Research Statement

Fan Bu, Ph.D.

> "In our lust for measurement, we frequently measure that which we can rather than that which we wish to measure... and forget that there is a difference." — George Udny Yules

The technological advancements of the "big data" era has brought us unprecedented access to rich data sources that take increasingly complex structures and formats. Data structures that transcend conventional tabular or matrix formats present additional challenges to data science endeavors and require advanced and specially designed methodology and techniques. Furthermore, despite evolving technologies in data collection and storage, there are still aspects of data that are not directly measurable and information that is not immediately available. Development of analytical tools should accommodate such partial observations or data missingness.

My research has thus been largely focused on statistical methodology and inference techniques for complex data structures while accounting for unobserved or latent information, mostly inspired by applications in health sciences and social sciences. Motivated by modern modalities of epidemic data sources that not only include case counts but also contact tracing (Figure 1) or viral deep sequences, I have developed models of temporal or spatial processes and dynamic networks to expand our understanding of infectious disease transmission and provide more reliable evidence base for public health efforts. Inspired by social science interests in latent patterns of interpersonal relations and information diffusion on social networks, I have adapted multivariate stochastic process models and latent factor approaches to provide in-depth insights into the dynamics of social processes.

Since joining UCLA as a postdoctoral fellow, I have further expanded my research scope to computational statistical methods for real-world evidence extraction and synthesis for solving massive-scale public health problems. Currently, I am a leading investigator in a collaborative contract between the OHDSI network (Figure 2) and FDA CBER [a] to develop Bayesian adaptive analysis methods for rapid and more reliable vaccine safety surveillance, aimed to monitoring new vaccines including COVID-19 vaccines.

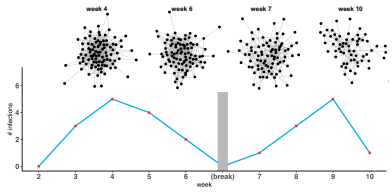

Figure 1: Example of modern epidemic data with contact tracing which uses Bluetooth tracking to record physical proximity of individuals (eX-FLU data [1]). Top panel visualizes the weekly aggregated contact networks; bottom panel plots new infection numbers obtained in weekly survey.

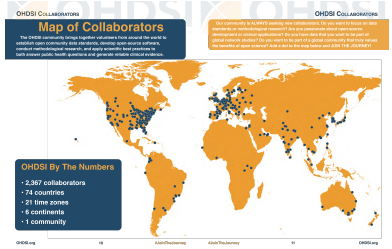

Figure 2: I am an active member of OHDSI (Observational Health Data Sciences and Informatics), a multi-national health science data and research network that aims at promoting open science and reproducible evidence discovery from large-scale observational health data. To date, OHDSI has collaborators from academia, industry and regulatory agencies in over 70 countries.

[a] The Center for Biologics Evaluation and Research (CBER) within FDA regulates biological products for human use under applicable federal laws. It conducts safety and effectiveness evaluations and monitoring for biological products (including vaccines) to ensure their safe and appropriate use.

Inference of partially observed stochastic epidemics on dynamic networks

With recent advancements in wearable device technologies and contact tracing, modern epidemiological studies have the capacity of learning transmission dynamics and evaluating intervention strategies at the individual level. Unlike traditional epidemic models (e.g., SIR or SIS models) that base on population-level aggregated counts,

I developed an individualistic framework of stochastic epidemics that depend on a dynamically evolving contact network while also accounting for contact behavior changes in response to an epidemic. Such framework flexibly utilizes individual-level disease and contact information (such as that collected in the eX-FLU study [1]) to account for heterogeneous transmission dynamics. To tackle the challenges of unobserved infection and recovery times in observational data, I developed efficient data-augmented inference algorithms using carefully designed conditional samplers that respect the constraints of dynamic networks. A fully Bayesian framework is developed in [2], and in an extended work [3], I designed a stochastic EM inference algorithm to infer individual exposure risks while accounting for disease latency based on a stochastic SEIR model (Figure 3). The extended framework can characterize the effects of preventative measures (such as vaccination and hand-washing) as well as the impact of a pandemic on social contact behaviors.

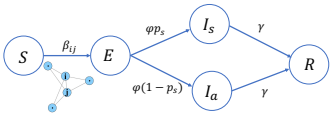

Figure 3: Diagram of an extended susceptible-exposed-infectious-recovery (SEIR) model. Compartment "E" is introduced to account for latency periods (e.g., for influenza, SARS and COVID-19), where the two "I" compartments are used to differentiate between (for example) symptomatic and asymptomatic cases. Transmission (from "S" to "E") depends on the individualistic dynamic contact structure.

