# CSCI 5352 Problem set 2

Jake Krol

February 2025

## 1 Q1

Define the closeness centrality as

$$C_i = \frac{1}{l_i}$$

where $l_i$ is the mean shortest path from node $i$ to all other $j \neq i$ nodes in the graph.

$$l_i = \frac{1}{n}\sum_{j}^{n} d_{ij}$$

For clarity, $d_{ij}$ is the shortest path from node $i$ to $j$. **The summation $l_i$ can be decomposed as the sum of the distances to nodes in each subnetwork $n_A$ and $n_B$.** Do this for both $l_A$ and $l_B$.

$$l_A = \frac{1}{n}(\sum_{j}^{n_A} d_{Aj} + \sum_{j}^{n_B} d_{Aj})$$
$$l_B = \frac{1}{n}(\sum_{j}^{n_A} d_{Bj} + \sum_{j}^{n_B} d_{Bj})$$

With the defined network, **notice that every distance from node $A$ to a node in subnetwork $n_B$ is just one plus the distance from $B$ to $j \in n_B$.** And vice versa, when traveling from $B$ to nodes $j \in n_A$. Therefore, we can rewrite these summations.

$$l_A = \frac{1}{n}(\sum_{j}^{n_A} d_{Aj} + \sum_{j}^{n_B} (d_{Bj} + 1))$$
$$l_B = \frac{1}{n}(\sum_{j}^{n_A} (d_{Aj} + 1) + \sum_{j}^{n_B} d_{Bj})$$

Use the fact that summation is distributive over addition.

$$l_A = \frac{1}{n}(\sum_j^{n_A} d_{Aj} + \sum_j^{n_B} d_{Bj} + n_B)$$
$$l_B = \frac{1}{n}(\sum_j^{n_A} d_{Aj} + n_A + \sum_j^{n_B} d_{Bj})$$

Finally, the relation

$$\frac{1}{C_A} + \frac{n_A}{n} = \frac{1}{C_B} + \frac{n_B}{n}$$

is proven by substiting $\frac{1}{C_A} = l_B$ and $\frac{1}{C_B} = l_B$

$$\frac{\sum_j^{n_A} d_{Aj} + \sum_j^{n_B} d_{Bj} + n_B}{n} + \frac{n_A}{n} = \frac{\sum_j^{n_A} d_{Aj} + n_A + \sum_j^{n_B} d_{Bj}}{n} + \frac{n_B}{n}$$
$$\frac{\sum_j^{n_A} d_{Aj} + \sum_j^{n_B} d_{Bj} + n_B + n_A}{n} = \frac{\sum_j^{n_A} d_{Aj} + n_A + \sum_j^{n_B} d_{Bj} + n_B}{n}$$

If we divide each side by the numerator and take the reciprocal of the result, then we would get $n = n$ which is obviously true, so the proposed relation must also be true.

# 2 Q2

## 2.1 Q2a

### 2.1.1 Q2ai criteria

Criteria: To do an edge double swap on a directed graph, choose two edges, $(u, v), (x, y)$, and **fix the inward stubs while swapping the outward stubs**: fix $v$,$y$ and swap $u$,$x$. The two **randomly chosen edges must not have equal inward nor outward stubs** since this would not transform a vertex-labeled graph; the resulting edges would be the same. Formally, for $(u, v), (x, y)$

$$\begin{aligned} &\text{for } (u, v), (x, y) \\ &\text{if } v \neq y \ \wedge u \neq x \\ &\text{then output } (x, v), (u, y) \end{aligned}$$

This ensures the in and out degree sequences $(k_{in}, k_{out})$ are preserved since $u, x$ keep the same outward degree while $v, y$ the inward and outward stubs stay on their original nodes.

### 2.1.2 Q2aii configurations

Let $f$ be the double edge swap operation. The output configuration is

$$f((u,v),(x,y)) = (\mathbf{x},\mathbf{v}),(\mathbf{u},\mathbf{y})$$

if inward stubs $v, y$ are fixed.

