from each other. Thus we still have a simple to understand unstructured base model which we can compare more structured data sets against.

However, it can be said that, in most real-world networks, weights are not assigned independently of the structure of the underlying graph. For example, thinking about a network of flights across different airports (with edges connecting airports and weights equal to the number of flights per day going across the connections), it is far more likely that there are more flights per day between nodes with high centrality in the network, which are likely to be hubs, than there is between two very remote regional airports (if there even is a connection between them).

To attempt to model this dependence of the weights on the underlying structure, we will assign the weights based on a distribution parametrized over the structure. One such way to accomplish this is to draw the weight for an edge $(u, v)$ from the Poisson distribution $Pois(d_u + d_v)$, where $d_u, d_v$ are the degrees of $u$ and $v$. As hubs have high degrees, this will act to weigh edges stemming from hubs more heavily than edges connecting sparsely connected nodes.

We now attempt to analyze the expected global clustering coefficient with such a dependence, to compare to the fully independent case. We take $\omega$ to be the sum of the weights in the wedge. Note that this is the same as taking $\omega$ as the average, as the division by $\ell$ cancels out in both numerator and denominator when evaluating the global clustering coefficient.

$$\begin{aligned}
E_G[C_\ell] &= E_G[E_{W_\ell}[C_\ell | W_\ell]] \\
&= E_G\left[E_{W_\ell}\left[\frac{1}{\sum_{x \in W_\ell} \omega(x)} \sum_{w \in W_\ell} \omega(w) P[w \text{ is closed} \,|\, \omega(w)]\right]\right]
\end{aligned}$$

Define $x_w$ as the endpoints of the wedge $w$. Let $E_{v,w}$ be the event that an endpoint node $v \in x_w$ does not have a connection to all other endpoints in $x_w$.

$$\begin{aligned}
P[w \text{ is closed} \,|\, \omega(w)] &= 1 - P\left[\bigcup_{v \in x_w} E_{v,w} \,\middle|\, \omega(w)\right] \\
&\geq 1 - \sum_{v \in x_w} P[E_{v,w} \,|\, \omega(w)] \\
&= 1 - \ell \, P[E_{v,w} \,|\, \omega(w)]
\end{aligned}$$

$$P[E_{v,w} \,|\, \omega(w)] = \sum_{d=0}^{n-1} P[E_{v,w} \,|\, \omega(w), d_v = d+1] P[d_v = d+1 \,|\, \omega(w)] \quad \text{[by marginalization over } d_v\text{]}$$

For $d \geq \ell$,

$$\begin{aligned}
P[E_{v,w} \,|\, \omega(w), d_v = d+1] &= P[E_{v,w} \,|\, d_v = d+1] \\
&= 1 - \underbrace{\frac{n-\ell-1}{n-1} \cdot \frac{n-\ell-2}{n-2} \cdot \frac{n-\ell-3}{n-3} \cdot \ \cdots \ \cdot \frac{d-\ell+1}{d+1}}_{(n-1)-d \text{ terms}} \\
&\leq 1 - \frac{n-\ell-1}{n-1} \cdot \frac{n-\ell-2}{n-2} \cdot \frac{n-\ell-3}{n-3} \cdot \ \cdots \ \cdot \frac{d-\ell+1}{d+1} \cdot \frac{d-\ell}{d} \\
&\leq 1 - \left(\frac{d-\ell}{d}\right)^{n-d} \\
&= 1 - \left(1 - \frac{\ell}{d}\right)^{n-d} \qquad \text{[assume that } d \geq n/c \text{ for some constant } c\text{]} \\
&\leq 1 - \left(1 - \frac{c\ell}{n}\right)^{n(1-\frac{1}{c})} \\
&\leq 1 - 4^{-c\ell(1-\frac{1}{c})} = 1 - 4^{-\ell(c+1)}
\end{aligned}$$

This result can then be substituted back into $P[E_{v,w} \,|\, \omega(w)]$ to get an upper bound, which can then be used in $P[w \text{ is closed} \,|\, \omega(w)]$ to get a further lower bound. However, evaluating or bounding $P[d_v = d+1 \,|\, \omega(w)]$ turns out to be difficult, and a far-fetched assumption $d \geq n/c$ is needed... The random graph may be too general of a structure, so it can indeed be the case that computing this bound is actually intractable. Intuitively however, the expected value of the global clustering coefficient with such a dependent weighting scheme should approach 1 as $n$ gets larger, since it is more likely that edges with higher weights correspond to higher degree nodes, which makes them more likely to close the wedge.

