Centrality Measure Based on Clusterings in the Effective Resistance Embedding Space

Kevin Bui, Joe Feng, Alto Senda, Justin Wang

MATH 191 Graphs and Networks, UCLA

March 19, 2015

Outline

Introduction

In our project:

- Review betweenness and closeness centrality measures.
- Introduce their variants and concept of electrical resistance embedding when visualizing a graph as an electrical network.
- Apply k-means algorithm on the effective resistance embedding to create a new centrality measure called k-means centrality.
- Compare the proposed k-means centrality to the existing centrality measures to see how discriminative it is for the nodes of the graph

Outline

Review of Centrality Measure

Recall that:

- Centrality measures attempt to quantify the importance of nodes, edges, or other network structures.
- A good choice of centrality measure depends on context; what does it mean to be ”most central”?
- A slight change to the definition of a centrality measure can have large changes.

Betweenness Centrality

Betweenness centrality measures the extent on which a node lies on the shortest path of a graph.

It is defined as follows:

$$c_B(v) = \frac{1}{n_B} \sum_{s,t \in V} \frac{\sigma_{st}(v)}{\sigma_{s,t}}$$

where

- $n_B = n(n-1)$ such that n is the number of nodes of a graph.
- $\sigma_{st}(v)$ denotes the number of shortest paths from s to t containing v.
- $\sigma_{s,t}$ denotes the number of shortest paths from s to t.

Closeness Centrality

Closeness centrality is based on the .

Closeness centrality is defined as:

$$c_C(v) = \frac{n_C}{\sum_{t \neq v} d_G(v,t)}$$

where

- $n_C = n - 1$.
- $d_G(v,t)$ denotes the length of a shortest path between v and t.

Graph as an Electrical Network

We may view the graph as an electrical network by corresponding the weight of each edge as the conductance. Since we are working with unweighted graphs, the conductance of each edge is 1.

The supply vector $b : V \to \mathbb{R}$ defines where current externally enters and leaves the network. For our case, if node s is the source and node t is the sink of the current, then we have

$$b_{st}(v) = \begin{cases} 1, & \text{if } v = s \\ -1, & \text{if } v = t \\ 0 & \text{otherwise} \end{cases}$$

Each edge is given an arbitrary orientation to account for direction of flow. The vector $x : \vec{E} \to \mathbb{R}$ is defined to be the current of the network.

Graph as an Electrical Network

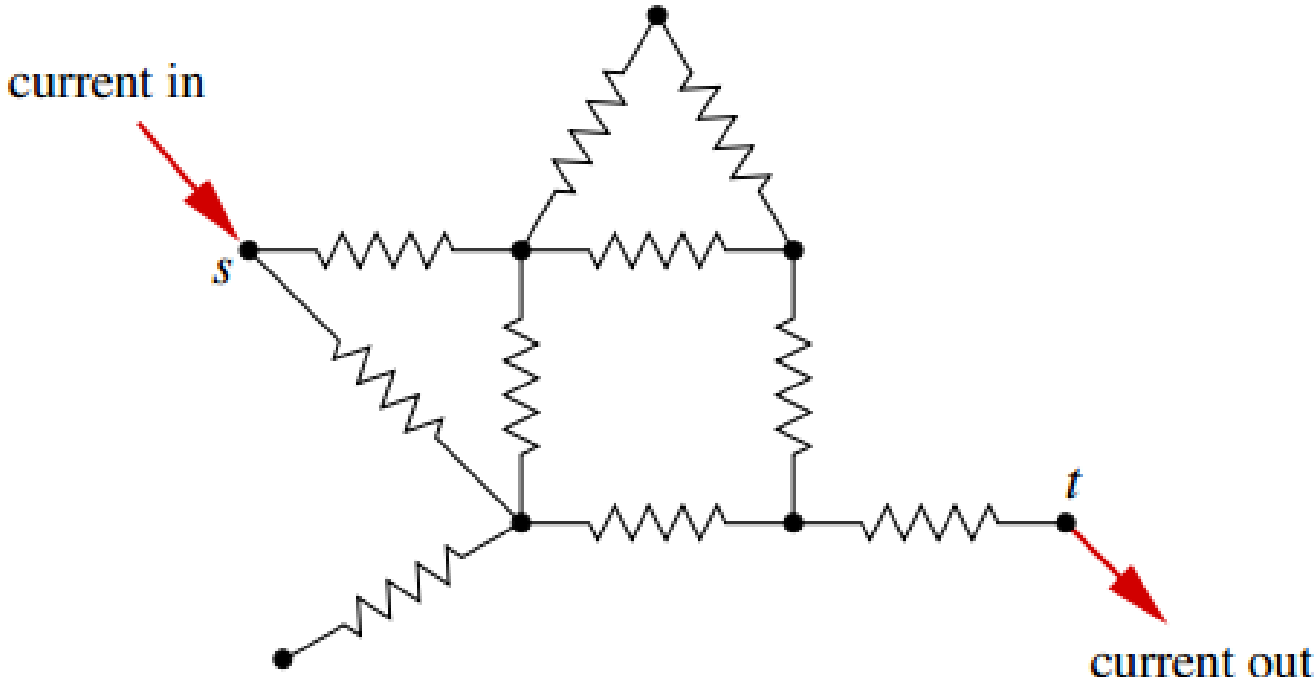

Figure: A simple electrical circuit

Graph as an Electrical Network

Currents are related to voltages or potential differences $\hat{p} : \vec{E} \to \mathbb{R}$ by $\hat{p}(\vec{e}) = x(\vec{e})$ for all $e \in E$, so the absolute potential of $p : V \to \mathbb{R}$ is assigned when $\hat{p}(v, w) = p(v) - p(w)$ for all $(v, w) \in \vec{E}$.

The absolute potentials can be computed directly by the system

$$Lp = b$$

where $L = D - A$ as the Laplacian matrix.

Since L is singular, we fix $p(v_1) = 0$ so that we have the following:

$$p = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \widetilde{L}^{-1} \end{pmatrix} b$$

where $\widetilde{L} \in \mathbb{R}^{n-1 \times n-1}$ is the matrix obtained from L by omitting the row and column of v_1.

Current-Flow Betweenness Centrality

In an electrical network, the analog of the fraction of shortest *st*-paths passing through a node is the fraction of a unit *st*-current flowing through it.

Given a supply b under a *st*-current, the throughput of a node $v \in V$ is defined as

$$\tau_{st}(v) = \frac{1}{2}\left(-|b(v)| + \sum_{e:v\in e} \left|x(\vec{e})\right|\right).$$

The current-flow betweenness centrality is defined as

$$c_{CB}(v) = \frac{1}{n_B} \sum_{s,t\in V} \tau_{st}(v)$$

where $n_B = (n-1)(n-2)$.

Current-Flow Closeness Centrality

To introduce the variant of closeness centrality on an electrical network, the current-flow closeness centrality is defined as follows:

$$c_{CC}(s) = \frac{n_C}{\sum_{t \neq s} p_{st}(s) - p_{st}(t)}$$

where $n_C = n - 1$.

Here $p_{st}(s) - p_{st}(t)$ is the effective resistance, which is an alternative measure of distance between s and t. We need an efficient way of computing it between each pair of nodes.

Effective Resistance

Spielman-Srivastava proposed the following method to approximate effective resistances:

1. Solve the systems $LX = B^T Q$
2. Effective resistance between nodes s and t is equal to $\|X_s - X_j\|_2$ where X_i is the ith row of X.

where

- L is the Laplacian
- B is the incident matrix
- Q is the random Johnson-Lindenstrauss projection of size $m \times O(\log n)$.

X is called the effective resistance embedding since each row corresponds to a node embedded in $n \times O(\log n)$-dimensional Euclidean space. We can apply k-means to it.

k-means Clustering

k-means clustering is the most basic but most widely used clustering technique. It partitions a dataset into k clusters $\{S_1, S_2, ..., S_k\}$.

The algorithm is as follows:

1. Randomly pick k data points and initialize them as centroids.
2. Calculate the distance between each data point to each centroid.
3. Assign each data point to the closet centroid.
4. Compute new centroids by calculating the mean of the data points in each cluster.
5. Repeat Steps 2-4 until convergence.

k-Means Centrality

We propose a new centrality measure based on applying k-means to effective resistance embedding. It is defined as the following:

$$c_K(s) = \sum_{k=2}^{L} \sum_{t \in V} \mathbb{1}_{\{\text{edge } st \text{ is a cut edge in a } k\text{-clustering on } \{S_1^{(k)}, S_2^{(k)}, ..., S_k^{(k)}\}\}}.$$

The cut edge is defined as an edge whose endpoints lie in two different partition sets and $\mathbb{1}$ is the indicator function.

Outline

Numerical Experiments on Synthetic Data

- We implement the centrality measures on two toy data sets.

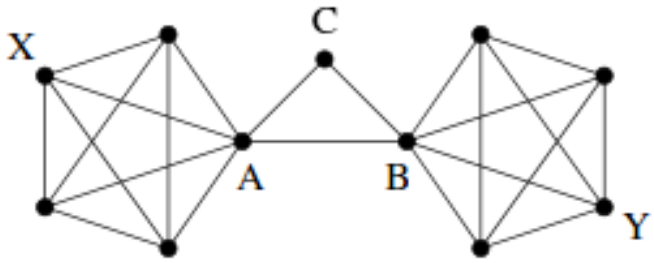

Figure: Toy Network

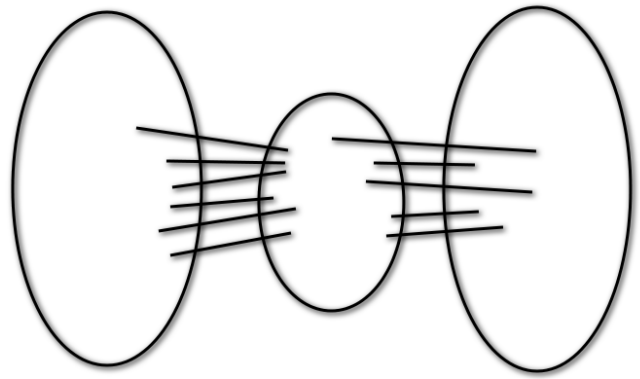

Figure: ”Tripartite” Graph

- For the following graphs, note that nodes with higher centrality measures are larger and are closer to violet on the spectrum and nodes with lower centrality measures are smaller and are closer to red on the spectrum.

