CS 446 – Final Project

BETWEENNESS CENTRALITY CALCULATIONS: USING GEODESIC PATHS VERSUS RANDOM WALK APPROACHES

Michael Kupperman and Rachel Sousa

ABSTRACT

The betweenness centrality of a biological network is important in determining the relative necessity of nodes within the graph topology by considering paths traversing the graph. Here, we examine and compare different notions of betweenness centrality of a transcription–factor–to–target-gene interaction network in the model species, *E. coli*. We contrast two approaches to the betweenness centrality: one considering the geodesic paths between nodes that the node of interest lies on and a biased random walks where the flow of information is not optimal.

INTRODUCTION

Betweenness centrality measures the extent to which a given node lies on a path between other nodes. Here, we inspect the betweenness centrality of an *E. coli* gene regulatory network from RegulonDB, a database at Universidad Nacional Autonóma de México (UNAM), using two contrasting methods.

The first method is the betweenness centrality described by Newman in his book *Networks: An Introduction*. Newman describes the betweenness centrality as the number of geodesic paths that a node lies on. Mathematically, Newman's betweenness centrality of node i is

$$x_i = \sum_{s,t} n_{s,t}^i$$

where n_{st}^i is an indicator function if node i lies on the geodesic path from s to t.

The second method is to calculate betweenness centrality using a biased random walk towards geodesic paths. Propagation of information through a biological system is typically imperfect, characterized by environmental noise and a reliance on chemical information rather than conscious agents. A random walk algorithm is implemented to explore different path options given that the node is biased in going towards a geodesic path.

METHODS

- The interaction network was simplified by filtering the network to include only target genes that are themselves transcription factors with at least one outgoing edge in the network and by eliminating self-edges.
- Newman's betweenness centrality of the graph was computed using the function *betweenness* method in *igraph*.
- We developed our own code to calculate the biased random walk towards geodesic path.
- A bias term β was set to describe the probability of selecting a shortest path. β values included: 0.5, 0.6, 0.7, 0.8, 0.9, 0.99.
- Then we iterate over all pairs of source and target nodes and return the betweenness centrality as a list of incidence weights for all vertices hit on the many random walks, divided by the total number of paths sampled.
- We compared the differences of the betweenness centrality of Newman's method and of the random walk method, with each value of β, to ensure that the biased walk method was stable (**Figure 1** and **Figure 3**).
- To compare the geodesic and the biased random walk betweenness centrality methods, we computed the linear regression of the centrality values. In **Figure 2**, the linear regression is somewhat linear, but there is noise distorting it.

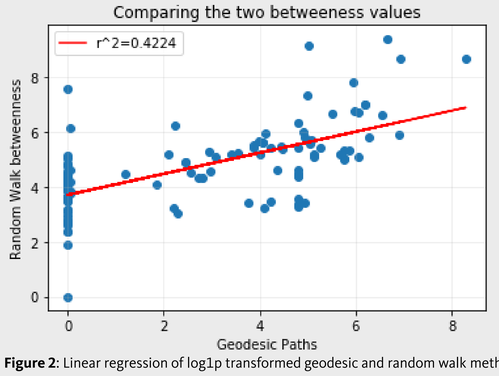

Figure 2: Linear regression of log1p transformed geodesic and random walk methods.

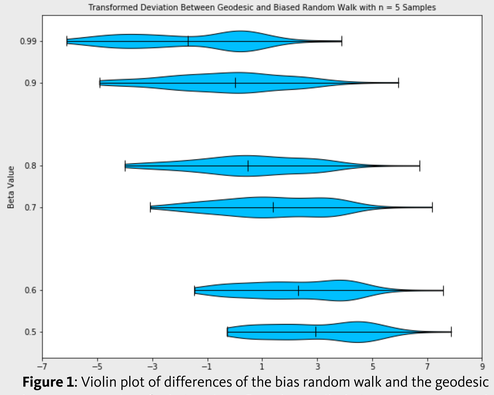

Figure 1: Violin plot of differences of the bias random walk and the geodesic betweenness values for beta values. 5 random walks between every two nodes.

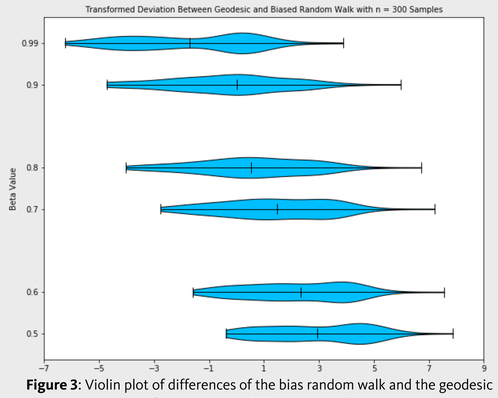

Figure 3: Violin plot of differences of the bias random walk and the geodesic betweenness values for beta values. 300 random walks between every two nodes.