#### 4.1.1 Application to Watts-Strogatz

We begin by taking the Watts-Strogatz model with $N$ nodes and $K = 1$, so that each node is only connected to two of it's immediate neighbors on the ring, but with no rewiring. Instead, two non-adjacent nodes $v$ and $w$ are connected by a *long* edge with probability $\frac{1}{Dist[v,w]^c}$, where $c$ is a constant greater than 0, and $Dist[v,w]$ is the minimum number of hops along the ring needed to reach $w$ from $v$. We assign a weight to each edge $(v,w)$ present to $Dist[v,w]$, so that edge weights along the ring are equal to 1, and weights of long edges are equal to their hop length.

We take one of the simpler weighted clustering coefficients, $C_2$ (triadic closure), which is easier to explicitly analyze over this structure. There are two types of substructures that can form triadic closures in this model: the 3-clique formed by 2 adjacent edges along the ring and a single long edge with weight 2 (type 1), or a 3-clique formed by 2 adjacent long edges with a difference in weight of one and a single ring edge (type 2).

Using the mean metric for $\omega$, we can compute bounds on the expectations of both the total weighted sum of all 2-wedges and the weighted sum of all closed 2-wedges.

There are $N$ possible type 1's that can arise, each with a probability of $\frac{1}{2^c}$, since that is the probability that the long edge of length 2 will exist. Each one contributes weights of $1 + \frac{3}{2} + \frac{3}{2} = 4$. Thus, type 1 triadic closures contribute an expected weight of $4N\frac{1}{2^c} = \frac{N}{2^{c-2}}$.

The computation for type 2's is more tricky, but can be bounded from above. The probability of two adjacent long edges stemming from the same node with weights $n$ and $n+1$ being both

present (and thus forming a type 2) is bounded above by $\frac{1}{n^{2c}}$ and below by $\frac{1}{(n+1)^{2c}}$.

$$\begin{aligned}
E[\text{Weights of all type 2's}] &\leq 2N \sum_{i=2}^{N/2} \frac{1}{i^{2c}} \left( \frac{(i+1)+1}{2} + \frac{i+1}{2} + \frac{i+(i+1)}{2} \right) \\
&= 2N \sum_{i=2}^{N/2} \frac{1}{i^{2c}} (2i+2) \\
&= 4N \sum_{i=2}^{N/2} \frac{1}{i^{2c-1}} + 4N \sum_{i=2}^{N/2} \frac{1}{i^{2c}}
\end{aligned}$$

Thus, $E\Big[ \sum_{w \in \widetilde{W_2}} \omega(w) \Big] \leq N \Big( \frac{1}{2^{c-2}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^{2c-1}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^{2c}} \Big)$.

$$\begin{aligned}
E\Big[ \sum_{w \in W_2} \omega(w) \Big] &\geq N + \underbrace{2N \sum_{i=2}^{N/2} \frac{1}{i^c} (4 \frac{1+i}{2})}_{\text{long edges}} \\
&= N \Big( 1 + 4 \sum_{i=2}^{N/2} \frac{1}{i^{c-1}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^c} \Big)
\end{aligned}$$

The latter inequality holds because we do not account for the weights of 2-wedges formed by two adjacent long edges.

$$\text{This gives us that } \frac{E\big[\sum_{w \in \widetilde{W_2}} \omega(w)\big]}{E\big[\sum_{w \in W_2} \omega(w)\big]} \leq \frac{N \Big( \frac{1}{2^{c-2}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^{2c-1}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^{2c}} \Big)}{N \Big( 1 + 4 \sum_{i=2}^{N/2} \frac{1}{i^{c-1}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^c} \Big)} = \frac{\frac{1}{2^{c-2}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^{2c-1}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^{2c}}}{1 + 4 \sum_{i=2}^{N/2} \frac{1}{i^{c-1}} + 4 \sum_{i=2}^{N/2} \frac{1}{i^c}}$$

