RESEARCH STATEMENT

Benjamin Leinwand

In recent years, researchers have introduced several modes of capturing and extracting information from brain scans. My research produces models that take advantage of the novel structures of these newly collected datasets. This involves developing theory for modeling dense, weighted networks and correlation networks, and applying these models to DTI or fMRI data. As many modalities collect signals over the course of a session, some projects require adapting tools from time series analysis. I incorporate insights from statistics, probability, and network science to address the unique challenges presented by each problem. I see methodology and applications as mutually beneficial endeavors, and prefer to work on both.

1 Networks

1.1 Frequency-Selective Network Vulnerability for Alzheimer's Disease [1]

The first branch of my research takes inspiration from structural brain networks, the physical connections between brain regions made up of white matter fibers. [1] uses changes in structural networks over time to assess a patient's risk of developing Alzheimer's Disease (AD) in the near future. Early diagnosis can improve treatment outcomes, but current early diagnostic tools are invasive. Our work integrates patient brain scan data together with other medical information from the Alzheimer's Disease Neuroimaging Initiative database, with data for each patient collected on average approximately once every six months. Longitudinal analysis is of particular interest in AD neuroimaging studies, since the measured subject-specific network changes allow us to develop connectome biomarkers for early AD diagnosis. We introduce a method that accounts for known heterogeneities in brain structure, while remaining flexible enough to handle new scans. We also identify markers that may indicate elevated risk for transitioning to AD by leveraging physical changes in particular regions of the brain to provide incremental information.

Our approach consists of two major steps. (1) Align all observed brain networks into a common graph spectrum domain. (2) Inter-subject variable selection in the graph spectrum domain. For this second step, we propose a novel sparse regression model to identify the critical frequency patterns along which observed relative within-subject longitudinal changes contribute to selective network vulnerability. By focusing on longitudinal changes to brain scans in a common space, we can incorporate benign subject-specific network structural information while looking for signatures which can hopefully generalize to a broader, unobserved clinical population. In particular, our procedure identifies four variables which are useful in predicting those patients who may transition to AD. The first two are CDRSB scores and changes in MMSE scores, well known tests used by clinicians to diagnose AD. The latter two included variables are derived from brain scans, and indicate that major oscillations occur at the temporal and parietal lobes, which are aligned with the default mode network. Our model uses LASSO to filter out irrelevant variables, so the selection of brain scan variables indicates that these scans contain information which incrementally improve the model's fit.

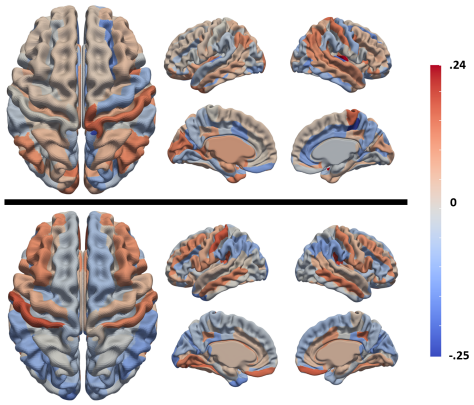

Figure 1: Visualizations of the variables derived from brain scans projected onto the cortical surface, with views of each hemisphere from above, left, and right. Relative changes along specified regions can predict elevated risk of transitioning to AD.

1.2 Modeling Dense, Weighted Networks [2]

While working on [1], I realized that many of the methods used to analyze other kinds of networks do not seem appropriate for studying brains. Existing methods are designed for sparse, unweighted graphs, like those documenting friendships in a social network. As nearly all regions in the brain are connected

to one another, though, structural brain networks are characterized by their density, and have weighted edges representing the number of white matter fibers connecting two regions. Dense, weighted networks often exhibit a community like structure, where although most nodes are connected to each other, different patterns of edge weights may emerge depending on each node's community membership. For example, Figure 2 illustrates the different connectivity patterns observed in a subject's left and right hemispheres, as well as the bipartite subnetwork between the two hemispheres. Our paper [2] proposes a new framework for modeling dense, weighted networks with potentially different connectivity patterns across different groups.