Numerical Experiments on Synthetic Data

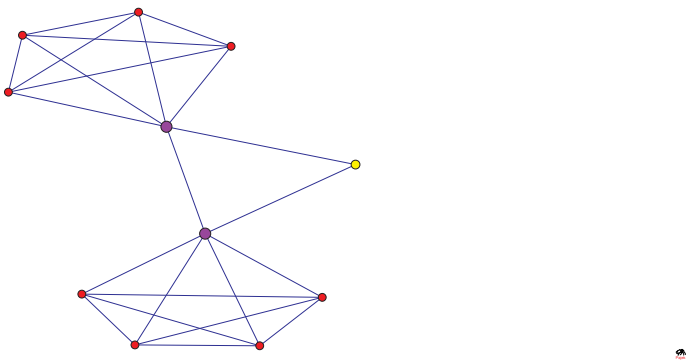

Figure: Closeness Centrality of Toy Network

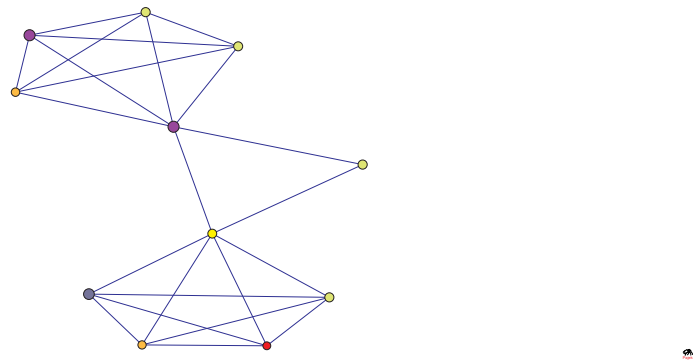

Figure: Current-Flow Closeness Centrality of Toy Network

Numerical Experiments on Synthetic Data

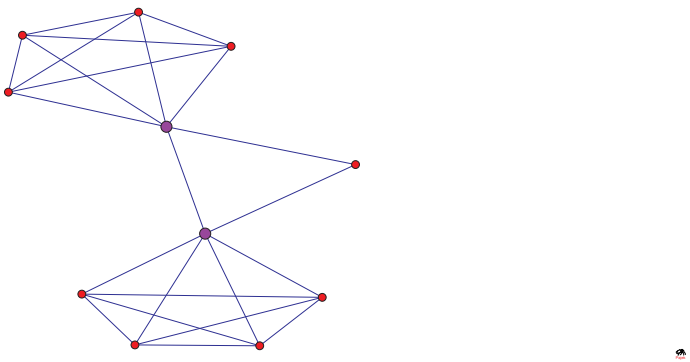

Figure: Betweenness Centrality of Toy Network

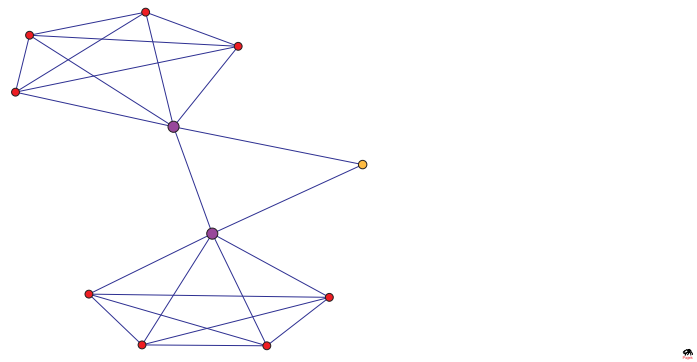

Figure: Current-Flow Betweenness Centrality of Toy Network

Numerical Experiments on Synthetic Data

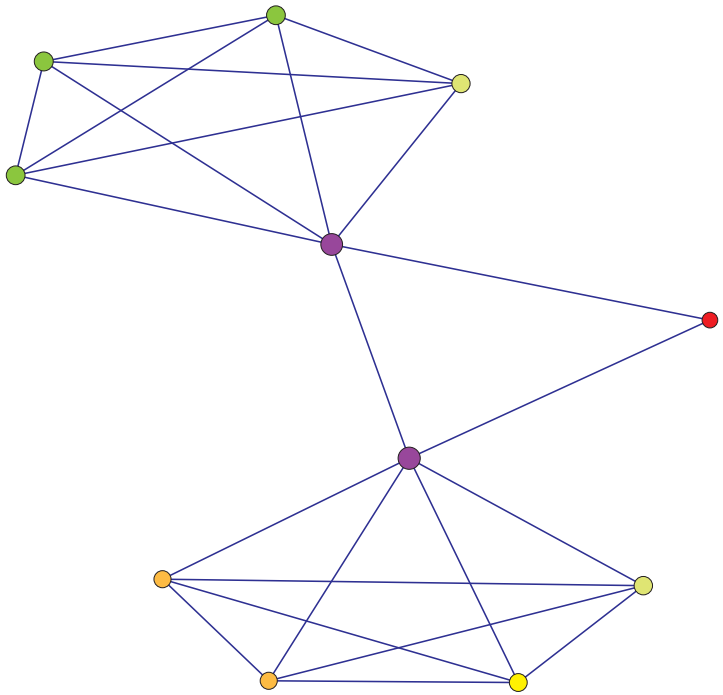

Figure: *k*-Means Centrality of Toy Network

Numerical Experiments on Synthetic Data

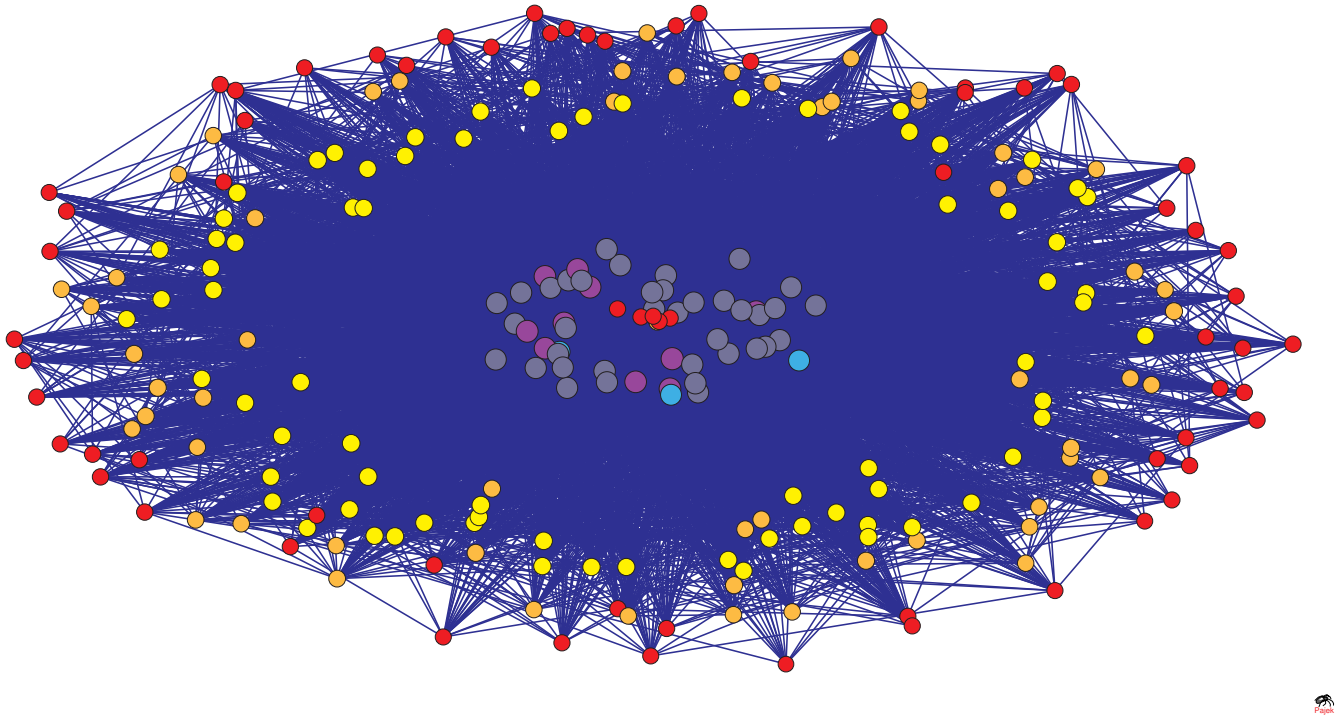

Figure: Closeness Centrality of ”Tripartite” Graph

Numerical Experiments on Synthetic Data

