Research Statement

Christina Durón

As an applied mathematician, my overall research interests lie in network analysis and network theory. Specifically, I focus on the development of new computational techniques to model, analyze, and explore relational data from a variety of fields (e.g., biological, social, transportation). As a graduate student at Claremont Graduate University, I co-developed a robust methodology that established the betweenness centrality network analysis as a valuable tool for identifying central genes unique to the tumor ecosystem. As a postdoctoral research scholar at the University of Arizona, I proposed a new shortest-path based centrality measure to identify super-spreader nodes in real-world networks. Additionally, I co-developed a parameter fitting procedure that utilizes a discrete time SIR model to estimate SIR epidemic network parameters on Erdös-Rényi networks. While my research has focused predominantly on network analysis involving biological applications, it has expanded to include work in measure theory, partial differential equations, and mathematics education.

1 Published Research

1.1 Combining Betweenness Centrality with Differential Expression

To accurately diagnose diseases and predict therapeutic responses, researchers must understand how genes interact with each other in different environments. One example may include an analysis of differential gene expression data for "healthy" and "diseased" samples that looks to address the question, *"What is the difference between the healthy and diseased groups?"*. Standard t-tests, such as those used to statistically test fold-change values, are problematic in that the analysis fails to incorporate the global structure of the data. While some differential expression analyses use pooled methods for estimating variability, the use of differential expression alone overlooks the important structure between genes.

In [1], the identification of unknown regulatory pathways essential to the maintenance of a tumor was a result of combining network centrality analysis and bioinformatic approaches. In this context, a centrality measure is defined as a function that numerically quantifies the level of influence of each gene (or node). Specifically, the betweenness centrality measure was used to identify *Etv5* as a key regulator in the development of optic glioma, where the betweenness centrality of gene x_i is defined as the fraction of shortest paths in the network that go through gene x_i.

RNA expression data from pediatric brain tumors were used to create two separate weighted networks of identical topological structure, one based upon healthy (or normal) and the other on diseased (or tumor) sets of samples. The weights on the network edges $w_{j,k}$ were given by the distance between two genes x_j and x_k, and were assigned using correlations between the RNA expression levels of each gene.

For each gene, its betweenness centrality value was calculated in both networks (Figure 1). Upon comparing each betweenness measure, it was determined that the role of the regulator *Etv5* had substantially changed. A series of independent experiments were performed to validate *Etv5* as a differentially-expressed tumor-specific gene at the RNA and protein levels. These results provided further evidence that *Etv5* and its associated target genes make up a potential regulatory network (Figure 2) in diseased tissue relative to their healthy counterparts.

1.2 Variability of Betweenness Centrality

Unfortunately, the network centrality and bioinformatic approach taken in [1] only provided an estimate of the true centrality measure of each node. Consequently, follow-up questions such as *"How accurate is this estimate?"* and *"How robust is the statistic to sampling variability?"* motivated a need to develop a more analytic method to gauge

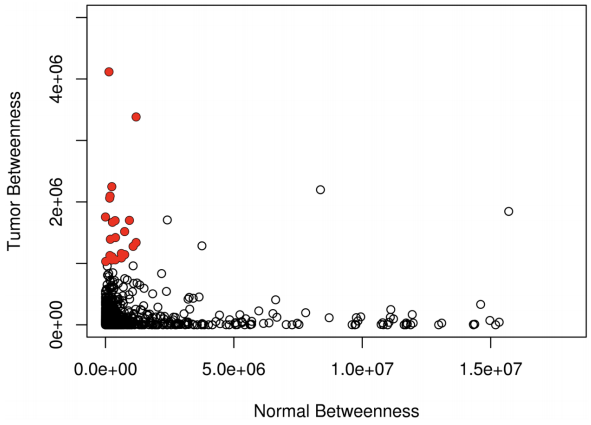

Figure 1: Filled (red) circles indicate genes whose betweenness measure is at least 1.1 times as large in the tumor network as in the normal network and either a tumor betweenness or normal betweenness value greater than 1e6.