For $c \gg 2$, the summations are all convergent and roughly equal to 0 each as $N$ approaches infinity, giving an upper bound for the ratio of $\frac{1}{2^{c-2}}$. However, once $c = 2$, the left-most summation in the denominator diverges as $N$ grows, bringing the upper bound down to 0 in the limit case. This behavior continues for $0 < c \leq 2$, since this summation diverges faster than any of the other ones. This is kind of counterintuitive, as one would expect the upper bound to actually shrink when $c$ increases, as less edges to close off 2-wedges will be present. This can be explained due to the fact that when $c$ decreases, the amount of long edges present increases, causing them to be counted more towards the total weighted sum of all 2-wedges, but it is still rare that they are adjacent and have a difference in length of exactly 1, so their weight ends up only in the denominator. Interestingly enough, when $c = 0$ (the graph is an $N$ clique), the leftmost summations in the numerator and denominator diverge at a faster rate than the other two, and are equal, giving an upper bound of 1 in the limit case (which is the actual value of the global clustering coefficient in this case).

## 4.2 Local Clustering Coefficient

Similarly, we can generalize a definition of the local higher-order weighted clustering coefficient for node $u$:

$$C_{\ell\omega}(u) = \frac{\sum_{w \in \widetilde{W}_\ell(u)} \omega(w)}{\sum_{w \in W_\ell(u)} \omega(w)}$$

with $W_\ell$ the set of $\ell$-wedges centered at node $u$, and $\widetilde{W}_\ell(u)$ the set of $\ell$-wedges centered at node $u$ which form an $(\ell+1)$-clique when edges are induced among the nodes. As before, the $\omega$ function defines the value of a wedge, which can use any of the four metrics already described.

One interesting question about this metric is whether the extremal bounds proved in Yin et al. (2018) proved for the unweighted version still hold for this weighted generalization. The unweighted

result states that for $\ell \geq 3$, $0 \leq C_\ell(u) \leq \sqrt{C_2(u)}$. We have found that this inequality does not hold for the weighted case.

A counterexample can be constructed that has the structure of the half-star, half-complete graph given in Yin et al. (2018). We will start with a specific example. Start with node $u$ and a complete graph on 5 vertices, with edge weights of 10. Connect $u$ to each of the nodes in this complete graph with edges of weight 1. Connect $u$ to five other nodes with edges of weight 10. Now we will calculate $C_{\ell\omega(u)}$, using the maximum value metric on the wedges. Using the definition, we find that

$$C_{2,max}(u) = \frac{20(1)}{20(1) + 25(10) + 25(10) + 20(10)} = 0.028$$

$$C_{3,max}(u) = \frac{60(10)}{60(10) + 100(10)} = 0.375$$

In particular, $C_{3,max}(u)$ is not bounded above by $\sqrt{C_{2,max}(u)} = 0.17$.

We can generalize this result by considering a node $u$ connected to a complete subgraph on $\frac{d}{2}$ nodes on one side and to $\frac{d}{2}$ other nodes on the other. We will use a weight of $b$ for the edges within the $\frac{d}{2}$-clique, a weight of $a$ for the edges connecting $u$ to the $\frac{d}{2}$-clique, and a weight of $b$ for the edges connecting $u$ to the other $\frac{d}{2}$ nodes. It can be found that

$$C_{2,max}(u) = \frac{(d-2)a}{(d-2)a + (3d-2)b}$$

$$C_{3,max}(u) = \frac{(d-4)b}{(d-4)b + (d)b} = \frac{d-4}{2d-4}$$

In particular, it is possible to force $C_{2,max}$ to approach 0 by setting $b \gg a$, and possible to force $C_{3,max}$ to $\frac{1}{2}$ by choosing a large $d$.

If we generalize this result to even higher orders of clustering, we find that

$$C_{n,max}(u) = \frac{b[\prod_{i=0}^{n-1}(\frac{d}{2} - i)]}{b[\prod_{i=0}^{n-1}(\frac{d}{2} - i)] + b[\prod_{i=0}^{n-2}(\frac{d}{2} - i)](\frac{d}{2})} = \frac{d - (2n-2)}{2d - (2n-2)}$$

Thus, for sufficiently large $d$, all weighted clustering coefficients for order $n \geq 3$ approach $\frac{1}{2}$, while $C_{2,max}$ remains arbitrarily small. What this example captures is that patterns of weighted clustering may not emerge until higher order of clustering are introduced. The flexibility of weighting means that the extremal bounds do not generalize well.