Our key contributions at the model level concern:

- Focus on dense weighted networks;
- Possibility of community structure;
- Nearly arbitrary distribution of edge weights, including allowing for multiple (potentially unique) distributions of edge weights connecting nodes in each pair of communities;
- Flexible degree correction [3] patterns while incorporating errors;

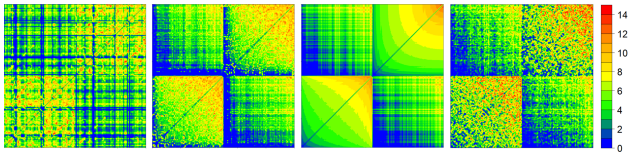

Figure 2: From left to right: a structural brain network of the log values of white matter fiber counts between 148 regions. The same network reordered by within community degree. The similarly reordered "estimate" of the structural network. A reordered bootstrap replicate network of the observed network.

For the last point, we introduce a class of "H-functions," which can map node "sociabilities" to edge weight orderings. $H : (0,1) \times (0,1) \to (0,1)$ is a monotonic function in both arguments such that $\iint_{H(x,y)\leq z} dxdy = z$, for all $z \in (0,1)$. If we give each node in the network a nodal (degree correction) attribute $\Psi \in (0,1)$, H-functions mandate that $H(\Psi_x, \Psi_y)$ is a Uniform $(0,1)$ random variable for all node pairs $x \neq y$. This is a rich class of functions, but we found one general construction to be both flexible and practical. Note that a random variable $F^{-1}(\Psi)$ has the CDF F for a $\mathrm{U}(0,1)$ random variable Ψ. Take two CDFs F_1, F_2 and let $F_{1,2}$ be their convolution CDF. Then $F_1^{-1}(\Psi_u) + F_2^{-1}(\Psi_v)$ has the same distribution as $F_{1,2}^{-1}(\Psi_{uv})$. This suggests setting

$$H(x,y) = F_{1,2}(F_1^{-1}(x) + F_2^{-1}(y)). \tag{1.1}$$

The choice of F_1 and F_2 will define the contours observed in the network, as shown in Figure 3. By allowing the H-function to depend on the communities to which nodes x and y belong, one can get differing patterns of connections, as can be observed in the third plot of Figure 2, where the bottom-left and top-right subnetworks look like different examples from Figure 3. To account for imperfect signal, since $H(x,y)$ is uniformly distributed, these values can be converted into the "normal" space, where normally distributed errors can be injected, with the magnitude of errors again depending on the communities of x and y, modulating the smoothness of the degree correction pattern in each subnetwork. Converting this new value back to uniform and drawing from the edge weight distribution of the relevant subnetwork, this process achieves the four stated aims above, and can produce qualitatively different networks than the kinds discussed in other works such as [4, 5].

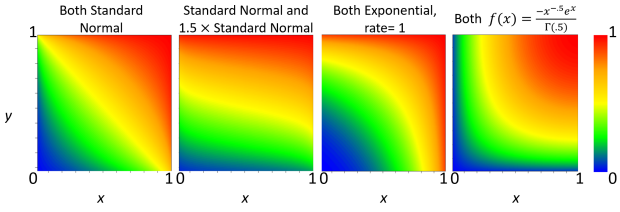

Figure 3: Plots of examples of H-functions.

In addition to the network generative model, we also introduce a measure which balances the desires for both larger estimated communities and homogeneity of the estimated communities. We develop algorithms to maximize this measure. Once community assignments have been estimated, we can use them to calculate "local" estimates of nodal features, then choose the best fitting H-functions and error magnitudes, providing an estimate of the degree correction patterns. We prove a concentration inequality on the distance between our estimates and the true underlying degree correction pattern of the network if the H-function is linear in the "normal" space, borrowing techniques from [6]. As a byproduct of our estimation techniques, we also develop a methodology for generating new networks of the same type as our observed data, including when some edges are missing. This may be beneficial when it is expensive or difficult to collect additional data.

1.3 Bipartite Networks and Recommender Systems [7]

Collecting complete data for large dense networks is challenging because the number of edges scales quadratically with the number of nodes. However, for networks where edges represent "affinities" between nodes, one can imagine that two entities may have an affinity even when they have not yet interacted. For example, inputs to recommender systems can be represented as a matrix where each row represents a user, each column represents a product (e.g. a movie or song), and each entry represents a user's rating of a given product. For services with large corpuses and subscriber bases, most of the entries of this matrix will be empty, so the true affinities are unknown. Matrices of this type can be seen as partially observed, dense, weighted, bipartite networks. One can use similarities across users and across products to recommend products to users. To adapt our work in [2] to this new setting, we account for missing values, utilize the bipartite structure of the network, and integrate notions of modularity into our proposed methodology.