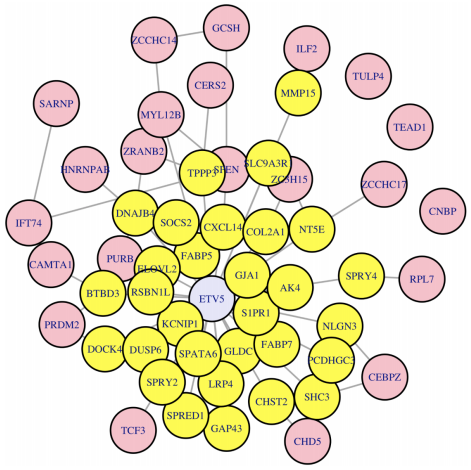

Figure 2: The subnetwork is comprised of *Etv5* (in lavender in the center), and its differentially-expressed targets (shown in yellow). The remaining central genes, identified by their high betweenness measures relative to the normal network, are shown on the periphery in pink.

the robustness of the betweenness centrality as a measure for identifying key regulators in situations when computational validation through experiments are not feasible.

In [2], two separate weighted networks of identical structure were created using each of the normal and tumor sets of samples from pediatric low-grade brain tumors following the procedure outlined in [1]. Then, the betweenness centrality measure was used to identify a set of essentially different genes whose role had substantially changed in the comparison of healthy to diseased states. More specifically, a gene was *essentially different* if it had a tumor betweenness value that was greater than $950,000$ and a tumor to control betweenness proportion greater than 1.5.

To address the variability of the betweenness statistic, two distinct perturbation techniques were then applied to the edge weights of each network: non-parametric bootstrapping and the addition of random noise to each gene-gene pair correlation. With each perturbation method, tumor and normal networks were simulated, and then confidence intervals for the difference in the tumor and normal log betweenness measures for each essentially different gene were constructed. If an essentially different gene had a confidence interval that excluded zero, then that gene was determined to be *statistically different* across the tumor and normal networks.

In order to measure the consistency of the method for identifying a set of essentially different genes, a sensitivity analysis of the ad hoc thresholding was performed, where the decisions for the particular thresholding values utilized in the identification of the set were substantiated with the level of accuracy,

$$\text{accuracy} = \frac{\text{number of genes that are both essentially different and statistically different}}{\text{number of genes that are essentially different}}$$

The sensitivity results (Figures 3 and 4) of the proposed framework suggest a general robustness of the betweenness centrality when used as a method for identifying genes essential to the functionality of a biological network.

1.3 Heatmap Centrality

Although nodes of high degree, high betweenness, and high closeness (another shortest-path based measure) have been identified as super-spreaders (i.e., nodes that are influential in the flow of information), the contribution of [3]

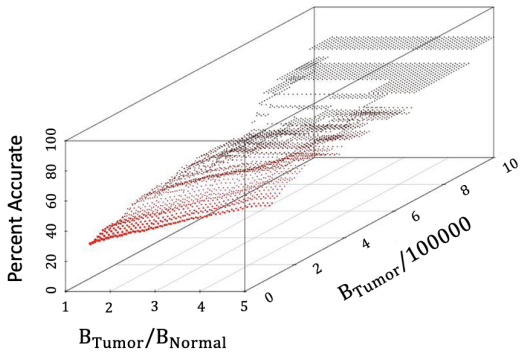

Figure 3: The essential gene identification accuracy associated with non-parametric bootstrapping, where the two planar axes represent the threshold values associated with the ratio of tumor betweenness to control betweenness, and the (scaled) tumor betweenness, while the height of the graph at the given point on the plane denotes the accuracy value.