This example can be constructed for the maximum weighting metric since the maximum metric is very sensitive: as soon as a large-weight edge is added to an $\ell - wedge$, the value of the entire wedge is immediately set equal to the weight of that edge. It is possible that other metrics, like mean and median, could be more sensitive to such variation in edge weights. Thus there could be extremal bounds for other types of weighting coefficients, which we leave as an open question.

### 4.2.1 Analysis of Local Coefficient on a Random Graph Model

We can analyze the local weighted higher-order clustering coefficient on a simple random graph model in order to demonstrate the kinds of patterns it can capture in a network.

To create our model, we will define two sets of nodes, $A$ and $B$. Nodes in set $A$ have no edges between them; nodes in set $B$ have edges for every pair of nodes. Given a node in $A$ and a node in $B$, there is a $p = \frac{1}{2}$ probability that there is an edge between them. We will define our graph to have $n$ nodes in set $A$ and $n+1$ nodes in set $B$. We will analyze the values of $C_2(v)$ and $C_{2,avg}(v)$ for this network under different weightings (with $avg$ referring to the arithmetic mean).

First, consider a node $v$ in set $A$. Since all of $v$'s neighbors are in set $B$, which is fully connected, $E[C_2(v)] = 1$. In fact, for any set of weights we put on the edges of this network, $E[C_{2,avg}(v)] = 1$.

Thus we will focus our analysis on nodes from set $B$.

Consider a node $v$ in set $B$. $v$ has expected degree of $n+\frac{n}{2}$. Thus the expected number of 2-wedges centered at $v$ is $\binom{\frac{3n}{2}}{2} = \frac{9}{8}n^2 - \frac{3}{4}n$. This breaks into $\frac{1}{2}n(n-1)$ wedges consisting of two $B-B$ edges, $\frac{1}{2}n^2$ wedges consisting of one $B-B$ edge and one $B-A$ edge, and $\frac{n}{4}(\frac{n}{2}-1)$ wedges consisting of two $B-A$ edges. The expected of number of closed wedges is $\frac{1}{2}n(n-1) + \frac{1}{2}n(\frac{n}{2}) = \frac{3}{4}n^2 - \frac{1}{2}n$. Thus the expected value of $C_2(v)$ is $\frac{\frac{3}{4}n-\frac{1}{2}}{\frac{9}{8}n-\frac{3}{4}} \to \frac{2}{3} = 0.67$. Thus, using the unweighted metric, nodes in set $B$ have a relatively high level of (order 2) clustering.

Now we can add weights to the model, and see how it affects the value of $C_{2,avg}(v)$ for a node $v$ in set $B$. One might assume that the nodes in set $B$, since they are so highly interconnected, form stronger bonds with each other than they would with the sparse nodes of set $A$. Thus we can assign a weight of 3 to every edge within set $B$, and a weight of 1 to every edge connecting set $B$ to set $A$. Now, using the average weighting metric, we get the expected total value of the 2-wedges centered at $v$ to equal $\frac{3}{2}n(n-1)+2n\frac{n}{2}+\frac{n}{4}(\frac{n}{2}-1) = \frac{21}{8}n^2 - \frac{7}{4}n$. The expected total value of closed 2-wedges equals $\frac{3}{2}n(n-1) + n(\frac{n}{2}) = 2n^2 - \frac{3}{2}n$. Thus the expected value of $C_{2,avg}(v)$ is $\frac{2n-\frac{3}{2}}{\frac{21}{8}n-\frac{7}{4}} \to \frac{16}{21} = 0.76$. Thus, when we add in a system of weighting that correlates high degree with high weights, we see that the expected weighted clustering around a node increases.