RESULTS

- We looked at the transformed deviations of the geodesic and the random walk betweenness centralities to determine the stability of our method via the number of random walks between every two nodes. Looking at **Figure 1** and **Figure 3**, we see that the biased random walk method is stable since the violin plots of each corresponding beta value are similar in shape, distribution, and median value.
- In the geodesic betweenness centrality histogram (**Figure 5**), we see that the most abundant betweenness centrality value is zero. However, in the biased random walk histogram (**Figure 7**), there are no centrality values of zero. We also see this trend in the scatter plots (**Figure 4** and **Figure 6**).
- This difference is due to the differing methods of calculations for the betweenness centrality. In the geodesic method, leaf nodes (vertex degree of 1) have betweenness of 0 as they do not lie on any geodesic path (proof left to the reader). With a random walk betweenness, there is a β^n probability of not walking through a leaf node when starting at the neighboring node. Then for n larger or β small, the average betweenness for the leaf nodes is nonzero.

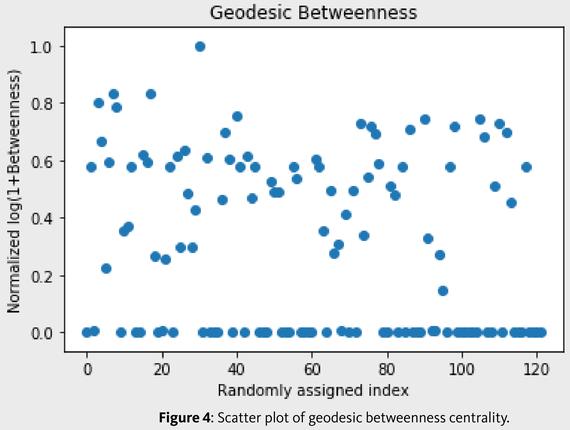

Figure 4: Scatter plot of geodesic betweenness centrality.

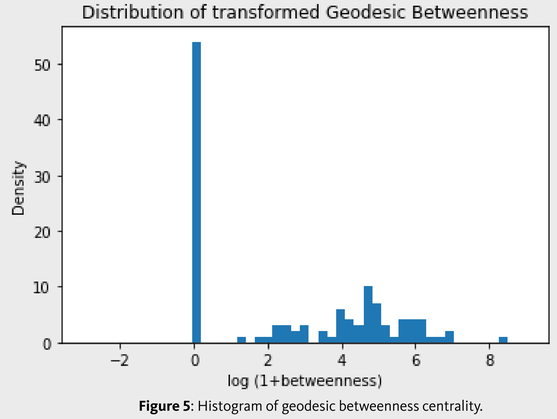

Figure 5: Histogram of geodesic betweenness centrality.

CONCLUSIONS

- There are several pros and cons of our biased random walk betweenness centrality calculation method.
- One pro of this method is that the random walk betweenness centrality returns only nonzero values eventually. That is, as n (the number of random walks sampled between pairs of points) becomes large, every node is eventually reached, removing zero values from the centrality calculation.
- Geodesic betweenness centrality is not a good predictor of biased random walk betweenness centrality. This was shown in **Figure 2.**
- The violin plots (**Figure 1** and **Figure 3**) suggest that the random walk betweenness centrality method is stable for any value of beta for any number of random walks greater than or equal to five.

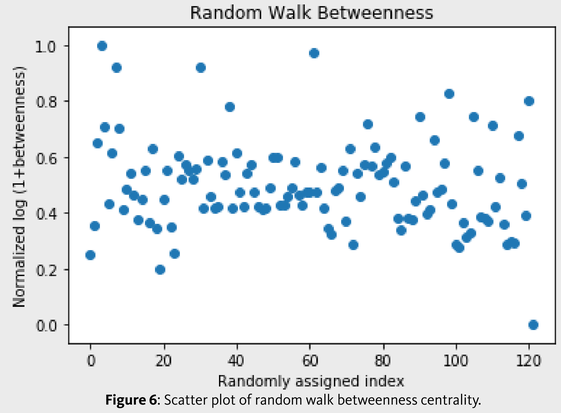

Figure 6: Scatter plot of random walk betweenness centrality.

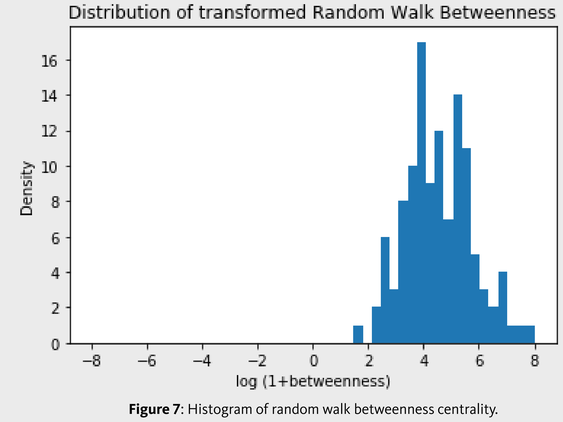

Figure 7: Histogram of random walk betweenness centrality.

REFERENCES

[1] Newman M. Networks: An Introduction. *Oxford University Press*, 2010.