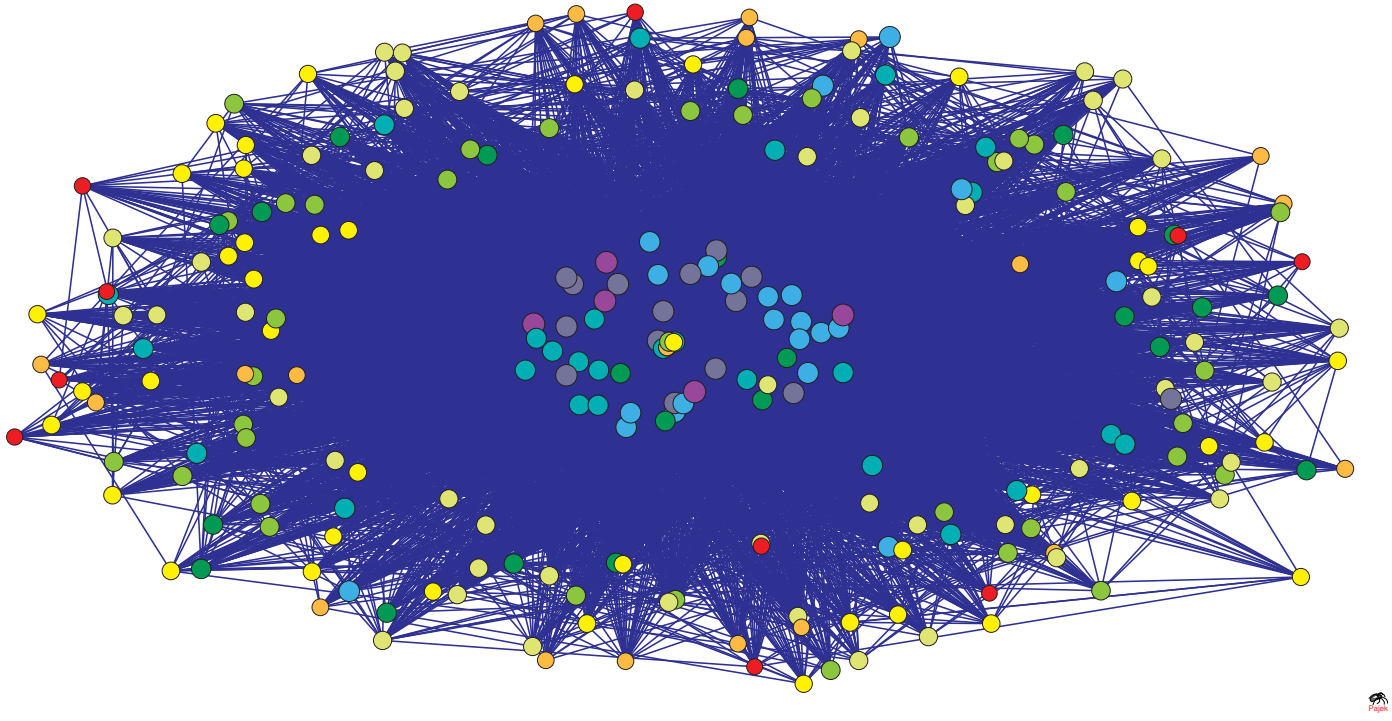

Figure: Current-Flow Closeness Centrality of ”Tripartite” Graph

Numerical Experiments on Synthetic Data

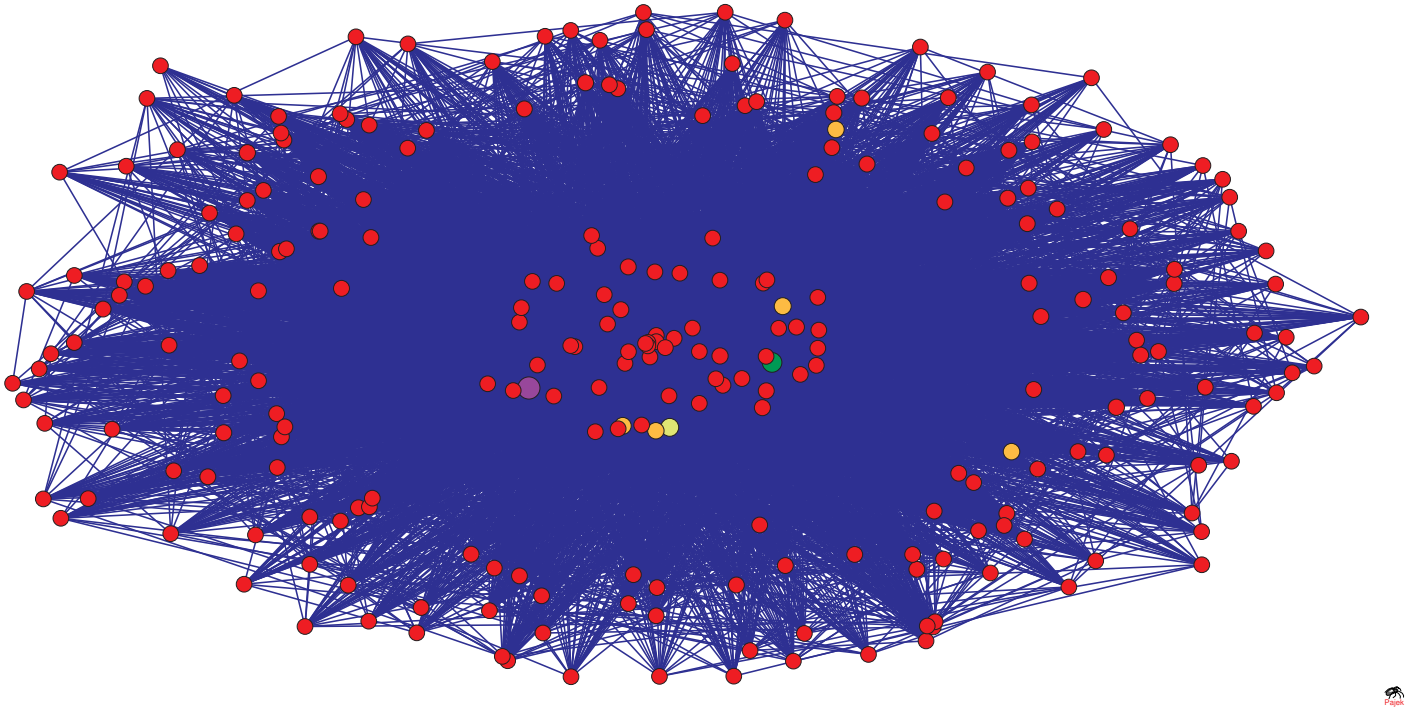

Figure: Betweenness Centrality of ”Tripartite” Graph

Numerical Experiments on Synthetic Data

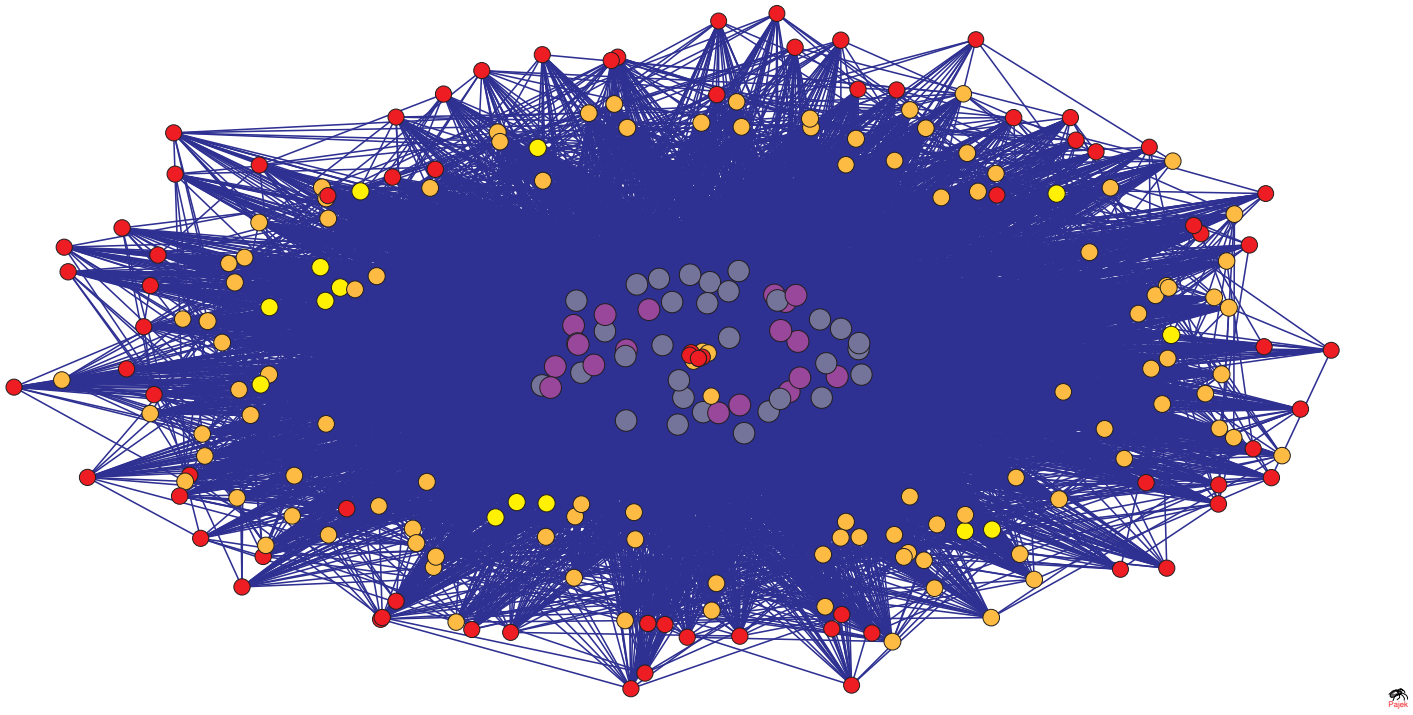

Figure: Current-Flow Betweenness Centrality of ”Tripartite” Graph

Numerical Experiments on Synthetic Data

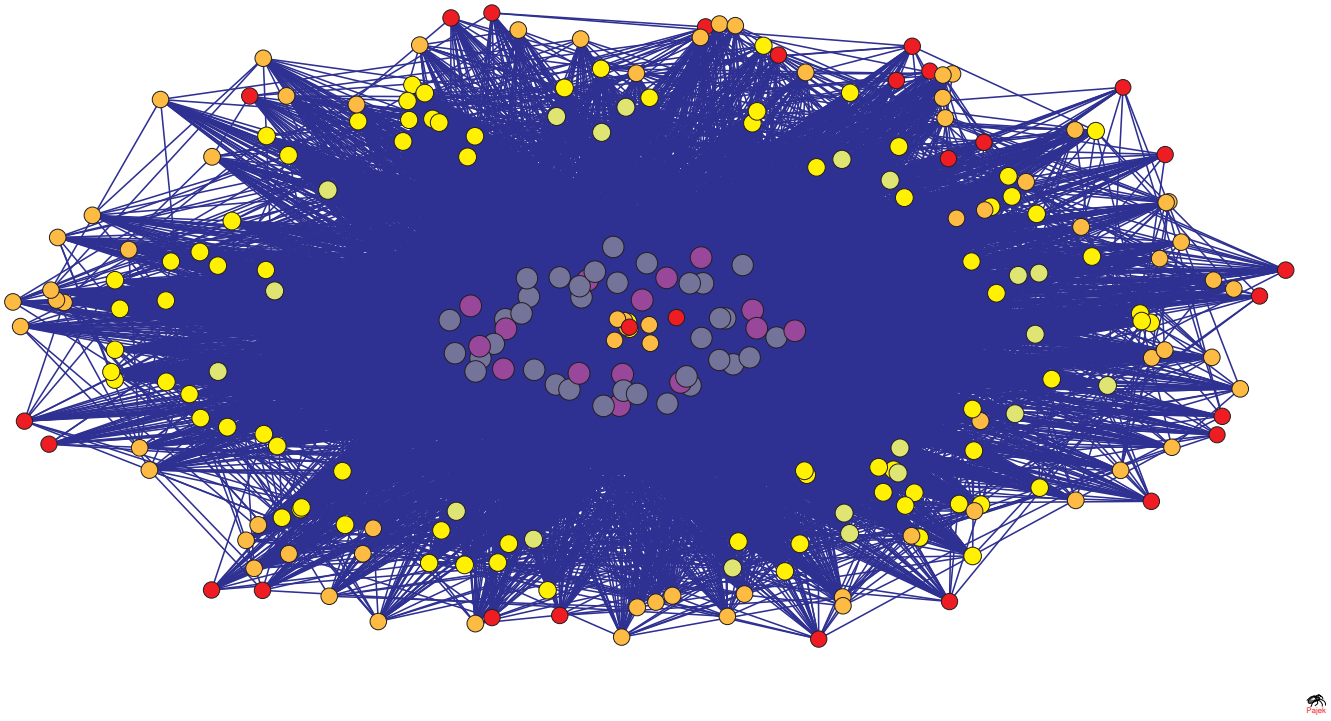

Figure: *k*-Means Centrality of ”Tripartite” Graph

Outline

Numerical Experiments on Real Data

- We compute the centrality measures on the Florentine network, consisting of the most influential families of Florentine, Italy, during the 15th century, and the social network of Reed College.