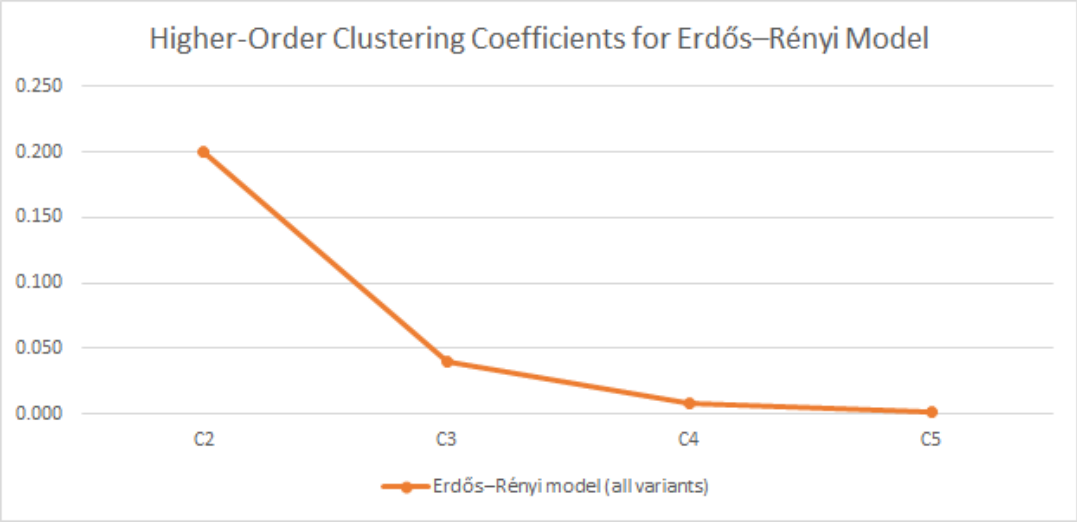
We can add an alternative style of weighting that correlates lower degree with higher weights. An interpretation of this phenomenon is that, if a node has less relationships, it can invest more resources into each relationship. If we assign a weight of 3 to every edge connecting set $B$ to set $A$, and a weight of 1 to every edge within set $B$, we can perform analogous analysis to yield that the expected value of $C_{2,avg}$ equals $\frac{n^2-\frac{1}{2}n}{\frac{15}{8}n^2-\frac{5}{4}n} \to \frac{8}{15} = 0.53$. Predictably, this scheme reduces the expected amount of weighted clustering around a node.

This example demonstrates the ways in which different weighting schemes can affect the clustering value around a node: if weights are associated with higher degree, it implies that a wedge made of higher relative weights is more likely to be closed. If weights are associated with lower degree, it means that a wedge made of lower relative weights is more likely to be closed. We repeated this analysis with a more detailed model ($p = 0.75$ for $B-B$ edges, $p = 0.5$ for $A-B$ edges, $p = 0.25$ for $A-A$ edges), and found that the same general patterns result.

# 5 Erdős–Rényi Graph

To verify our findings above, we generated a 1000 node Erdős–Rényi graph with an edge-formation probability of .2. Additionally, we added weights uniformly on the edges within the range [0, 20]. As expected, our model produced clustering coefficients which were independent of the weighting $\omega$ used, and consistent with the results of Yin et al. (2018), as shown in the figures below.

| **Graph Model** | **Weighting** | **C2** | **C3** | **C4** | **C5** |
|---|---|---|---|---|---|
| Erdős–Rényi | unweighted | .200 | .040 | .008 | .002 |
| Erdős–Rényi | mean | .200 | .040 | .008 | .002 |
| Erdős–Rényi | median | .200 | .040 | .008 | .002 |
| Erdős–Rényi | maximum | .200 | .040 | .008 | .002 |
| Erdős–Rényi | minimum | .200 | .040 | .008 | .002 |




# 6 Bitcoin Alpha Trust Network

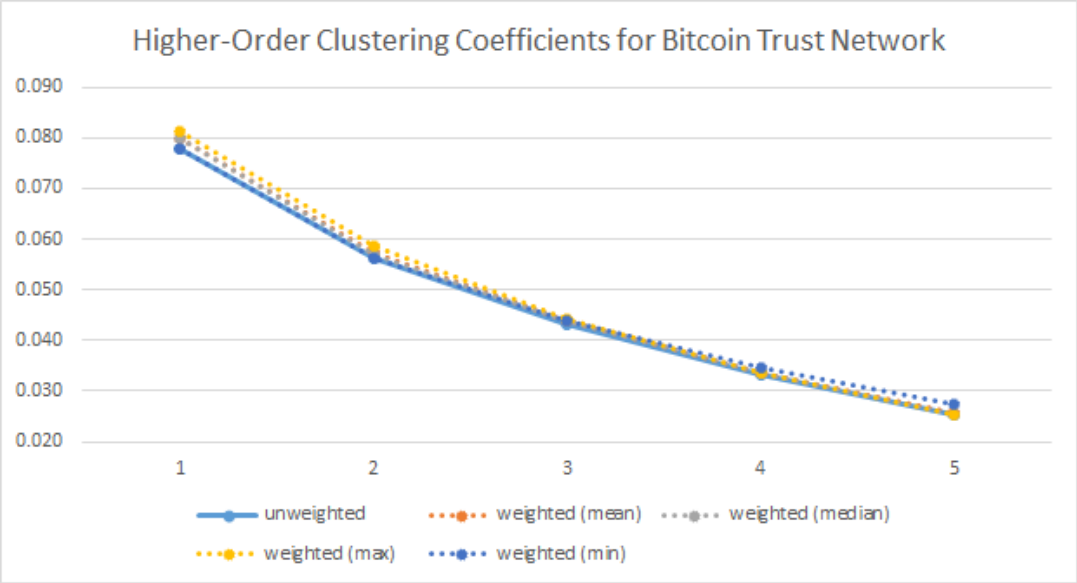
We tested our model on weighted data available through the the Stanford Large Network Dataset Collection, specifically **bitcoin-alpha** [2]. This network is a who-trusts-whom network of users who trade Bitcoin on the Bitcoin Alpha platform. As presented in the dataset, the graph is directed, where a directed edge $(u, v)$ with weight $w$ conveys that user $u$ interacted with user $v$ and assigned $v$ a reputation of $w$ where $w$ is an integer value ranging from -10 (very untrustworthy) to 10 (very trustworthy).