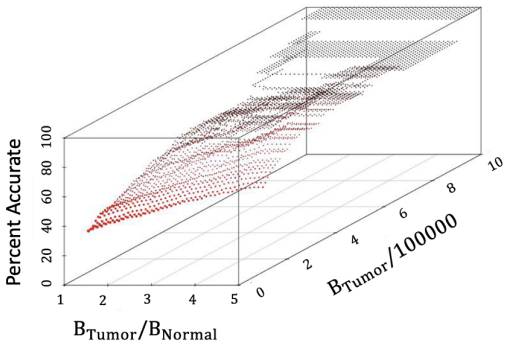

Figure 4: The essential gene identification accuracy associated with adding noise to the edge wights, where the two planar axes represent the threshold values associated with the ratio of tumor betweenness to control betweenness, and the (scaled) tumor betweenness, while the height of the graph at the given point on the plane denotes the accuracy value.

was the proposal of the *heatmap centrality*, a measure which utilizes features from all three centrality measures to strike a balance between accuracy and algorithmic simplicity in the identification of influential nodes. Motivated by a different interpretation of the "shortest path" between two nodes, this work explored the properties of the proposed centrality as a potentially viable measure in the identification of super-spreaders within real-world networks.

The heatmap centrality of a node is defined, simply, as the difference in the node's farness and the average farness of its neighbors, where *farness* is the sum of the shortest distances from a node to all other nodes in the network. The theoretical intuition behind the proposed measure is that a node is likely to lie on the shortest paths for several pairs of nodes within the network if the node's farness is smaller than the average of its neighbors. If a node and all of its neighbors have a similar farness, then information can flow through any of those nodes and neither of the nodes may be more influential than the others.

To verify the effectiveness of the heatmap centrality among the most commonly used centrality measures, two experiments (a comparison of CPU time in seconds and the correlation of the nodal rankings) applied to simulated scale-free networks and three experiments (a comparison of the top-10 ranked nodes, the average spreading influence of the top-10 ranked nodes using a susceptible-infected epidemic model, and the correlation of the nodal rankings) applied to four real-world scale-free networks were conducted. In short, the results indicated that the heatmap centrality may be executed in acceptable amount of CPU time (Figure 5), can successfully identify the top-10 influential nodes, and possesses a strong correlation with the betweenness centrality measure (Figure 6).

2 Publications Submitted for Peer-Review

2.1 Mean-Field Approximation for SIR Epidemics

The stochastic nature of epidemic dynamics on a network makes studying them directly very challenging. One avenue to reduce the complexity is a mean-field approximation of the dynamics; however, the classic mean-field equation has been shown to perform sub-optimally in many applications. The work in [4] aimed to address the disparity between the classic mean-field equation and simulations of the SIR (susceptible-infective-recovered) epidemic model on Erdös-Rényi networks by, first, proposing a new infection function f that describes how many susceptible individuals are, on average, infected during one time step,

$$f(S, I, \beta, d) = S(t) \cdot \left(1 - (1 - \beta)^{d \cdot I(t)}\right)$$

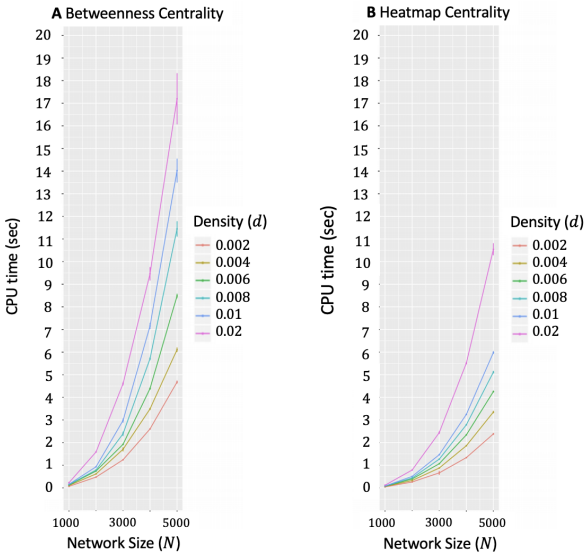

Figure 5: The CPU time (in seconds) of the both the (A) betweenness and (B) heatmap centrality measures required to calculate the value of each node in the scale-free networks of size N and density d averaged over 100 iterations. The standard deviation of the CPU times at each point is included.