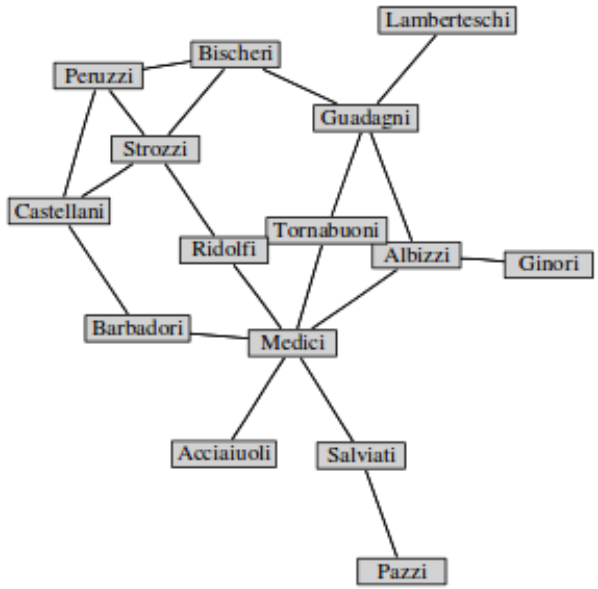

Figure: Florentine Graph

Numerical Experiments on Real Data

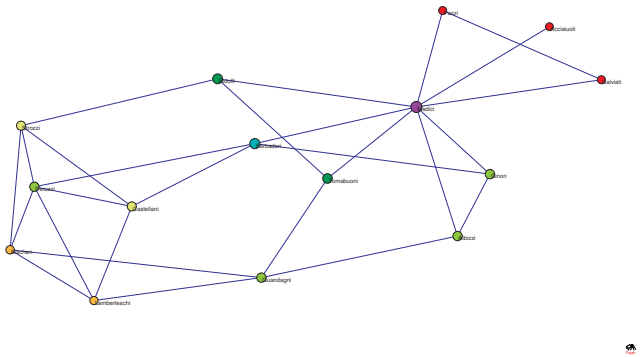

Figure: Closeness Centrality of Florentine Network

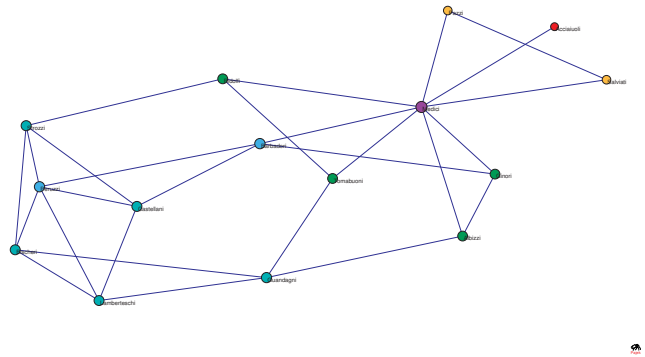

Figure: Current-Flow Closeness Centrality of Florentine Network

Numerical Experiments on Real Data

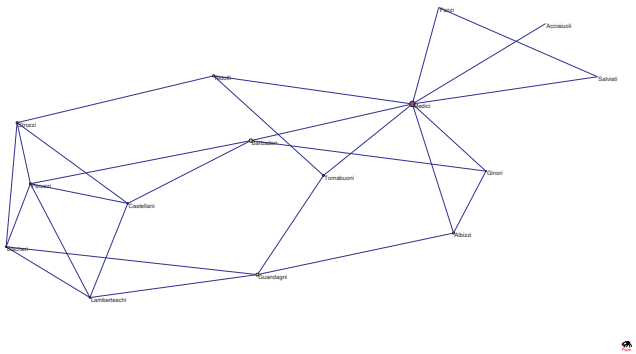

Figure: Betweenness Centrality of Florentine Network

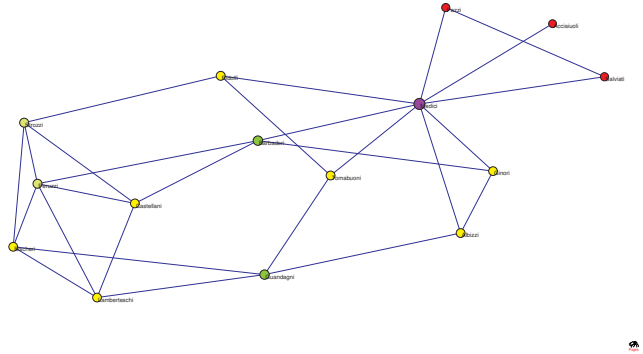

Figure: Current-Flow Betweenness Centrality of Florentine Network

Numerical Experiments on Real Data

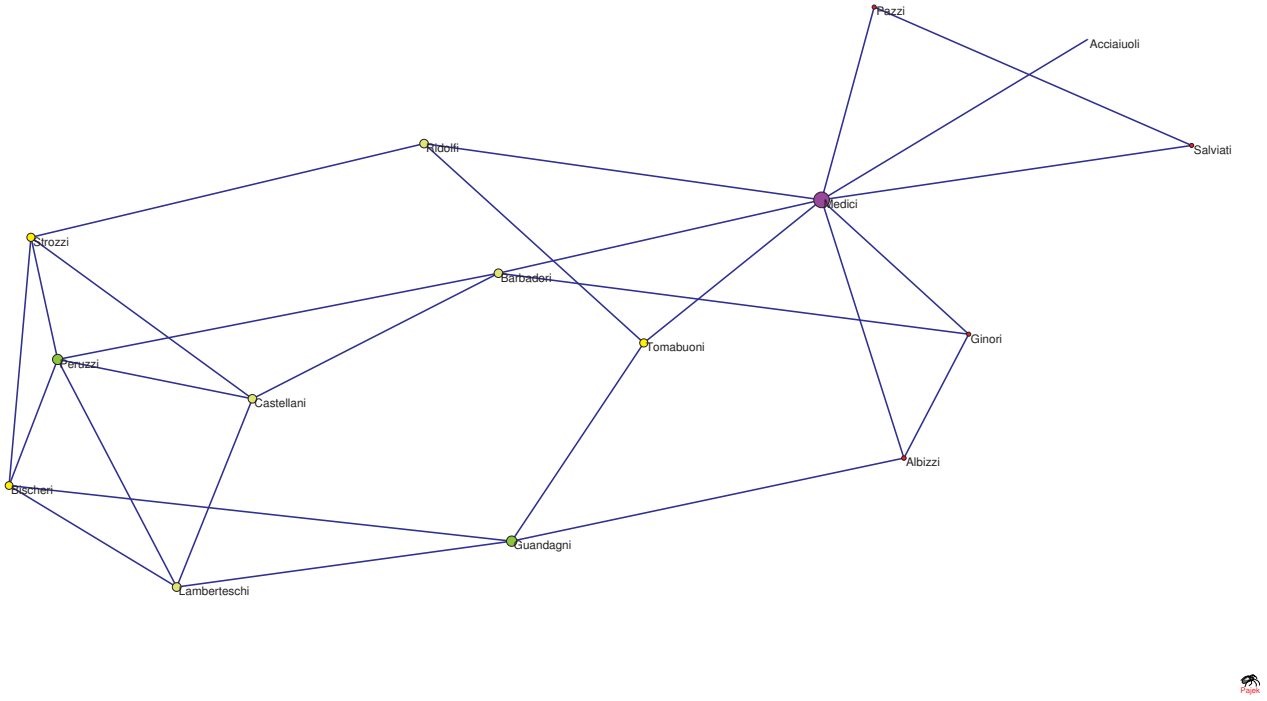

Figure: *k*-Means Centrality of Florentine Network

Numerical Experiments on Real Data

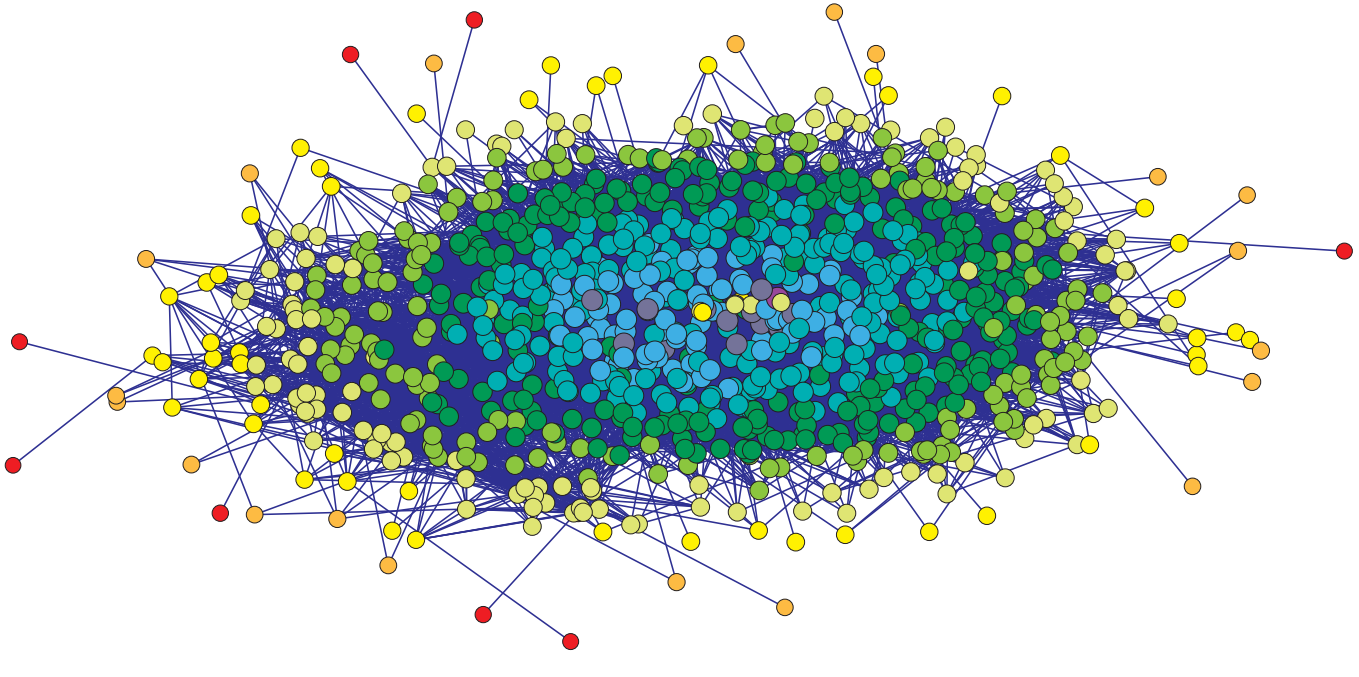

Figure: Closeness Centrality of Reed Network