To make this graph compatible with our model, we first added 10 to all edge weights, ensuring non-negative edge values. Additionally, we reinterpreted the graph representation of the data such that it was undirected and thus compatible with our model. More explicitly, for all directed edges $(u, v)$ and $(v, u)$, we merged these edges into an undirected edge $(u, v)$ corresponding to the minimum reputation $u$ and $v$ assigned to each other, modeling mutual trust.

We hypothesized that components with high mutual trust are more likely to cluster, which would be observable through higher clustering coefficients. We include data below.

## 6.1 Data

| **Graph Model** | **Weighting** | **C2** | **C3** | **C4** | **C5** | **C6** |
|---|---|---|---|---|---|---|
| bitcoin-alpha | unweighted | .078 | .056 | .043 | .033 | .025 |
| bitcoin-alpha | mean | .080 | .057 | .044 | .034 | .026 |
| bitcoin-alpha | median | .080 | .057 | .044 | .034 | .025 |
| bitcoin-alpha | maximum | .081 | .059 | .044 | .034 | .025 |
| bitcoin-alpha | minimum | .078 | .056 | .044 | .035 | .027 |




### 6.2 Discussion

Comparing the unweighted and weighted clustering coefficients for the Bitcoin Alpha trust network, we observe a slight increase in clustering coefficients for all weighted graph models when compared to the unweighted graph model. Specifically when using the minimum weighting metric, we notice that the difference in clustering coefficients is most prominent for higher indexes of the clustering coefficient. This seems to suggest that mutual trust does increase the tendency for components to cluster, although to a very slight degree. However, this cannot be definitive due to the limited size of the graph and the minimal impact of using a weighted model.

Perhaps more interestingly, for large indexes of clustering coefficients, the Bitcoin Alpha trust network shows a much higher degree of clustering than the randomly generated Erdős–Rényi model. However, this difference is not apparent when only looking at lower-order clustering coefficients, specifically triadic closure.

## 7 Friendship Network

Using data collected on students from the University of California, we tested our model on a friendship network of roughly 2,500 students [1]. Each student was asked to indicate their 5 closest female friends and 5 closest male friends. A directed graph was formed from this data where a directed edge $(u, v)$ represents that student $u$ listed student $v$ as one of their closest friends. Additionally, students were asked to report how often they interacted with the friends they listed, providing edge weights for the friendship network.