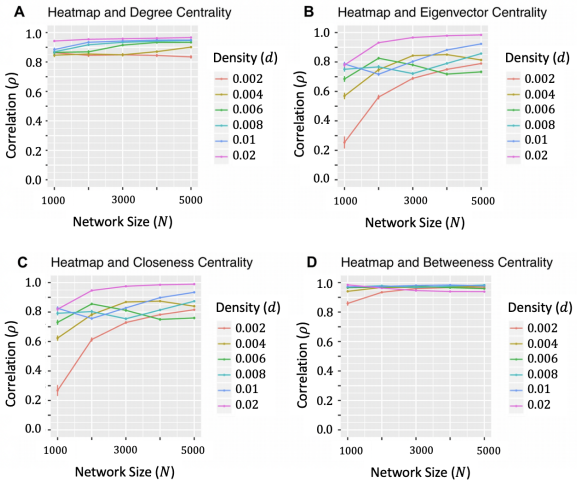

Figure 6: The value of the Spearman-rank correlation coefficient ρ for the rankings with respect to the (A) heatmap and degree, (B) heatmap and eigenvector, (C) heatmap and closeness, and (D) heatmap and betweenness centrality measures applied to each simulated scale-free network of size N and density d. The standard deviation of ρ at each point is included.

where $S(t)$ and $I(t)$ denote the number of susceptible and infective nodes, respectively, at time t, β denotes the probability of infection, and d denotes the network density. The inclusion of this infection function yielded the following discrete SIR model:

$$\begin{aligned} S(t+1) &= S(t)(1-\beta)^{d\cdot I(t)} \\ I(t+1) &= I(t) + S(t)\left(1-(1-\beta)^{d\cdot I(t)}\right) - (\mu+\rho)I(t) \\ R(t+1) &= R(t) + \rho I(t) \end{aligned} \tag{1}$$

where $R(t)$ denotes the number of recovered nodes, and μ and ρ denote the probability of succumbing and recovering from the disease, respectively.

To create a correspondence between the parameters of the discrete SIR epidemic model and SIR epidemics on the ER network model, first, a parameter set (d, β, μ, ρ) was selected. Then, for $N = 1000$ and $N = 5000$ node networks, 100 networks were created for each density and a single epidemic was simulated on each network. In an attempt to remove stochastic variations, the average of the 100 simulated epidemics was taken to create a single 30 day set of S, I, and R data for the 1000 and 5000 node networks. Next, the discrete SIR model (1) was fit to the single 30 day data and the best fit parameters, $(\tilde{\beta}, \tilde{\mu}, \tilde{\rho})$, were determined. To gauge the accuracy of the fitting procedure, the parameters input into the ER network, (β, μ, ρ), were compared to the best fit parameters from the discrete time model, $(\tilde{\beta}, \tilde{\mu}, \tilde{\rho})$.

The results (Figure 7) suggested that for the discrete SIR model, the modified mean-field equation using the proposed infection function and the Erdós-Rényi network simulations were consistent as the density of the network increased. Furthermore, the parameter fitting procedure had improved accuracy in the estimation of the network epidemic parameters as the average degree of the network increased.