Vice versa, double edge swap also works for directed graphs if $u, x$ are fixed and $v, y$ are swapped instead.

$$f((u,v),(x,y)) = (\mathbf{u},\mathbf{y}),(\mathbf{x},\mathbf{v})$$

### 2.1.3 Q2aiii checks

Checks

- **Check that no self loops are created** by verifying outward stubs are not from the same node $x \neq u$.
- **Check that the graph does not become multiedge** by verifying that transformed edges are not already in the graph: $(x,v) \notin E_G \wedge (u,y) \notin E_G$. Where, $E_G$ is the edge set of the untransformed graph.

## 2.2 Q2b

### 2.2.1 Q2bi criteria

The degree sequence and node type of a bipartite graph is preserved by a double edge swap which **swaps stubs of nodes in the same set for new edges**.

Formally, given $(u,v),(x,y)$ such that $u,x \in V_1$ and $v,y \in V_2$ ($V_1 \cap V_2 = \emptyset$) swap $v, y$ to get $(u,y),(x,v)$.

Ensure **each edge has a pair of nodes from the disjoint node sets**.

- $u, x \in V_1$
- $v, y \in V_2$

Next, **ensure nodes within the same set, $V_i$, are distinct** to guarantee the swapped edges do not simply replace each other.

- $x \neq y$
- $u \neq v$

### 2.2.2 Q2bii configurations

Let $f$ be the double edge swap operation. The output configuration is

- $f((u,v),(x,y)) = (\mathbf{u},\mathbf{y}),(\mathbf{x},\mathbf{v})$

### 2.3 Q2biii checks

**Verify a multigraph was not created** by ensuring the new edges did not exist in the original edge set, $E_G$, before swapping.

$$(u, y), (x, v) \notin E_G$$

# 3 Q3

## 3.1 Q3i

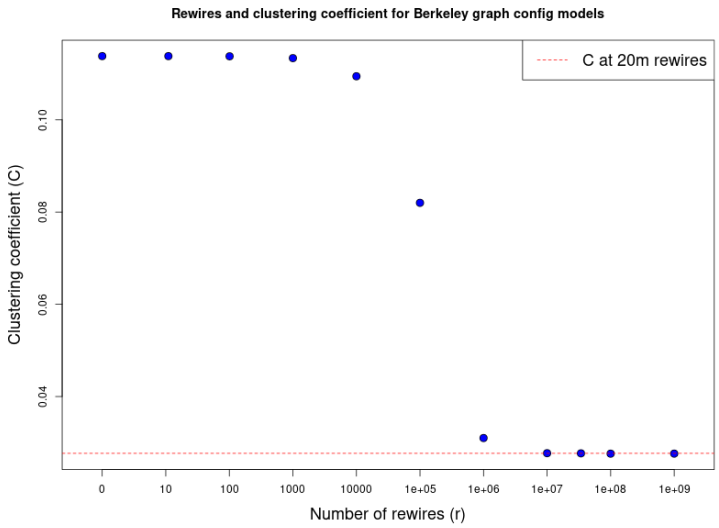


Figure 1: Clustering coefficient of Berkeley graph after double edge swaps (rewires).

The original **clustering coefficient** (C) of the Berkeley Facebook network is **not explained by its degree sequence since many, random double edge swaps cause C to decrease**. With no rewires, $C \approx 0.11$, yet at 10,000 rewires $C$ begins to noticeably decrease. And, after $20m$ ($m :=$ edges) rewires $C$ asymptotes to $\approx 0.028$. Meaning, the Berkeley network had a higher C compared to random graphs with the same degree sequences.