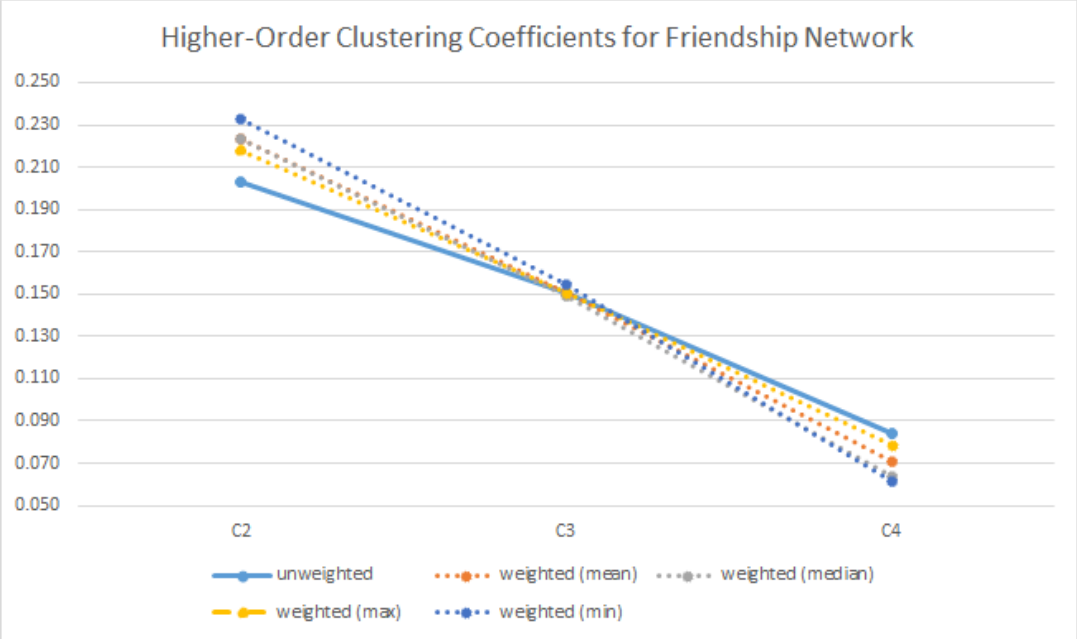
To make this graph compatible with our model, we needed to reinterpret the data such that it could be represented as an undirected graph. To do so, we created a graph where the weights of undirected edges corresponded to the closeness of two friends. If a directed edge $(u, v)$ existed in the original graph, but the edge $(v, u)$ did not, we removed any connections between the two. This is because close friendships by their nature must be reciprocated. Additionally, if a directed edge $(u, v)$ existed in the original graph with weight $w_1$ and the directed edge $(v, u)$ existed with weight $w_2$, the undirected graph would have an undirected edge $(u, v)$ with weight $\min(w_1, w_2)$. Again, this goes back to the premise that close friendships must be reciprocated. If one friend thinks very highly of the other but that is not reciprocated, it's probably not a very close friendship.

We hypothesized that regardless of the index of the clustering coefficient, or the weighing metric used, the weighted clustering coefficients would be greater than the unweighted clustering coefficients. The reasoning for this hypothesis is mostly intuitive: the closer that friends are, the more likely they are to share mutual friends. Close friends often share common interests, so it seems

likely that they will share many common friends. And it seems natural enough that this would generalize for higher orders of clustering as well. However, this hypothesis turned out to be very far from the truth, as shown in the data below.

## 7.1 Data

| Graph Model | Weighting | C2 | C3 | C4 |
|---|---|---|---|---|
| friendship network | unweighted | .203 | .150 | .084 |
| friendship network | mean | .223 | .150 | .071 |
| friendship network | median | .223 | .149 | .064 |
| friendship network | maximum | .218 | .151 | .079 |
| friendship network | minimum | .232 | .155 | .061 |




## 7.2 Discussion

Observing the weighted clustering in the friendship network, we notice a stark divergence from our hypothesis. When examining triadic closure, we see that using weighted higher-order clustering results in a higher clustering coefficient, regardless of the weighting metric used. However, for larger indexes of the clustering coefficient, the opposite result is seen. For this reason, we believe this network provides an interesting example where weighted higher-order clustering provides insights into the network which neither higher-order clustering or weighted triadic closure provided on their own.

Without taking a more qualitative view of the friendship network, it is difficult to pinpoint the exact reason why higher indexes of the clustering coefficient are lower when incorporating weights. So, it's important to note that any explanations which follow are merely conjecture.