Numerical Experiments on Real Data

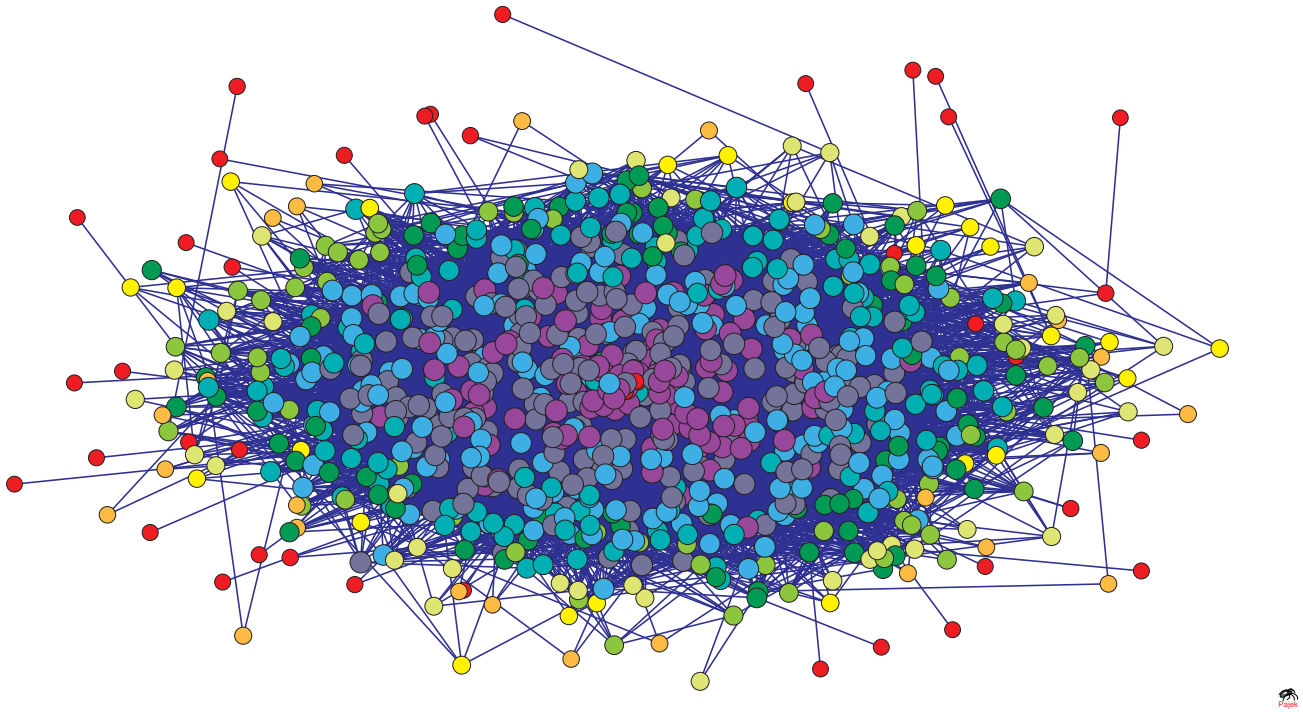

Figure: Current-Flow Closeness Centrality of Reed Network

Numerical Experiments on Real Data

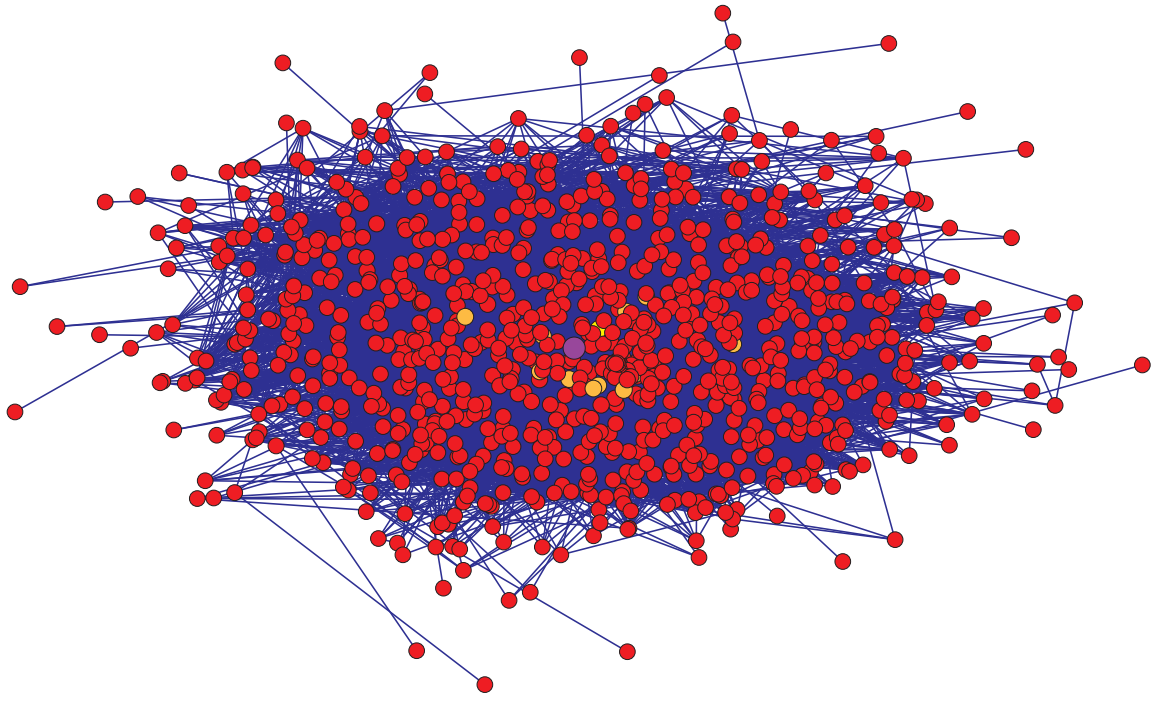

Figure: Betweenness Centrality of Reed Network

Numerical Experiments on Real Data

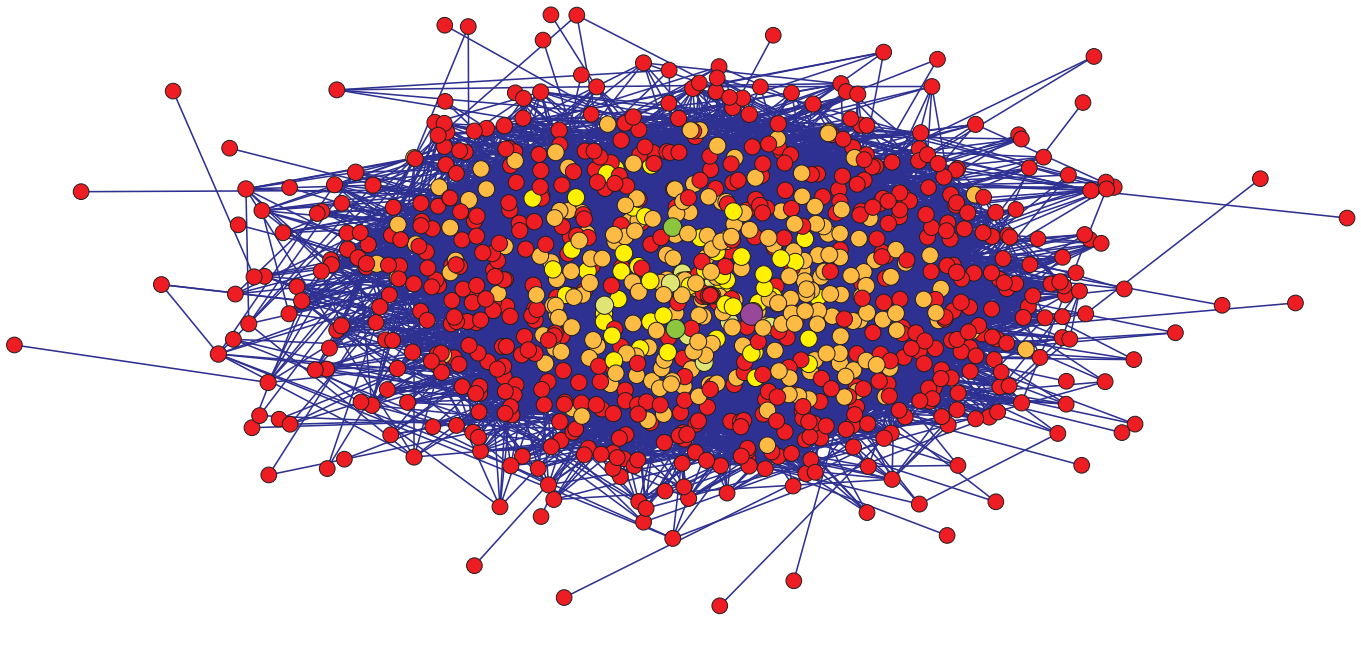

Figure: Current-Flow Betweenness Centrality of Reed Network

Numerical Experiments on Real Data

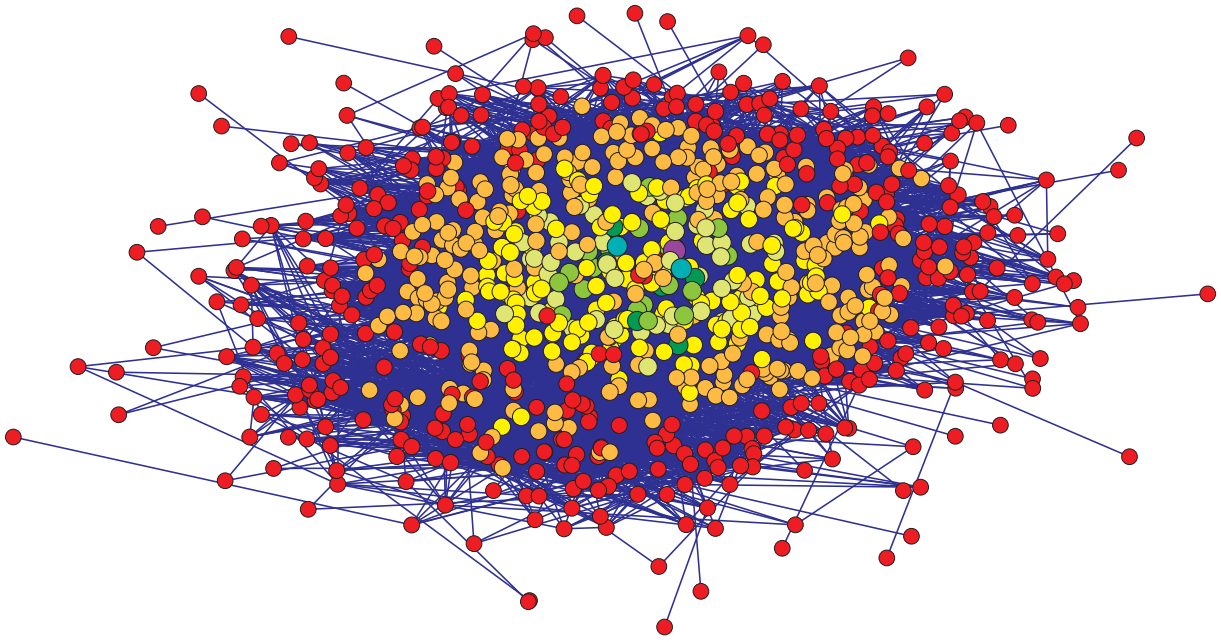

Figure: *k*-Means Centrality of Reed Network

Outline

Discriminative Measure

We assign a discriminative measure δ to each centrality measure:

$$\delta = \sum_{i=1}^{n} \left[\frac{v}{\| v \|} \right]_i, \qquad \textit{where} \quad v = \frac{\mathbb{D} \cdot \mathbb{1}}{n}$$