1.4 Multilayer Network of Governmental Agreements [8]

Inspired by a dataset of governmental agreements over 26 years in 33 service types, we examine the evolution of organizational relationships by building a multilayer network [9] where nodes are agreement participants. In contrast to brain networks, researchers have greater intuition about cause and effect in this network, and can interview participants in the network. Additionally, the dataset is inherently incomplete, as it can be difficult to capture exogenous shifts in the broader environment which often drive agreement behavior. However, like brain networks derived from low resolution scans, the measured nodes in this network may in fact represent many individual, independent actors, resulting in seemingly inconsistent patterns of agreements. Network statistics can detect otherwise hidden quirks in network topologies, identify latent issues, and eventually improve governance. As the team includes researchers across math, public administration, and psychiatry, we are also interested in translating lessons from one domain, like multilayer EEG networks [10], to a seemingly unrelated topic like public management [11, 12], and vice versa.

1.5 Future Work

Agency in Brain Networks: [13, 14] model the spread of brain disease as a reaction-diffusion process characterized by differential equations, but the brain is not an inert medium. Implicit in network centrality metrics is the notion that each node's edges are valuable. Inspired by economics, we treat each node's edges as their literal capital stock, incorporate nodal features and constraints based on network structure, and examine the equilibrium. For example: does the equilibrium induced by nodes acting in their individual interests result in deadweight loss relative to a strategy imposed by a central planner? It's worth exploring how predicted neurodegeneration paths from reaction-diffusion type models differ from degeneration predicted by each region "fending for itself" as well as that predicted by central planner models. Determining which model appears to best describe the true process can help clinicians develop personalized, targeted treatments.

Multilayer Mouse Bacteria Networks: Community detection for a multilayer network can depend on choices of both data structure and clustering methodology, where the optimal choices are not known a priori. Using data consisting of bacteria concentrations in five different organs from five different mice, we can construct a 25-layer network where we have several potentially conflicting sources of information. In addition to expecting similarity between layers representing the same organ or the same mouse, we also have information about bacterial biology which could potentially aid bacterial clustering. However, there may be no available methodology which can cluster organs, mice, and bacteria as we might expect. In this circumstance, is there a principled approach to community detection? What "consensus" information and what incremental information can be attained by taking several different approaches?

2 Time Series and Other High-Dimensional Modeling

2.1 Two Sample Tests for High-Dimensional Autocovariances [15]

Another subset of my research concerns analyzing brain activity data collected in fMRI studies that measure blood flow in the brain. Multivariate time series data $X_t = (X_{1,t}, \ldots, X_{d,t})$, $t = 1, \ldots, T$, with a large

number d of univariate component series $X_{j,t}$, called high-dimensional time series (HDTS), are prevalent in fMRI studies where a component series represents a signal at a particular brain location, or ROI [16]. Signals from many ROIs are collected at regular intervals to create many concurrent time series. The focus of the work is to determine whether two distinct samples have the same pattern of evolution, by testing whether their autocovariance functions (ACVFs) are equal.

In [15], we adapt two existing tests for high-dimensional means (sup and sum tests), introduce PCA tests, and compare their performance across three regimes (sparse models, dynamic factor models, and a combination of both). Sup tests use the entry in the estimated ACVF (up to a certain lag) which, averaged over t displays the largest difference between the two series. Sum tests sum over the squared values of the differences for all entries in the estimated ACVFs, again averaged over t and up to a certain lag. Based on the test design, one would expect sup tests to perform better on sparse models. PCA tests, on the other hand, are tailored specifically for latent factor models; by concatenating two HDTS and taking the resulting estimated "pooled" covariance matrix, we can use PCA to get a "factor series" for each HDTS, which we treat as low dimensional stationary time series. Using a limiting distribution on the difference between the estimated ACVFs of identical factor series (up to a lag), we propose a test statistic to detect differing ACVFs in the original HDTS.

Surprisingly, though it is not theoretically justified in all cases based on the described construction, the PCA test performs best in simulations across all regimes. The PCA test almost always has the greatest power of all tests, and the test also demonstrates the appropriate size in all settings. This last point is important, as in an analysis of fMRI data trying to distinguish induced anxiety from induced anger using functional connectivity patterns, the PCA test picks up statistically significant differences in more subjects than the other tests. As noted, PCA tests are the least researched of the described methods, so these results provide reason to continue developing theory about these tests, particularly in the sparse and combined regimes.