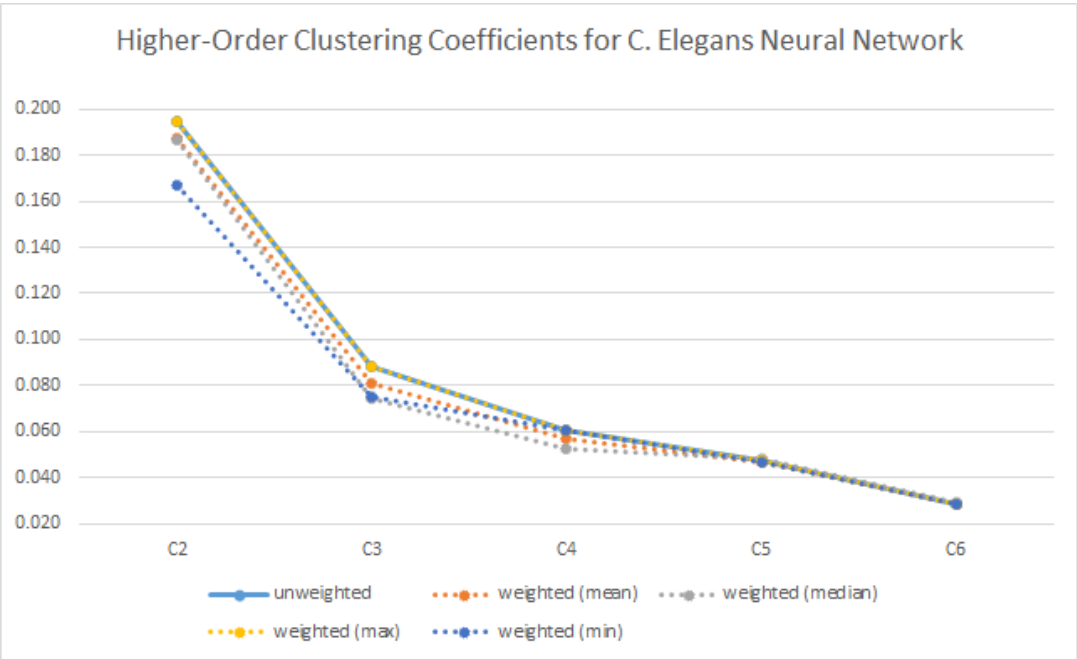
Although the weights of wedges are directly influenced by the weights of edges, they're still being affected indirectly by the nodes which make up the wedge. We predict that highly social people (those who tend to interact often with others) are more likely to have diverse friend groups which are unlikely to form large clusters. These highly social people, due to their increased level of interactions, are members of wedges with higher weights. And thus, larger indexes of clustering coefficients are skewed downwards when weighting is introduced.

# 8 C. Elegans Neural Network

Finally, we examined a dataset which included two-dimensional spatial positions of rostral gangilia neurons within the neural network in *C. elegans*, a type of worm [3]. Weights in this graph are the number of synapses and gap junctions between neurons. In order to fit within the constraints of our model, we removed direction from edges, combining weights of opposing edges when they existed. Results for this network are shown below.

## 8.1 Data

| Graph Model | Weighting | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| C. Elegans neural network | unweighted | .194 | .088 | .060 | .047 | .029 |
| C. Elegans neural network | mean | .187 | .081 | .057 | .047 | .029 |
| C. Elegans neural network | median | .187 | .074 | .053 | .048 | .029 |
| C. Elegans neural network | maximum | .194 | .088 | .060 | .047 | .029 |
| C. Elegans neural network | minimum | .167 | .075 | .069 | .047 | .029 |



### 8.2 Discussion

For this network, we noticed that using the max weighting metric had no effect on clustering coefficients. All other weighting metric produced differing results, the results being more prominent with the minimum weighting metric. Most interestingly, we see that weighting clustering has an effect for low cluster orders but little to no effect for larger orders of clusters. Thus, this network provides a useful example for why extending weighted clustering beyond triadic closure can provide further insights into a network's structure. More explicitly, we see that the decrease in clustering when incorporating weights is very shallow, and while it may seem significant when looking at triadic closure, plays much less of a role for higher orders of clustering.

## 9 Future Directions

A large limitation of our model is that it does not support directed graphs. And many networks, especially weighted ones, are directed. In fact, all of the weighted data sets we used were directed and needed to be reinterpreted as undirected networks. Further extending our model to support directed graphs is an insightful and potentially very useful direction for future research.

## References


[1] Adolescent health network dataset – KONECT, April 2017.

[2] Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 221–230. IEEE, 2016.

[3] Jure Leskovec and Rok Sosič. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1, 2016.

[4] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social Networks*, 31(2):155 – 163, 2009.

[5] Hao Yin, Austin R. Benson, and Jure Leskovec. Higher-order clustering in networks. *CoRR*, abs/1704.03913, 2017.