## 3.2 Q3ii

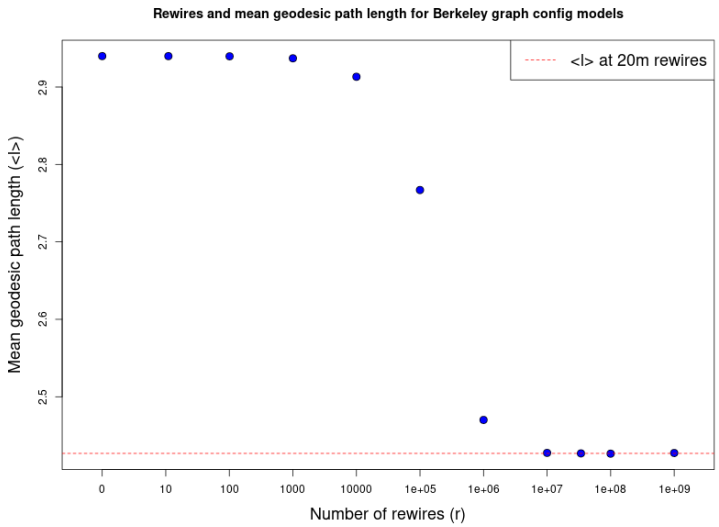


Figure 2: Mean geodesic path length of Berkeley graph after double edge swaps (rewires).

Similar to C, the original **mean geodesic path length ($\langle l \rangle$) of the Berkeley Facebook network is not random since many, random double edge swaps causes $\langle l \rangle$ to decrease**. With no rewires, $\langle l \rangle \approx 2.94$, yet again after 10,000 rewires there's a noticeable decline in $\langle l \rangle$. After $20m$ rewires, $\langle l \rangle$ asymptotes to $\approx 2.4$. So, randomized networks with preserved degree sequence reduced the mean geodesic path length by about half a node.

# 4 Q4

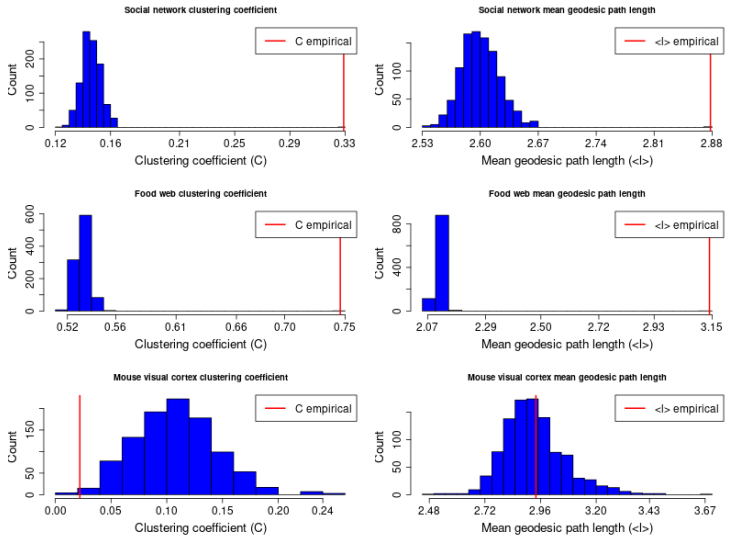


Figure 3: Null distributions of clustering coefficients ($C$) and mean geodesic path length ($\langle l \rangle$) for configuration models of social, food-web, and connectome network. Empirical values of $C$ and $\langle l \rangle$ for each network is denoted by a red, vertical line.

Preface: A social, a food-web, and a connectome network were found using the Colorado Index of Complex Networks (ICON) database and downloaded from the original sources. For each graph and each statistic, a null distribution was generated using configuration models. The first configuration model was sampled after $20m$ ($m :=$ number of edges) degree-preserving, double-edge swaps (rewires). After $20m$ rewires, the next configuration model was sampled after $2m$ more rewires, then a new config model was sampled after $2m$ more rewires ... The sampling was repeated until 1000 null networks were generated. For each config model network, the clustering coefficient ($C$) and the mean geodesic path length ($\langle l \rangle$) were computed.

## 4.1 Q4 social

The **social** network chosen was the within-organization Facebook friendships (2013). **i) C** of the network was **high** ($C \approx 0.33$) and **atypical** relative to the null distribution (Figure 3, top-left), also $\langle$**l**$\rangle$ was **high** ($\langle l \rangle \approx 2.88$) and **atypical** relative to the null distribution. **ii)** Therefore, the **structure** of the social network has relatively more triangles and nodes are slightly more distance to one another (on average) compared to a random network with the same degree sequence. **iii)** I **hypothesize** that this social network has more triangles

because Facebook actively encourages this structure by recommending mutual friends. And, I **hypothesize** that maybe the higher $\langle l \rangle$ is explained by social grouping behavior. Where, social groups with conflicting interests could create long paths in the network.

