Research Statement

Daniel K. Sewell

My primary research area is in the analysis of network data, with secondary interests in infectious disease (ID) and clustering. Network analysis interfaces very naturally with infectious disease, and I have both brought network analysis tools to answering ID questions and developed new methodology motivated by ID problems. Clustering is a widely applicable branch of analysis; its intersection with networks, however, yields unique methodological challenges. Below I provide a summary of my work in these three domains, highlighting the connections between ID, clustering, and network analysis. Beyond these, I have also engaged consistently in interdisciplinary work in community health yielding six publications since 2021, as well as in general Bayesian methodology (Sewell 2024).

Infectious Disease

I am PI on a multi-PI NIH project called, "Statistical and Agent-based Modeling of Complex Microbial Systems: A Means for Understanding Enteric Disease Transmission Among Children in Urban Neighborhoods of Kenya" (R01TW011795-05), nicknamed PATHOME. I am also co-I on the CDC MInD Healthcare funded project "Contact Network Transmission Modeling of Healthcare Associated Infections" which produces both interdisciplinary and methodological research.

As PI of the PATHOME project, I have, together with microbiologists, behavioralists, and field staff, developed large study protocols (Baker et al., 2023) and protocols for implementing geotracking of infants and animals (Busienei et al., 2024). Much of this work focuses on the analysis of microbiological data (Baker et al., 2022; Hoffmann et al., 2022; Gutema, Cumming, et al., 2024; Gutema, Okoth, et al., 2024, under review).

Other interdisciplinary work I have been involved with has focused on hospital epidemiology. Many such projects have utilized big data to obtain insights into healthcare associated infections (HAIs) and their risk factors (Miller et al., 2022; Miller et al., 2023; Miller et al., 2024). One such area I am branching into is health disparities research, defined by the CDC as "preventable differences in the burden of disease, injury, violence, or opportunities to achieve optimal health that are experienced by socially disadvantaged populations." This important and persistent issue merits increased attention by interdisciplinary teams, including biostatisticians. I have ongoing work utilizing large data to decompose differences in health outcomes such as lenth of inpatient stay and HAIs between racialized groups, ethnicities, and genders. These projects tend to use large claims databases, but big data can also come in the form of high frequency data. Our group has collected sensor mote data in a variety of settings in order to better understand how healthcare workers behave and act to transmit HAIs. This line of research has devised algorithms optimizing healthcare operations (Hasibul et al., 2021), and informed agent-based modeling (Li et al., 2024); I am leading ongoing work constructing a regional-wide agent-based model (ABM) to better understand the spread of HAIs both within and between healthcare facilities and the potential effects of various mitigation measures. This research has been presented to the CDC at a grantee meeting in 2022 and two virtual calls in 2023 and 2024.

It is obvious that networks are at the heart of ID, and they have played an integral part in, e.g., the ABMs mentioned above. As my team began to study the regional spread of HAIs,

we realized that new methodology was required to address deficiencies in the data, leading us to develop a novel network imputation method (Justice et al., 2021). More generally, ID and networks together have led to the development of new statistical methodology. I developed a novel approach to pooled testing in order to make universal screening more feasible for smaller or resource-poor organizations by leveraging information using even coarse contact network data (although technology is making high fidelity data easily obtainable) (Sewell 2022a). While this relied on a slow simulated annealing algorithm, follow-up work yielded massive increase in speed. This was achieved by building off the autologistic actor attribute model to derive an objective function which was then maximized using a constrained divisive clustering method* (Sewell 2022b). Another important methodological area where networks and ID intersect is the problem of source detection. This problem, in the realm of ID, tries to identify where an infectious disease was first introduced into a population. This is of vital importance for understanding how epidemics begin, transmission occurs, and pathogens evolve. Currently Haomin Li and I have obtained promising results on improving the current state-of-the-art source detection algorithms and are working on writing these results up.

Clustering

I am engaged in ongoing work using latent class analysis (clustering methodology for discrete data) to discover common eneteric pathogen exposure profiles across infants in low income neighborhoods in Kenya, and to determine if these exposure profiles differ in enteric infection outcomes of these children. However, most of my work in clustering since 2021 has revolved around methodological developments.