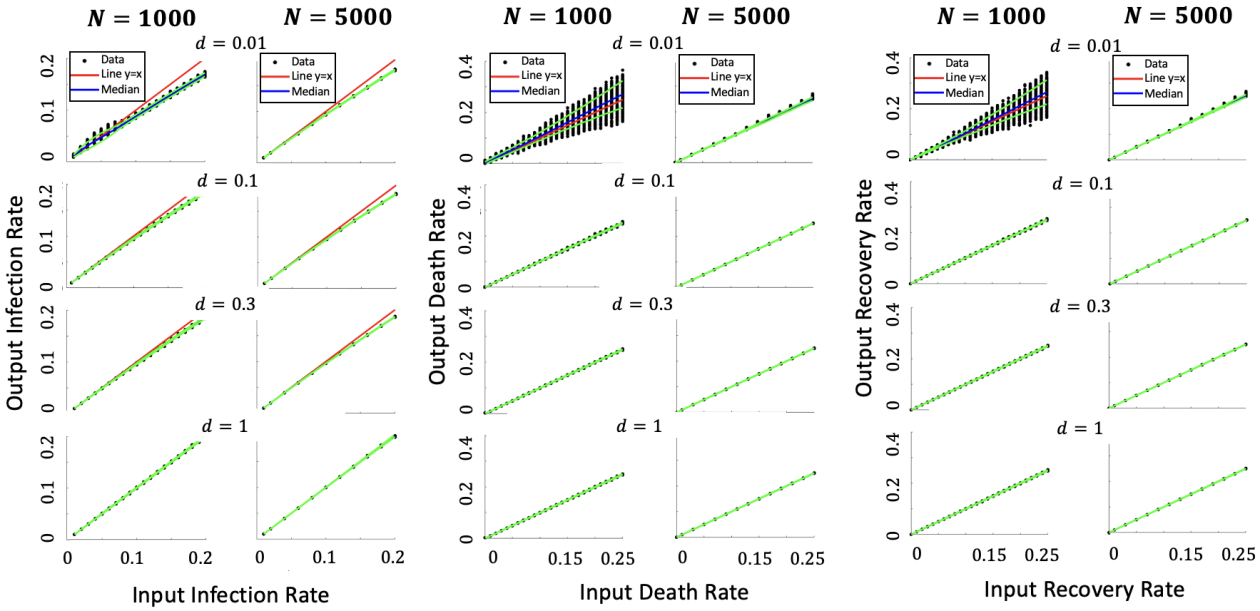

Figure 7: The correspondence of the infection, death, and recovery probability, respectively, between the parameters used to create the network data ('input') and the parameters found through the difference equation fitting procedure ('output') is plotted in blue for various network densities with $N = 1000$ and $N = 5000$ nodes and various densities d. The green lines denote a 95% confidence interval for the true parameter value.

3 Ongoing Research

3.1 Wasserstein Metric and Burn-in

Under standard conditions on the Markov chain, for any starting value X_0, the distribution of X_n converges to the stationary distribution π as $n \to \infty$. Yet, if the starting value is not in a high-density region, then the samples at the earlier iterations may not be close to the stationary distribution. To address this issue, the common practice is to *burn-in* by discarding early iterations in the chain. This collaboration aims to develop an algorithm that utilizes the Wasserstein metric, a distance function defined between probability distributions, to analytically determine the value of burn-in.

3.2 Wave Equation and Disease Transmission

In [5], a framework was developed to compute eigenvalues and eigenvectors of arbitrary order for a general metric graph (a finite set of nodes connected by oriented edges on which a metric is assigned), which was then used to compute solutions of linear PDEs on graphs. By applying their framework to analyze disease transmission on social networks, this collaboration will explore what eigenvalues mean in the context of disease transmission and how the shape of a social network will affect disease transmission.

3.3 Symmetry Module for Math Circles

In Spring 2021, this collaboration designed a 6-part module on symmetry for the online use in the Tucson Math Circle. Using a scaffolded, hands-on approach, new and old mathematical topics with various group-based activities were incorporated into each session. Currently, the goals, the activities used, how the students interacted with the material, and a reflection of the encountered successes and difficulties of each session are being detailed with the intention of making this module applicable for use in other Math Circles.

