Research Statement

Christina Durón

My overall research interests lie in network analysis and network theory. Specifically, I focus on the development of 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 network centrality measure is defined as a function that numerically quantifies the level of influence of each gene (or node). Specifically, 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 were assigned using correlations between the RNA expression levels of each gene pair.

For each gene, its betweenness centrality value was calculated in both networks (Figure 1), where the betweenness centrality of a gene is defined as the fraction of shortest paths in the network that go through that gene. Upon comparing each betweenness measure, it was determined that the role of the regulator *Etv5* had substantially changed in diseased tissue relative to its healthy counterpart. 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 identified *Etv5* as a key regulator in the development of optic glioma and provided further evidence that *Etv5* and its associated target genes make up a potential regulatory network (Figure 2).

1.2 Heatmap Centrality

Although nodes of high degree, high betweenness, and high closeness (another centrality measure based upon shortest paths) have been identified as super-spreaders (i.e., nodes that are influential in the flow of information), the contribution of [2] was the proposal of *heatmap centrality*, a measure which utilizes features from all three network 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.

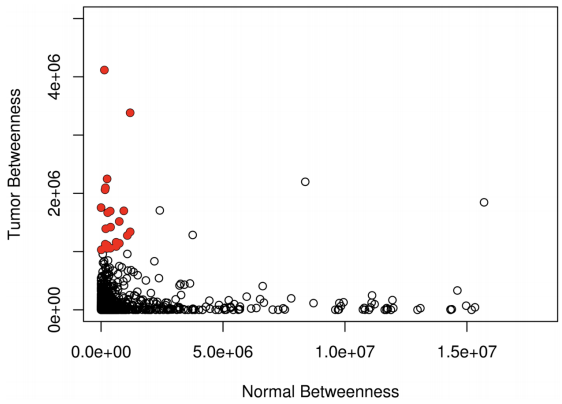

Figure 1: Filled (red) circles indicate genes whose betweenness measure is at least 1.1 times as large in the tumor network than in the normal network, and either a tumor betweenness or normal betweenness value greater than $1e6 = 10^6$.

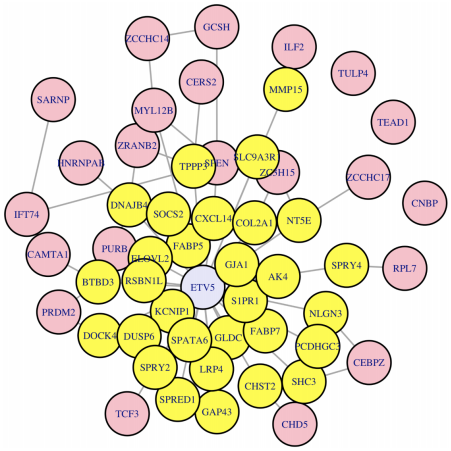

Figure 2: The subnetwork is comprised of *Etv5* (in lavender, center), and its differentially-expressed targets (in yellow). The remaining central genes (in pink) were identified by their high betweenness measures relative to the normal network.

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 none of the nodes may be more influential than the others.

To verify the effectiveness of 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 an acceptable amount of CPU time (Figure 3), can successfully identify the top-10 influential nodes, and possesses a very strong correlation with the betweenness centrality measure (Figure 4).

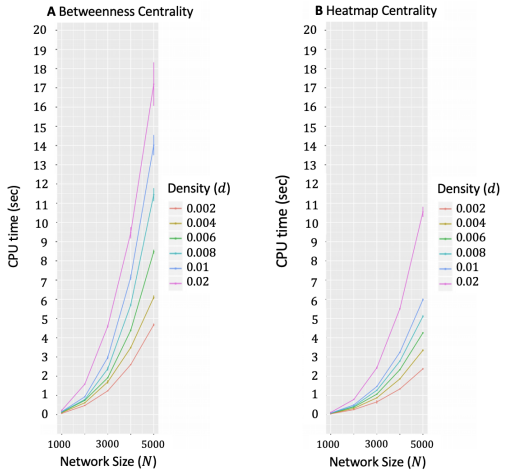

Figure 3: The CPU time (in seconds) of both the (A) betweenness and (B) heatmap centrality measures required to calculate the value of each node in the scale-free networks of size N and density d averaged over 100 iterations. The standard deviation of the CPU times at each point is included.

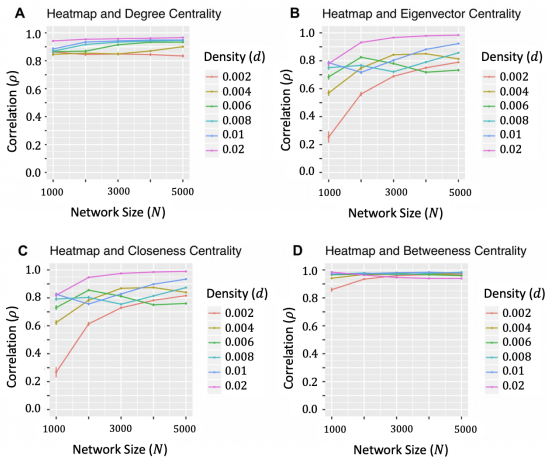

Figure 4: The value of the Spearman-rank correlation coefficient ρ for the rankings with respect to the (A) heatmap and degree, (B) heatmap and eigenvector, (C) heatmap and closeness, and (D) heatmap and betweenness centrality measures applied to each simulated scale-free network of size N and density d. The standard deviation of ρ at each point is included.