## 4.2 Q4 food-web

The **food-web** network chosen was the Chengjiang Shale food-web depicting predation relationships of ancient organisms. **i)** both empirical **C** and $\langle \mathbf{l} \rangle$ are **large and atypical** compared to the null distributions. **ii)** Similar to the social network, high $C$ and $\langle l \rangle$ implies that the network **structure** has many triangles, for any existing triads, and the average distance between nodes relatively long compared to random graphs with this degree sequence. **iii)** I **hypothesize** that the large amount of triangles is likely due to the nature of the study and how the network was constructed. The network was originally directed and weighted to score putative predator-prey relationships, and I think it makes sense for the study to be exhaustive in trying to compare all pairwise species relations. The few missing links of all pairwise connections could be explained by obvious evidence that species do not have a predator-prey interaction, such as a very large arthropod and a single cell organism. Also, I **hypothesize** that the high empirical $\langle l \rangle$ is explained by the hierarchical nature of food-webs where there are "top" predators and "bottom" preys. Using the same example, I would expect a very large arthropod and a single cell organism to be relatively distant in the network. Therefore, this distance is not random.

## 4.3 Q4 connectome

The **connectome** network chosen was a mouse visual cortex connectome. **i)** The empirical **C** is **small and atypical** compared to the null distribution while $\langle \mathbf{l} \rangle$ is a **middling, typical** value. **ii)** Therefore, the network **structure** has very few triangles, low $C$, for the number of triads, and mean geodesic paths is typical where each node pair can reach other in about 3 steps. **iii)** I **hypothesize** that the low $C$ is explained by the tree-like structure of visual cortex. I think the visual cortex is a hierarchical system where many neurons process the raw incoming information from the eye, and "parent" neurons aggregate information from children neurons into higher-order concepts like shapes. And, tree networks do not have many edges within levels, so this may explain the low $C$. Next, I **hypothesize** that the typical $\langle l \rangle$ is explained by the the tree-like structure. Trees are efficient network structures to carry information long distances by traversing minimal edges. And, a $\langle l \rangle$ ranging from 2 to 3 is very common in networks. Only particular graph structures (such as a first order ring) have a relatively high $\langle l \rangle$.

# 5 Q5

## 5.1 Q5a

| Family name | Harmonic centrality |
|---|---|
| Acciaiuoli | 5.92 |
| Albizzi | 7.83 |
| Barbadori | 7.08 |
| Bischeri | 7.20 |
| Castellani | 6.92 |
| Ginori | 5.33 |
| **Guadagni** | **8.08** |
| Lamberteschi | 5.37 |
| **Medici** | **9.50** |
| Pazzi | 4.77 |
| Peruzzi | 6.78 |
| Pucci | 0.00 |
| Ridolfi | 8.00 |
| Salviati | 6.58 |
| Strozzi | 7.83 |
| Tornabuoni | 7.83 |

Table 1: Harmonic centralities of all families in Medici network

The **Medici** family has the highest harmonic centrality (9.50), and the **Guadagni** family has the second highest harmonic centrality (8.08). The claim that the Medici family's power is due to their centrality in social/business networks is supported by this network since Medici has the highest harmonic centrality. And intuitively, harmonic centrality quantifies the average distance of a node to all others by averaging its reciprocal shortest distance; meaning, the Medici family was a good social/business "route" for other families. And, other families seeing this optimal route to social/business success will want to connect with Medici to advance their own status, creating yet another close path to Medici.

For the Guadagni family with the second highest harmonic centrality, I think their influence is probably less distinguishable since other families have similarity centrailty (e.g., Ridolfi [8.00] and Albizzi [7.83]). Therefore, the value of creating a connection with Guadagni is less important since this relationship could maybe be substituted with other families to achieve a similar path through the social network.