Together with a PhD student and colleague, we developed a novel Bayesian hierarchical clustering model (Burghardt et al., 2022). Our approach provides both agglomerative and divisive methods that outperformed the current state-of-the-art on benchmark datasets. In ongoing work, we have created a non-trivial extension to longitudinal data, and applied it to discovering Parkinson's Disease subtypes.

Clustering on network data is a challenging problem, in large part because there is no natural coordinate system for network data, and hence most orthodox methods cannot be applied. The following are two advances in this area. First, we applied sparse finite mixture models to a model-based edge clustering algorithm (which will be described below), using a variational generalized EM algorithm to perform estimation which automatically selects the number of clusters in the data (Pham and Sewell, 2024). Second, in ongoing work with Haomin Li, we have brought the notion of a "noise cluster" into the network realm- to my knowledge for the first time- when clustering the edges of a network (Li and Sewell, 2024+, invited revision to *CSDA*).

Network Analysis

As previously described, networks have played an integral part in my ID research, such as when we created a bipartite network of $\sim 17M$ inpatient visits to analyze the transmissibility of *C. difficile* in hospital settings (Justice et al., 2022). I have also brought network analytic tools to bear on a study of how financial incentives create and maintain collaborations between healthcare organizations (Heeren et al., 2022).

*I point out here that this particular problem involved all three focal areas of my research: Network analysis, ID, and clustering.

One blossoming avenue of research is the investigation of how healthcare teams communicate (Tu et al., 2024). I am a PI of the multi-PI study, "Connected Cancer Care: EHR Communication Networks in Virtual Cancer Care Teams" (1R01CA273058-01), which aims to understand the complex communication that occurs over the EHR system within teams and between teams focused on a patient's care (this type of system is often referred to as a team-of-teams). There are myriad methodological challenges here, starting with how to construct the network for each patient, who ought to be included in the network, and how to relate the patient-level networks to patient-level scalar outcomes.

Two works cited above- Pham and Sewell (2024), and Li and Sewell (2024+)- build off of a highly novel network modeling paradigm that focuses on the edges of the network, rather than the nodes (Sewell 2021). This lends scalability (computational cost grows $\mathcal{O}(n)$ vs. $\mathcal{O}(n^2)$, where n is the number of nodes in the network), as well as a great deal of interpretability that was previously lacking. I am continuing to build off of this work, including the ongoing source detection project with Haomin Li alluded to above.

One last area of research within the broad umbrella of network analysis is that of the network autocorrelation model (NAM). The NAM captures how networked individuals' outcomes affect one another. Li and Sewell (2021) demonstrate how the estimation of the influence parameter of the NAM is affected by network topology, and provide a comparison between two frequentist estimators and one Bayesian. In work under review, Pham and Sewell (2024+) demonstrate how homophily (the process in which similar actors form edges with one another) provides a backdoor pathway leading to severe bias that does not decay asymptotically, and proving under certain conditions the distribution to which the outcome converges. Other ongoing work with Scott Cleven is creating a new way to understand and measure network mediation which borrows ideas from geographically weighted regression.

Evidence of Research Quality

As evidence of a national and international reputation, I have been invited to give talks at the Social Networks and Beyond Conference (2022), the U.S. Centers for Disease Control and Prevention (2022, 2023, 2024), the Joint Statistical Meetings (2023, 2024), and the IMS International Conference on Statistics and Data Science (2024, but declined due to health reasons). Upon invitation I have joined the Working Group for Model-Based Clustering (2021-present), a small group of the world's experts on the titular topic. I have also been asked to participate in several webinars: Office of the Vice President for Research Webinar on COVID-19 (2020), College of Public Health Webinar on Face Coverings (2020), and the ICHI Podcast[†] (2024). I have also been invited to write an editorial on Advances in network data science for the *Journal of Data Science* (Chen et al., 2023).

I have also consistently obtained a high level of external funding: 82%, 48%, 65%, 70%, and 81% for AY2020-21, 2021-22, 2022-23, 2023-24, and 2024-present respectively. I have been PI on two of these grants. Through this external funding, I have supported 15 RA-years starting in Fall 2021.

[†]This podcast highlights articles published in Infection Control & Hospital Epidemiology.