$\mathbb{D}$ is the distance matrix whose ij-entry is the Euclidean norm of the centrality measures of the pair of nodes i and j.

Larger discrimination measure better distinguishes different node properties

Comparison of Discriminative Measure

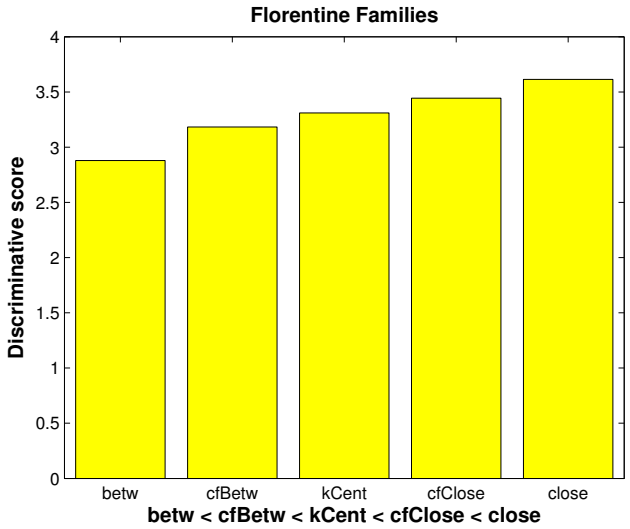

Figure: Discriminative Measures of Florentine Families Network

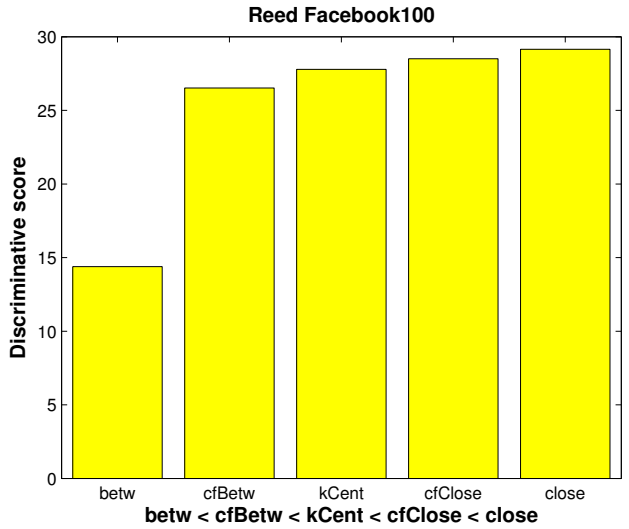

Figure: Discriminative Measures of Reed Facebook Network

Correlation Between Centrality Measures and Node Degree

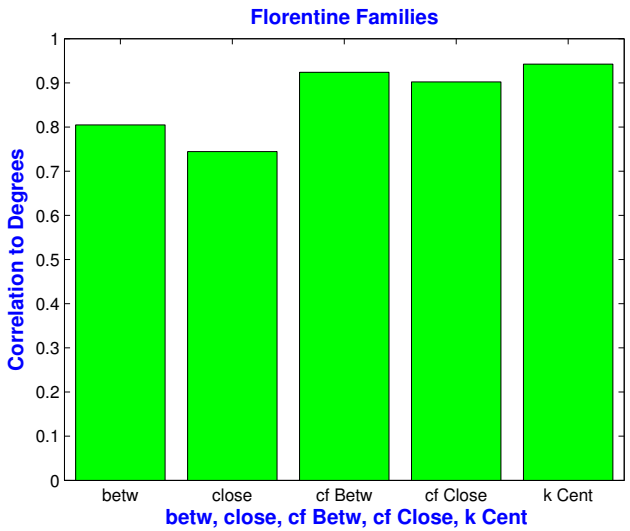

Figure: Correlation for Florentine Families Network

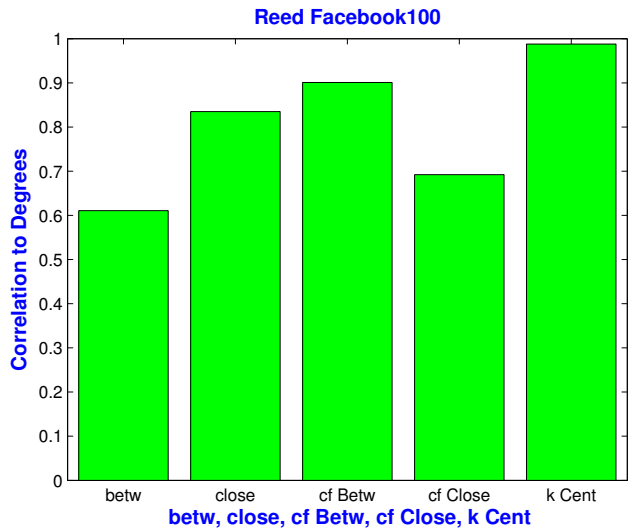

Figure: Correlation for Reed Facebook Network

Two-Cluster Separation for the ”Tripartite” Graph

We analyze how effective each measures are by trying to recover two clusters from the ”tripartite” graph.