2.2 Detecting Functional Connectivity Changes in fMRI Data [17]

A natural extension of the above work is to determine if there are "change points" within the same sample. A change point is a time point that separates the sample such that the data collected before that time has different properties from the data collected afterwards. Without accounting for change points, recovered connectivity patterns based on a scan are susceptible to spurious effects. Furthermore, while the earlier work compares two series, a sample may contain *multiple unknown* change points, requiring us to both identify when changes occur, and to repeatedly test whether connectivity patterns differ significantly before and after these proposed change points.

Our work [17] discusses three tests for detecting change points. Dynamic Connectivity Regression (DCR) looks for a significant reduction in BIC when splitting the data at that point. The BIC is calculated based on an estimate of the precision matrix of (a rescaled version) of the original HDTS. Furthermore, if a change point is detected, the process is recursively repeated on the newly separated subsegments of the HDTS until no new change points are detected. Max-type methods are similar to the sup tests above, focusing on the largest difference in ACVFs along a local window around each time point. PCA methods are also very similar to those described above, and can use binary segmentation or a sliding window. Each of these detection methods has a corresponding test for detecting differences between induced subsegments.

Using both simulated data and collected fMRI data, we assess the performance of each proposed method. No method clearly outperforms the others in general. However, there are persistent differences, as DCR and PCA using binary segmentation are less likely to pick up false positives than other methods. Along with other methodological suggestions in the paper, this may also guide clinicians toward more careful analysis. Nonetheless, additional work may be required to improve detection of true change points.

2.3 Hypocells [18]

Cellular differentiation is a key gene expression modulation process that enables multicellular life. An archetypal example of differentiation is hematopoiesis, or the process by which the multiple types of blood cells are generated from a self-renewing pool of stem cells. Research into the nature of hematopoeisis, especially recent work in single-cell RNA sequencing, has allowed for the discovery of many basic principles in differentiation through cell trajectory inference, the process of tracing out a medial path that cells traverse.

In contrast to previous methods, we propose to learn a model of differentiation that allows us to simulate the entire differentiation paths of *individual cells*, not merely analyze trends of the "average" cell in a differentiation trajectory. This shift of focus mandates that "more optimal" policies capture additional dynamics in the observed dataset beyond a "typical" cell's behavior. We ground this approach in a statistical formulation of differentiation in gene expression space. We formalize trajectory inference as a partially observable Markov process and present a task-centric redefinition of cellular differentiation trajectory modeling. We present this viewpoint, evaluation metrics, and new exploration capacities enabled by this recasting of trajectory inference. Our redefinition entails a complete, end-to-end *in silico* framework for *causal* analysis of gene expression trajectories in individual, hypothetical cells (hypocells), which, for experimental reasons, cannot be directly observed with current or near-current technologies.

2.4 Future Work

Recovering Gene Regulatory Networks (GRNs): We can incorporate the work discussed earlier into a larger framework. By fixing a (real or hypothetical) GRN, tools like [19] can generate a simulated dataset of many cells' gene expressions at a particular time. After simulating hypocell trajectories as in Section 2.3, change points within these trajectories may demarcate discrete modules of the estimated Gene Regulatory Network, which can be compared to the known initialized GRN. As real GRNs may include different dynamics for different cell types, careful analysis of simulated data can provide not only improved techniques, but also inform how to conduct more effective experimentation.

Change points for dense, weighted networks: As the model in Section 1.2 relies on operations in the "normal" space, it naturally lends itself to time series analyses. One possibility is to allow each edge to take a random walk in the latent "normal" space. Another is to convert the Ψ value for each node to a standard normal random variable that takes a random walk, which be translated back to a uniform random variable at each time step. By observing the same network at multiple times, one can look for times where the structure of the network dramatically changes, such as when community structures change, or if the network uses different H-functions to define connectivity patterns. Both of these changes might be observed in an fMRI when a subject switches tasks.

3 Other Work

Working on a smaller, fun side project provides an opportunity to step away when I need a break from my main research projects. I usually come back to "real" work re-energized and excited. I plan to seek out other side projects in the future, especially regarding topics that can interest the non-statisticians in my life.