## 5.2 Q5b

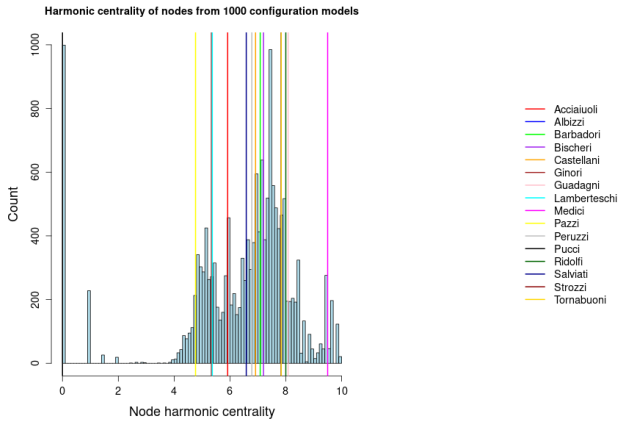


Figure 4: Histogram of harmonic centrality of nodes from 1000 configuration models. Empirical harmonic centralities of families in the Padgette Florentine families network are denoted by vertical, colored lines.

Since the double edge swap procedure performs randomization while preserving the original degree sequence, it can be used to study the extent to which the degree sequence alone explains the harmonic centrality of the Padgette Florentine network. Using a double edge swap procedure on Padgette Florentine families network, 1000 configuration models were generated, and the harmonic centrality of each node in each model is concatenated into a single vector create a null distribution (Figure 4). For clarity, the first configuration model used $20m$ double edge swaps and all successive configuration models were sampled by performing $2m$ more swaps after the previous.

The null distribution of harmonic centralities shows that **most of the observed harmonic centralities in the Padgette Florentine network can be explained solely by the degree sequence, with the Medici, Guadagni, and Pazzi families potentially being exceptions**. First, many of the families have harmonic centralities which occur frequently in the random, configuration models, and this implies these centrality values are common even in completely random networks with the same degree sequence. Next, the Medici, Guadagni, and Pazzi families are in reltively low mass regions of the null distribution. Meaning, the harmonic centralities for Medici, Guadagni, and Pazzi are less likely to be random than other observed centralities. Yet, it is certainly possible to observe these values under purely random circumstances since if we convert this count distribution to a probability mass function, then these probabilities are well above zero.

Also intuitively, it is guaranteed that each config model will have a node with harmonic centrality of 0. I defined the harmonic centrality of nodes with 0 degree as 0, and the degree sequence is fixed. This explains the 1000 count of

nodes with harmonic centrality of 0 in the null distribution (Figure 4).

# 6

# 7 Extra credit 7

- I **read** "Configuring random graph models with fixed degree sequences" (Fosdick, et al. 2017).
- The **research question** is to understand how graph spaces (simple, loopy, multigraph, and loopy multi-graph) and graph type (vertex-labeled or stub-labeled) effect randomly generated configuration models.
- The **approach** is to showcase how configuration model sampling differs for eight different graph spaces using three empirical networks as a case study. The eight common graph spaces are simple, loopy, multi-graph, and loopy-multigraph further stratified by whether the graph is vertex/stub-labeled.
- What the **paper did well** was supplying all pre-requisite knowledge to understand the study. The explanation of graph spaces and how choosing a target graph space can change the results of sampling was conveyed well by Figures 1 and 2. Similarly, the motivation and explanation of using Markov Chain Monte Carlo sampling was detailed and informative. In general, I think the paper was well above average in its detail. In my field, it's common to have papers only present new findings with minimal explanation of the foundational lemmas that the work is based upon.
- I do not see any major gaps that the **paper could improve on**.
- As mentioned in the paper, I think a useful **extension** of the paper is to find ways to sample configuration models for weighted graphs, where weights are not easily translated into multiple edges.

# 8 References


- Fosdick, B. K., Larremore, D. B., Nishimura, J., & Ugander, J. (2017). Configuring random graph models with fixed degree sequences. arXiv. https://arxiv.org/abs/1608.00607


# 9 Code

GitHub.