4 Future Research with Potential for Undergraduate Collaboration

I would like to work with undergraduate students to develop fundamental skills in the knowledge and conduct of mathematical research areas and in their applications. I would like to help them develop as professionals in their ability to write and in their articulation of their knowledge, for example, by writing papers and presenting at conferences or professional meetings with funding agency directors. Overall, I would like to show them that there exist tools to be learned that are beyond what is taught in the classroom that not only impact the scientific community, but also directly impact their future as a successful individual.

My research can incorporate undergraduate involvement through a variety of projects. To begin, since the heatmap centrality was shown to have a strong correlation with the betweenness measure, an extension of the work developed in [3] would be to improve the computational complexity of the heatmap measure. As of now, the heatmap centrality possesses the same time complexity $\mathcal{O}(Nm)$ given a network with N nodes and m edges, so developing a heuristic method to estimate this centrality would be advantageous in the analysis of information flow within networks. Since the heatmap measure is dependent upon shortest paths, students would start reviewing work that has developed faster methods to estimate these paths [6, 7].

Another project that allows for undergraduate research focuses on using matrix tools for mining networks, such as singular value decomposition (SVD), to identify the most important sub-network. While there exist many algorithms to detect sub-networks, SVD has been predominantly utilized in the analysis of network data. For example, a methodology to construct a partial SVD of the adjacency matrix (whose elements indicate the presence of an edge) associated with the network has been used to identify a subset of the most important nodes [8]. This work may be extended to rank both the most important nodes and edges of a network, thereby constructing the most essential substructure. Given that SVD is a time-intensive algorithm, more efficient algorithms to compute SVD may be used to improve the practicality of its utilization in identifying sub-networks.

Given the applicability of network science to a variety of fields, the impact of my research has the potential to be extensive. Furthermore, the fundamental concepts of network theory make research in this field very accessible to the undergraduate level, and may serve as an excellent first area of research for students. I hope to attract students from different academic backgrounds (including mathematics, physics, engineering, biology, and sociology) and leverage their training in core courses (e.g., calculus, probability, statistics, and linear algebra) to ensure success in my research program.

References

[1] Y. Pan, C. Durón, E. C. Bush, Y. Ma, P. A. Sims, D. H. Gutmann, A. Radunskaya, and J. Hardin, "Graph complexity analysis identifies an ETv5 tumor-specific network in human and murine low-grade glioma," *PLoS ONE*, vol. 13, no. 5, p. e0190001, 2018.

[2] C. Durón, Y. Pan, D. H. Gutmann, J. Hardin, and A. Radunskaya, "Variability of Betweenness Centrality and its Effect on Identifying Essential Genes," *Bulletin of Mathematical Biology*, vol. 81, no. 9, pp. 3655–3673, 2019.

[3] C. Durón, "Heatmap centrality: A new measure to identify super-spreader nodes in scale-free metworks," *PLoS ONE*, vol. 15, no. 7, p. e0235690, 2020.

[4] C. Durón and A. Farrell, "A Mean-Field Approximation of SIR Epidemics on an Erdös-Rényi Network Model," *Bulletin of Mathematical Biology*, Submitted June 2021.

[5] M. Brio, J.-G. Caputo, and H. Kravitz, "Spectral solutions of PDEs on networks," *arXiv preprint arXiv:2104.15048*, 2021.

[6] B. Li, G. Si, J. Ding, and F. Wang, "A faster algorithm to calculate centrality based on Shortest Path Layer," in *2017 29th Chinese Control and Decision Conference (CCDC)*, pp. 6283–6290, IEEE, 2017.

[7] A. Saxena, R. Gera, and S. Iyengar, "Fast Estimation of Closeness Centrality Ranking," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 80–85, 2017.

[8] J. Baglama, C. Fenu, L. Reichel, and G. Rodriguez, "Analysis of directed networks via partial singular value decomposition and Gauss quadrature," *Linear Algebra and its Applications*, vol. 456, pp. 93–121, 2014.