3.1 Winning an Election, Not a Popularity Contest [20]

In two of the last six U.S. Presidential elections, the winner received fewer popular votes than his opponent. We ask what is the smallest percentage of the popular vote a candidate can garner while still winning the election? Using Monte Carlo methods and Integer Programming, we show that using the voter turnout in 2020 or the "voting eligible population" [21] in each state, a candidate can win an election with under 23% of the popular vote. By redistributing the population across states, it would be possible to win with just 15.6% of the popular vote. By taking even further liberties with the rules, we show there is no theoretical lower bound on the proportion of popular vote a candidate would need to win the electoral college. We also examine more "realistic" scenarios by approximating each party's minimal support in each state. While Republicans can win the election getting a smaller vote share than Democrats could, a Democrat could still win the election while losing the popular vote by more than seven points.

References

[1] **B. Leinwand**, G. Wu, and V. Pipiras, "Characterizing frequency-selective network vulnerability for alzheimer's disease by identifying critical harmonic patterns," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4, IEEE, 2020.

[2] **B. Leinwand** and V. Pipiras, "Block dense weighted networks with augmented degree correction," *arXiv preprint arXiv:2105.12290*, 2021.

[3] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, p. 016107, Jan 2011.

[4] C. Aicher, A. Z. Jacobs, and A. Clauset, "Adapting the stochastic block model to edge-weighted networks," *ICML Workshop on Structured Learning*, 2013.

[5] T. P. Peixoto, "Nonparametric weighted stochastic block models," *Physical Review E*, vol. 97, p. 012306, Jan 2018.

[6] M. Noroozi, R. Rimal, and M. Pensky, "Estimation and clustering in popularity adjusted block model," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 83, p. 293–317, Apr 2021.

[7] **B. Leinwand** and V. Pipiras, "Bipartite augmented degree correction with applications to recommender systems," *In preparation.*

[8] **B. Leinwand**, K. Albrecht, F. Zheng, A. Campbell, J. Thomas, and P. Mucha, "Multilayer network analysis of iowa governmental agreements," *In preparation.*

[9] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.

[10] V. Müller, D. Perdikis, T. von Oertzen, R. Sleimen-Malkoun, V. Jirsa, and U. Lindenberger, "Structure and topology dynamics of hyper-frequency networks during rest and auditory oddball performance," *Frontiers in Computational Neuroscience*, vol. 10, p. 108, 2016.

[11] B. Nowell, M. C. Hano, and Z. Yang, "Networks of networks? Toward an external perspective on whole networks," *Perspectives on Public Management and Governance*, vol. 2, no. 3, pp. 213–233, 2019.

[12] H. Yi, Y. Yang, and C. Zhou, "The impact of collaboration network on water resource governance performance: evidence from China's Yangtze River Delta region," *International Journal of Environmental Research and Public Health*, vol. 18, no. 5, p. 2557, 2021.

[13] T. Aoki and T. Aoyagi, "Scale-free structures emerging from co-evolution of a network and the distribution of a diffusive resource on it," *Physical Review Letters*, vol. 109, no. 20, p. 208702, 2012.

[14] J. Zhang, D. Yang, W. He, G. Wu, and M. Chen, "A network-guided reaction-diffusion model of at [n] biomarkers in Alzheimer's Disease," *arXiv preprint arXiv:2009.04854*, 2020.

[15] C. Baek, K. M. Gates, **B. Leinwand**, and V. Pipiras, "Two sample tests for high-dimensional autocovariances," *Computational Statistics & Data Analysis*, vol. 153, p. 107067, 2021.

[16] H. Ombao, M. Lindquist, W. Thompson, and J. Aston, *Handbook of Neuroimaging Data Analysis*. CRC Press, 2016.

[17] C. Baek, M. Gampe, **B. Leinwand**, K. Lindquist, J. Hopfinger, K. M. Gates, and V. Pipiras, "Detecting functional connectivity changes in fMRI data," To appear in *Brain Connectivity*, 2021+.

[18] E. Robson, **B. Leinwand**, and V. Pipiras, "Hypocells: a machine learning framework for in silico simulation of cellular differentiation," *In preparation.*

[19] R. Cannoodt, W. Saelens, L. Deconinck, and Y. Saeys, "Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells," *Nature Communications*, vol. 12, p. 3942, Dec 2021.

[20] **B. Leinwand**, P. Ge, V. Kulkarni, and R. Smith, "Winning an election, not a popularity contest," *Significance*, vol. 18, no. 4, pp. 24–29, 2021.

[21] M. P. Mcdonald, "United States elections project." http://www.electproject.org.