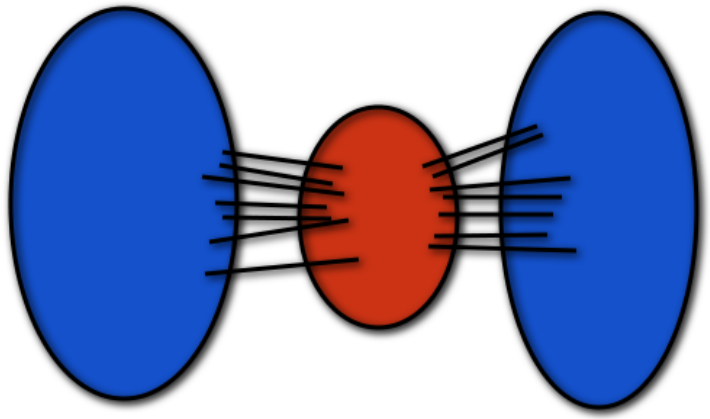

Figure: Two Clusters of the ”Tripartite Graph”

We have 200 nodes in blue and 50 nodes in red. We also added some noise, that is some edges between the two separate blue groups. k-means is applied to partition the vector of centrality measures into two clusters.

Partitioning on Closeness and Betweenness Centralities

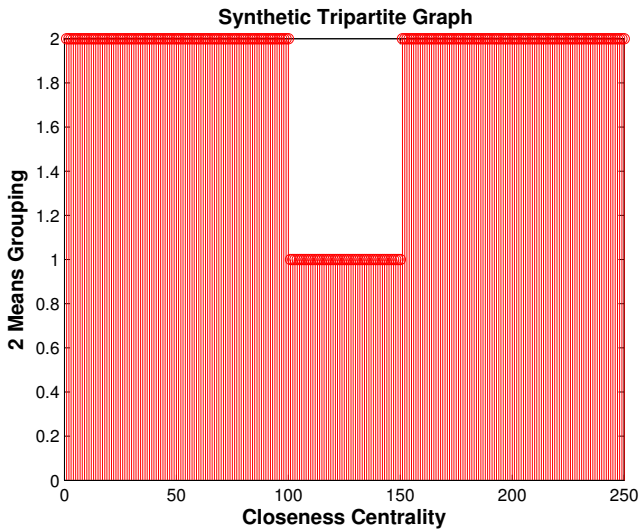

Figure: Closeness Centrality Partitioned into 2 Clusters

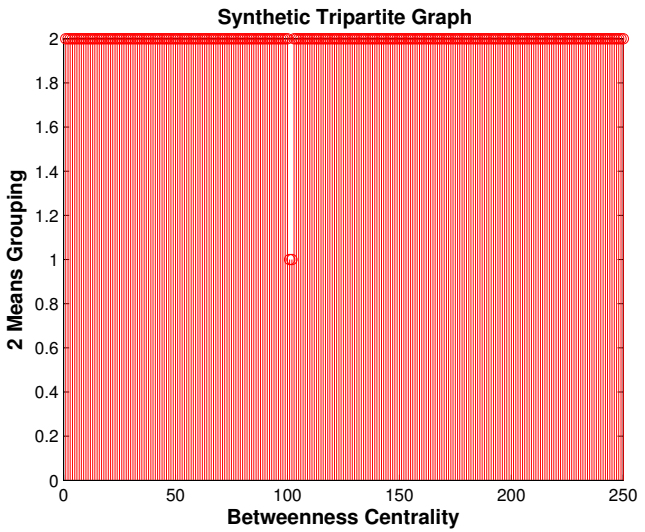

Figure: Betweenness Centrality Partitioned into 2 Clusters

Partitioning on Current-Flow Closeness and Current-Flow Betweenness Centralities

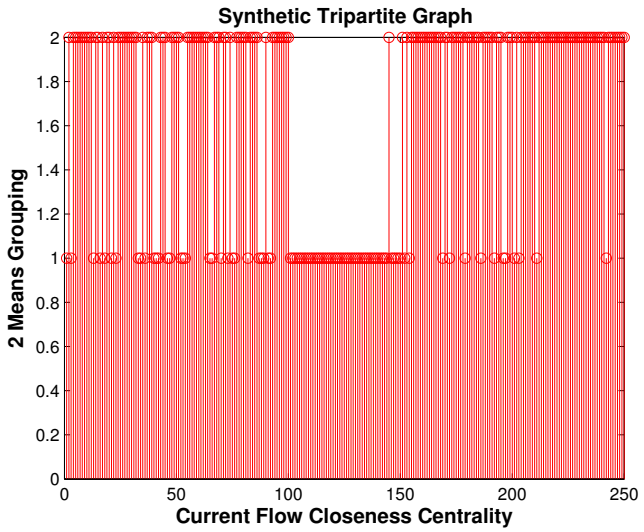

Figure: Current-Flow Closeness Centrality Partitioned into 2 Clusters

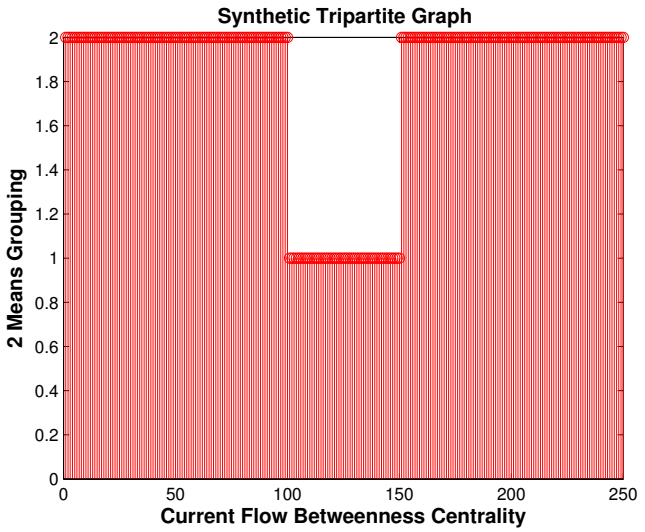

Figure: Current-Flow Betweenness Centrality Partitioned into 2 Clusters

Partitioning on *k*-means Centrality

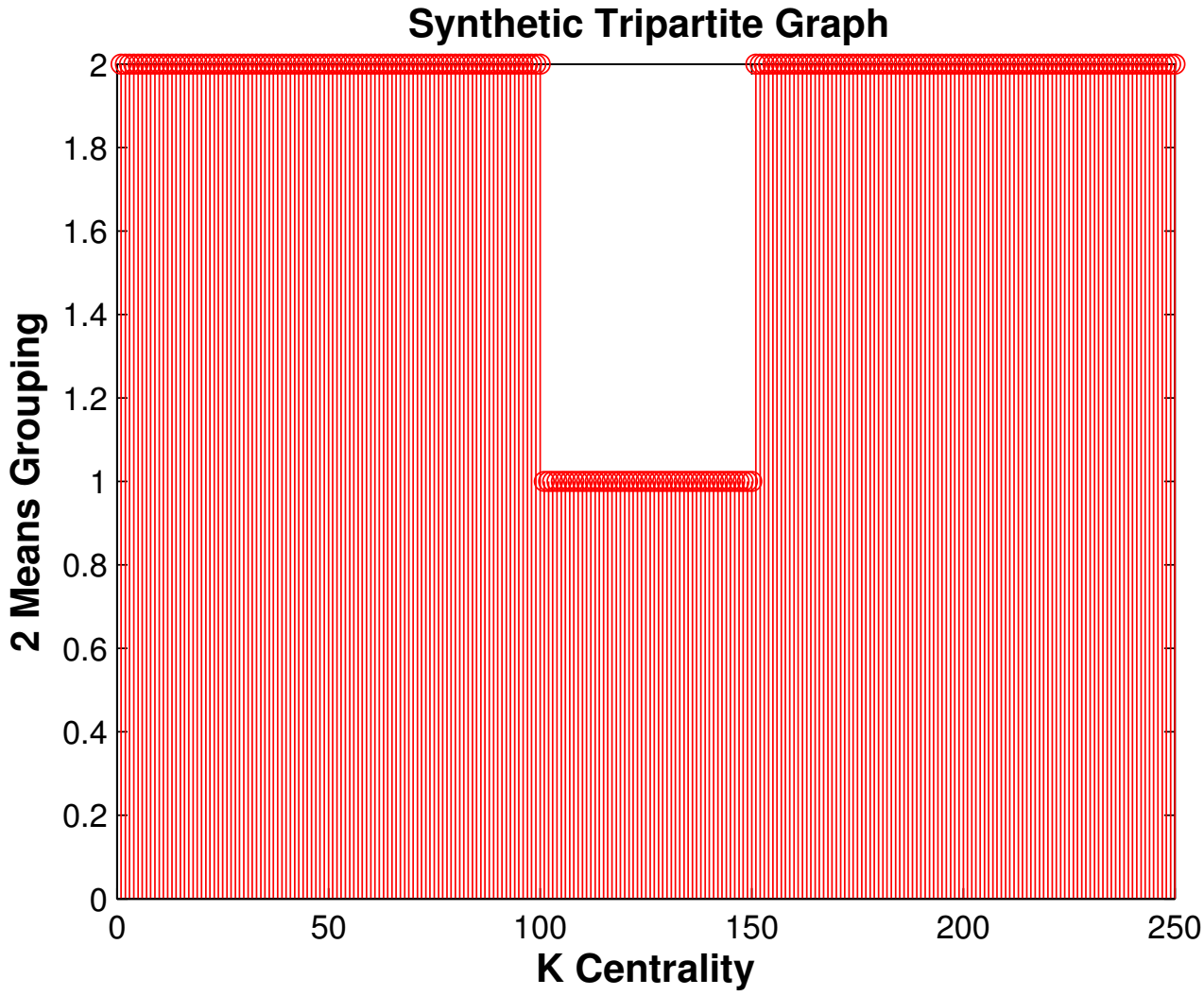

Figure: *k*-means Centrality Partitioned into 2 Clusters

Average Distance Test on "Tripartite Graph"

- We sort the vector of centrality measures in descending order.
- We divide the vector into two groups of size p and q such that $p + q = n$ and compute the distance matrices D_1 and D_2 of the two groups.
- We plot $\lambda = \frac{\sigma_1 + \sigma_2}{2}$ for each cut point p where

$$\sigma_1 = \frac{\sum_{k=1}^{p} [D_1 \cdot \mathbb{1}]_k}{\binom{p}{2}}, \quad \sigma_2 = \frac{\sum_{k=1}^{q} [D_2 \cdot \mathbb{1}]_k}{\binom{q}{2}}$$

- λ is expected to obtain its minimum at $p = 50$ since that is where we divide the nodes with high centrality measures and nodes with low centrality measures into their own groups.

