Goodness of Fit of Network Data
Peter Liu, Senior, Statistics
Mentor: Shane Lubold
Mentor: Tyler McCormick

Network goodness of fit (GOF), which deals with determining if an observed network could have been drawn from a particular graph model, is a common method to identify network structures and a prerequisite of further network analysis. In this talk, we will present a method that constructs a test statistic using advances in Random Matrix Theory, which traces the limiting behavior of the leading eigenvalue of the observed graph’s adjacency matrix to perform network GOF. We show that under many network models - such as the Degree Corrected Stochastic Block model (DCSBM), the Beta model, and the Aggregated Relational Data (ARD) - our method performs well in identifying the correct model with high precision and low computation time. We further extend our method to the popular Latent Space model, and develop an algorithm that fast-and-accurately predicts the underlying dimension of the true model. With precision grows with network sizes, our algorithm enables better estimations for large-scale networks with minimal computational cost.

Principal Component Analysis Reveals Political Leaning of US States Tied to Economic Status and thus Everyday Life
Joia W (Joia) Zhang, Junior, Pre-Sciences
Mentor: Jerry Wei, Statistics
Mentor: Abel Rodriguez, Statistics

In recent decades, partisanship between the Democratic and Republican parties in the US has grown, resulting in congressional gridlock and economic stagnation. Untangling the relationship between partisanship and major economic factors such as household income, homeownership, population, and poverty has the potential to solve these problems. Our hypothesis is that a state’s political leaning is not associated with these factors. Our data contained the political leaning of all 50 states and DC in the 2008 presidential election, alongside 71 economic variables in the aforementioned four categories of annual household income (1984-2018), homeownerships (1986-2014), population (2000-2005), and poverty rate (1990, 2000). We used principal component analysis (PCA) to condense the 71 dimensions into two dimensions for visualization, principal components (PCs) 1 and 2. They explained 48% and 13% of total variance. In the plot between the two PCs, dots represent states (and DC) colored by political leaning. The visualization revealed that a state’s political leaning is strongly tied to the 4 economic categories. Thus, we rejected our hypothesis. Political party and economic status are related, but determining how they are related is a limitation of our research. Further work must be done to determine whether it is political leaning that determines economic status or economic status that determines political leaning, or both. Unraveling the relationship between politics and economics can provide insights into the symptoms,
causes, and possible solutions to the US’s growing polarization. Keywords: partisanship, household income, homeownership, poverty, population, education

SESSION T-7B

MATHEMATICS & URBAN DEVELOPMENT

Session Moderator: Tamre Cardoso, Statistics

A Consulting Tale: Analysis of Schirmer Tear Test Data for Pigtail Macaques
Samantha Shimogawa, Fifth Year, Statistics: Data Science
Mentor: Serena Young, Veterinary Services, WaNPRC

Dry eye is a common disease of older adults that produces symptoms ranging from mild discomfort to visual disturbances. Treatments range from lifestyle changes, use of eye drops, to surgical interventions. A nonhuman primate model of severe dry eye disease has been developed using rhesus macaques and employs the use of the Schirmer 1 Tear Test (STT-1). Pigtail macaques would likely be similarly suited as a model for this disease, however, no STT-1 values have been published for this species. Our objectives are to determine the normal range of STT-1 values in apparently healthy pigtail macaques, while considering age, sex, and type of sedation. The data consists of STT-1 values in each eye on 218 pigtail macaques, along with who performed the test, type of sedative used, and the age and sex of each monkey. Analysis of these data using t-tests, ANOVA and linear regression methods indicate that STT-1 values vary significantly between left and right eyes, type of sedation, and staff conducting the test. STT-1 values do not vary significantly by sex but may vary by age. We used the observed relationships to determine an overall 95% baseline interval of STT-1 values for healthy pigtail macaques, as well as individual 95% baseline STT-1 intervals that depend on sedation type and age.

SESSION T-7B

MATHEMATICS & URBAN DEVELOPMENT

Session Moderator: Tamre Cardoso, Statistics

Infectious Disease Modeling
Harper Zhu, Senior, International Studies, Biochemistry
UW Honors Program
Mentor: Abel Rodriguez, Statistics
Mentor: Anna Neufeld, Statistics, Washington

Against the backdrop of Coronavirus spreading on a worldwide scale, public health policy became a priority among universities worldwide. Statistical modeling emerged as a solid tool to visualize the role of mask-wearing and social gathering in the spread of this global pandemic. This research aims to simulate the social network of college students living on-campus and how that will impact the spread of Coronavirus within the college campus. We design a simple network model that approximates the type of social interactions occurring on a college campus, which relies on assumptions about people’s living situations (e.g., how many roommates they have, how many people have in-person work). We simulate the disease spreading through the network model by allowing individuals to move between Susceptible, Exposed, Infected, and Recovered states. The transition probabilities between states are determined by an individual’s social interactions and mask-wearing habits. We created an interactive data visualization tool where the users will be able to adjust the parameters to explore the impact of each parameter on the dynamics of the disease. The visualization tool indicates that even if a large percentage of people wear masks, a social gathering such as a party can lead to increased transmission of the disease. The result of this research will be able to approximate the pandemic trend within the campus to inform individual students and policymakers at the university and shape further decisions (e.g., Whether or not to attend social gatherings or initiate university re-opening).
SESSION T-7B

MATHEMATICS & URBAN DEVELOPMENT

Session Moderator: Tamre Cardoso, Statistics

Join Room

3:10 PM to 4:00 PM

* Note: Titles in order of presentation.

Analyzing Temporal Trends in Leukemia Incidence using Knots in Nonlinear Regression
Alejandro Fabian Gonzalez, Freshman, Business Administration
Mentor: Michael Pearce, Statistics
Mentor: Abel Rodriguez, Statistics

Statistical models are fundamental to identify and understand cancerous tendencies and properties in our bodies. Much of the current research focuses on the relationships between binary gene expressions and cancer incidence, which often leads to uninterpretable models due to complex relationships between gene expressions. Instead, using knot identification and analysis in nonlinear modeling creates more interpretable trends. Using data by age, sex, and race from the National Cancer Institute, we analyze leukemia incidence in the period 1975-2017 using regression splines, a technique that partitions the model into several piecewise functions at various knots in the covariate space. Knot locations are chosen to provide interpretable results and minimize the least squared error, which allows for inference based on techniques from linear regression. After ANOVA forward selection for the polynomial regression model and general cross-validation for the natural cubic spline, the knot points converged on an interval between 1985 to 1986. This suggests that the female cancer incidence rate developed an exponential cancer growth in an interval of 1 year. Therefore, to oppose future exponential incidence increases in female rate, conducting medical research for genomic or environmental causation factors will be more explicit and accelerated due to the specificity of the 1985 to 1986 time interval.