2 Publications Submitted for Peer-Review

2.1 Mean-Field Approximation for SIR Epidemics

The stochastic nature of epidemic dynamics on a network makes their direct study 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 [3] aimed to address the disparity between the classic mean-field equation and simulations of the SIR (susceptible-infective-recovered) epidemic model on Erdös-Rényi (ER) networks by, first, proposing a new infection function f that describes how many susceptible individuals are, on average, infected during one time step,

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

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, the discrete SIR model (1) was fit to epidemic data simulated on the network and the best fit parameters, $(\tilde{\beta}, \tilde{\mu}, \tilde{\rho})$, were determined using the least-squares criterion. 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 suggested that for the discrete SIR model, the modified mean-field equation using the proposed infection function and the ER 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.

2.2 Symmetry Module for Math Circles

Wallpapers are two-dimensional repetitive patterns that can be defined in terms of their 'generating tile', or the smallest polygon that can be found such that the entire wallpaper can be generated through an application of any of the four fundamental symmetries (e.g., reflective, rotational, translational, and glide). As symmetry is a natural property that children see on a daily basis, a 7-part module on symmetry was designed for online use in the junior section (grades K – 6) of the Tucson Math Circle (TMC), an outreach program where K – 12 students can engage with challenging and complex mathematical concepts.

Using a scaffolded, hands-on approach, the module outlined in [4] opened with reflective symmetry and culminated in the deconstruction of wallpapers into their generating tiles. Each session saw the incorporation of old and new mathematical topics through various group-based and interactive activities administered entirely online through Zoom and Miro, a free website that allows all users to participate and share ideas digitally on a communal infinite whiteboard. The final discussions were facilitated through the use of the TMC Widget, a freely accessible web-based application designed and created specifically for the symmetry module. Using the TMC Widget, students can select a wallpaper, highlight a sub-portion, and repeatedly copy, rotate, move, and reflect the highlighted piece in an attempt to recreate the original wallpaper.

With the intention of making the module applicable for use in other Math Circles, a lesson plan was developed for each session that discussed its goals and described the motivation, setup, and steps of each activity. In addition, a reflection on the successes and difficulties encountered during each session were included such that future session leaders could modify their lesson plans accordingly.

3 Ongoing Research

3.1 Wasserstein Metric and Burn-in

A common approach to learn about a probability distribution is to draw samples from it. In statistics and machine learning, this is often done using Markov chain Monte Carlo (MCMC) algorithms which construct Markov chains by accepting or rejecting the candidate samples as the new state of the chain. 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 work aims to develop an algorithm that utilizes the Wasserstein metric, a distance function defined between probability distributions, to determine the value of burn-in for univariate and multivariate distributions.

3.2 The Diffusion of Epidemics on Metric Graphs

It has long been known that epidemics can travel along transportation routes, such as major roads and highways. In [5], the speed of epidemic propagation on a line and half-plane was studied using a classical SIR model with diffusion. In [6], a PDE-ODE model was proposed to study epidemic dynamics where each node of the graph had a standard SIR model, and the edges between nodes were given by the heat equation with Robin-like boundary conditions at the nodes. This research will expand upon the work of [5] by extending the half-plane to a bounded domain to study the diffusion of epidemics on metric graphs (i.e., networks in which the notion of physical distance is defined on each edge), where the node conditions of [6] will be imposed.

4 Future Research with Potential for Undergraduate Collaboration

I plan to engage undergraduate students to develop fundamental skills in the knowledge and conduct of mathematical research areas and in their applications. My goal is 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 program officers. Overall, I will 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 very strong correlation with the betweenness measure, an extension of the work developed in [2] would improve the computational complexity of the heatmap measure. As of now, the heatmap centrality possesses the same time complexity $\mathcal{O}(Nm)$ given a scale-free network with N nodes and m edges, so developing a heuristic method to estimate this centrality more would be advantageous in the analysis of information flow within real-world networks. Since the heatmap measure is dependent upon shortest paths, students would start reviewing work that has developed faster methods to estimate these paths [7, 8].

Another project that allows for undergraduate research focuses on matrix tools for mining networks, such as the 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 [9]. 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 (e.g., chemistry, finance, and social sciences), the impact of my research has the potential to be extensive. Furthermore, the fundamental concepts of network theory make research in this field accessible to the undergraduate level, and may serve as an excellent first area of research for students. I hope to attract undergraduates 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, "Heatmap centrality: A new measure to identify super-spreader nodes in scale-free metworks," *PLoS ONE*, vol. 15, no. 7, p. e0235690, 2020.

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

[4] N. Fider, C. Durón, and D. Pfeffer, "From Mirrors to Wallpapers: A Virtual Math Circle Module on Symmetry," *Journal of Math Circles*, Submitted August 2021.

[5] H. Berestycki, J.-M. Roquejoffre, and L. Rossi, "Propagation of Epidemics Along Lines with Fast Diffusion," *Bulletin of Mathematical Biology*, vol. 83, no. 1, pp. 1–34, 2021.

[6] C. Besse and G. Faye, "Dynamics of epidemic spreading on connected graphs," *Journal of Mathematical Biology*, vol. 82, no. 6, pp. 1–52, 2021.

[7] 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.

[8] 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.

[9] 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.