Average Distance Test for Closeness and Betweenness Centralities

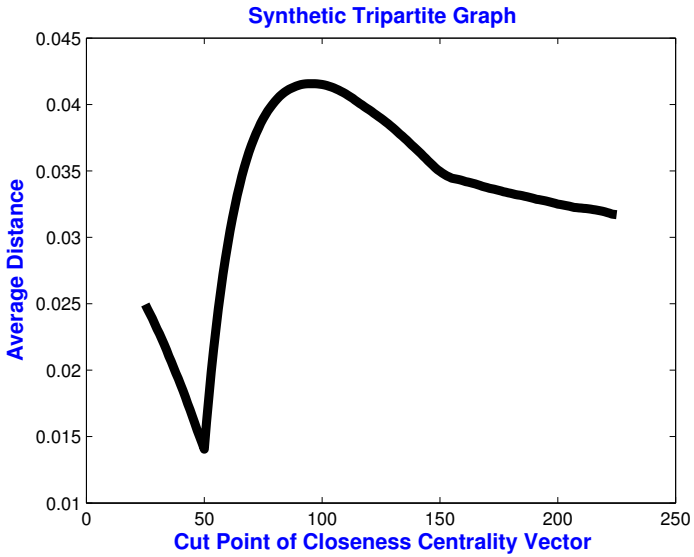

Figure: Average Distance for Closeness Centrality

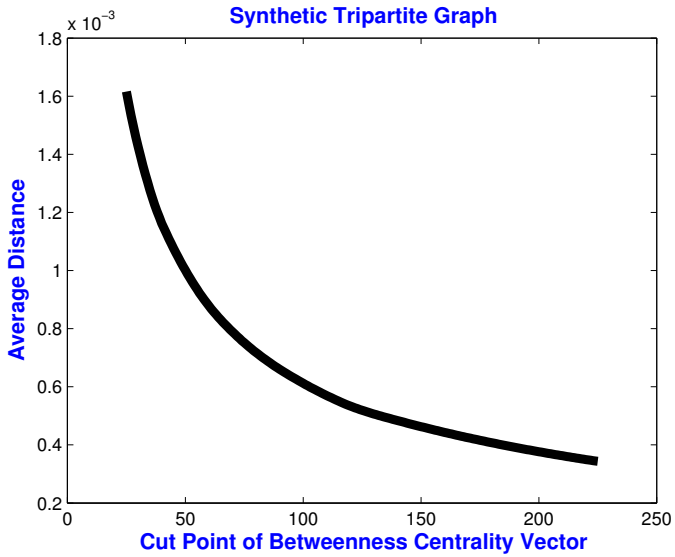

Figure: Average Distance for Betweenness Centrality

Average Distance Test for Current-Flow Closeness and Current-Flow Betweenness Centralities

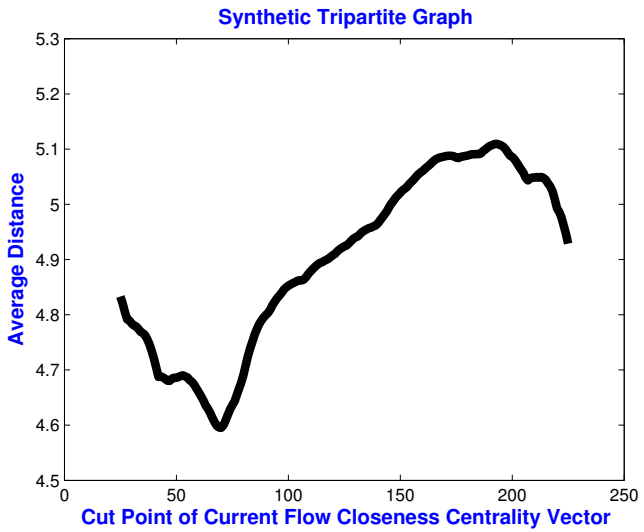

Figure: Average Distance for Current-Flow Closeness

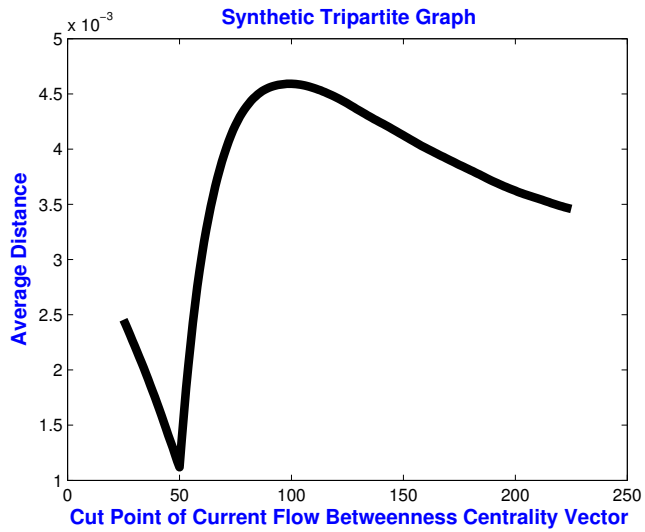

Figure: Average Distance for Current-Flow Betweenness

Average Distance Test for k-Means Centrality

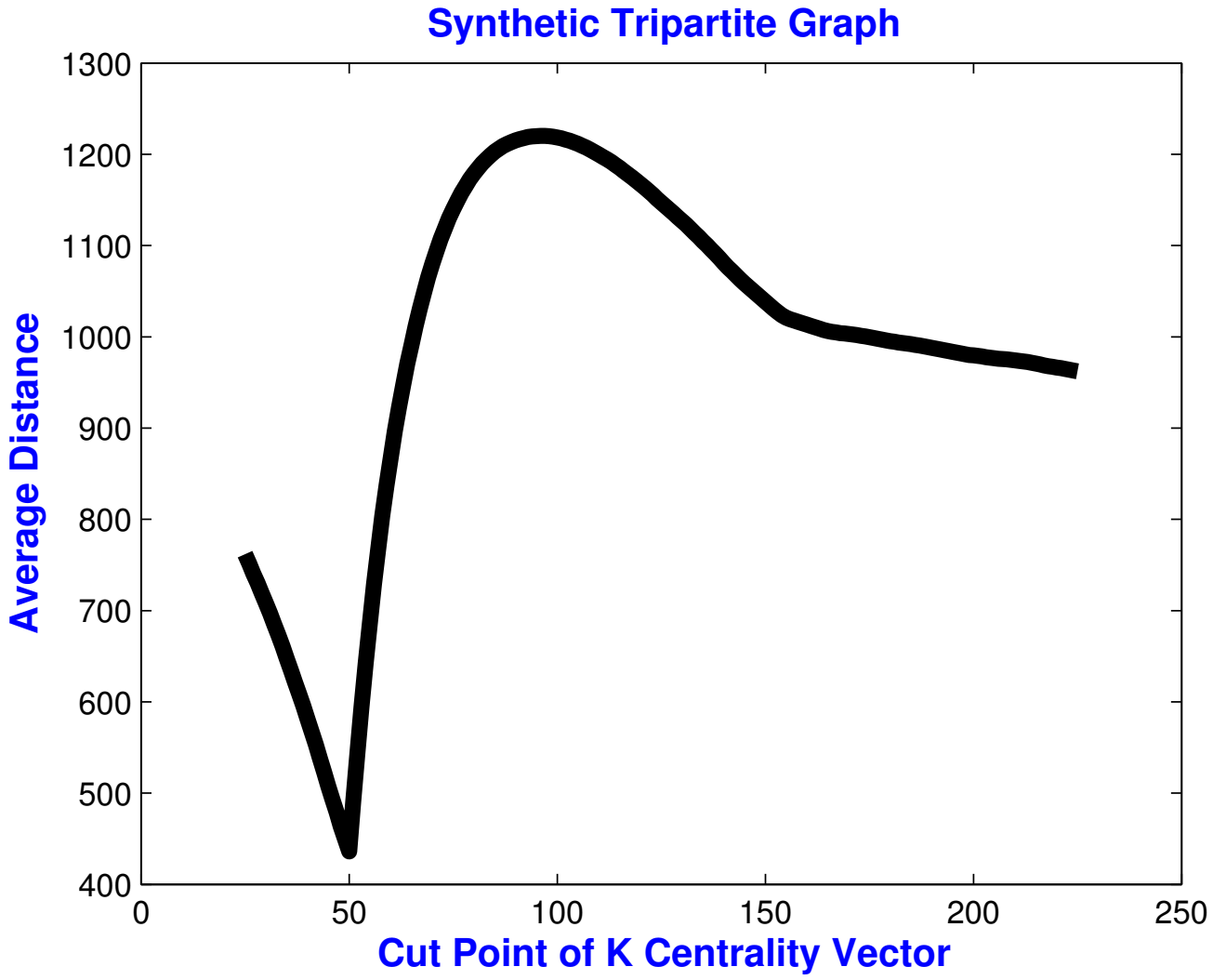

Figure: Average Distance for k-means Centrality

Outline

Summary of Our Work

- *k*-means centrality ranks in the middle in terms of discriminative measure since it measures less than closeness and current-flow closeness but more than betweenness and current-flow betweenness.
- *k*-means centrality has the highest correlation to node degree among the five centrality measures.
- Closeness, current-flow betweenness, and *k*-means centrality successfully partitioned the ”tripartite” graph into two clusters.
- *k*-means centrality is able to discriminate nodes with high centrality measures from nodes with low centrality measures through the average distance test.