Inferring HIV transmission age structure from high-throughput viral sequencing data

Modern phylogenetic analysis on high-throughput viral sequencing data provides information on the likelihoods of unobserved disease transmission links and directions between potential sources and recipients. Inference of "who infected whom" for HIV transmission is essential for identifying high-risk groups and thus prioritizing limited resources for more effective interventions. The current standard practice is to assume a fixed threshold on the phylogenetic likelihoods to draw conclusion on transmission directions without uncertainty quantification [4]. To provide a more flexible and statistically principled approach, I developed a Bayesian hierarchical spatial Poisson process model [5] to fully account for the uncertainty in HIV transmission links and directions while retaining more power from lower-likelihood data points that were discarded in the fixed threshold approach. Moreover, this framework enjoys much improved computational efficiency by assuming a continuous spatial process instead of a manually discretized model in previous work. The proposed method learns a transmission flow surface between age groups and pinpoints high-frequency infection sources for vulnerable demographic groups such as young women in Africa (Figure 4).

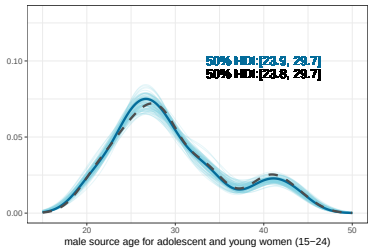

Figure 4: Age distribution of male sources for infections in young women from Rakai, Uganda between 15 to 24, learned by two variants of our proposed method. Young women were typically infected by men between 24 and 30, but there is a noticeable mass for much older men around 40, implying a "sugar daddy" phenomenon for sexual interactions and HIV transmissions. Data analysis was conducted in collaboration with the PANGEA HIV consortium [6].

Learning latent patterns of interactions in social processes

It is of great interest in social sciences to understand the underlying structures in inter-personal interactions, especially the latent patterns that cannot be explained by observed covariates. Latent factor models [7, 8, 9] provide powerful tools to capture such patterns by projecting high-dimensional social network objects into lower-dimensional vectors. In a work that studied collaboration dynamics of multi-team work groups [10], we adapted a dynamic latent factor model [11] to learn the evolution of latent social positions between team members

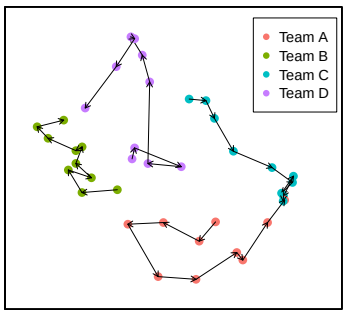

Figure 5: Learned dynamic latent positions of mutual dependence between individuals across four teams over the span of 10 weeks (dynamic changes denoted by arrows). From this visualization we can see that Teams A and C develop a closer dependence relationship over time while Teams B and D stay relatively closely dependent throughout.

and across different teams (see example in Figure 5), and how those latent factors affect team members' psychological perceptions and collaboration performance. In another work that introduced a social network metric for basketball game success [12], I combined a latent factor model with a spatio-temporal point process to describe higher-order patterns in basketball player passes that can distinguish a victory from a loss.

When inter-personal links are not directly observed, it is a challenging task to infer the relational influence and information flow between individuals in social processes. In the context of information diffusion, we developed a marked multivariate Hawkes process model to uncover the latent replying structure of online conversations by leveraging both the temporal and textual information [13]. The high-dimensional latent structure is inferred using an efficient variational Bayesian inference scheme, which can further uncover influential users in group conversations (Figure 6) and original sources of misinformation on social media.

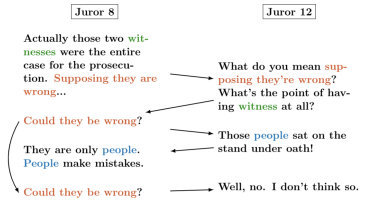

Figure 6: Example of latent conversation structure uncovered by our method, using the text transcript of the movie "12 Angry Men". Our model infers the high social influence of Juror 8 (the protagonist) by identifying how he steers the conversation flow, as implied by borrowed vocabulary (same-colored text) between him and another juror.

Ongoing and future research agenda

In collaboration with the FDA CBER BEST Initiative, I am developing Bayesian sequential analysis methods to improve vaccine safety surveillance procedures. I have proposed a joint statistical framework that can adaptively analyze large-scale sequential data without the need to pre-specify a surveillance schedule, while correcting for bias induced by systematic error in observational datasets; both are major challenges and bottlenecks faced by regulatory agencies. Results of the study [b] show that the new framework offers flexibility, accuracy and efficiency while providing more transparency and interpretability of real-world evidence (Figure 7). In a follow-up work, I will investigate theoretical properties of Bayesian sequential tests and explore unified frameworks of frequentist and Bayesian tests.

As an active member of OHDSI, I am very interested in developing statistical and computational techniques for evidence synthesis over a federated data network, as many large-scale healthcare databases are hosted at multiple locations and distributed data analysis has to be conducted with limited information shared between sites. An important direction of current and future work is to develop and adapt Bayesian hierarchical and non-parametric methods for meta analysis, especially for analysis models with partial likelihoods or likelihood-free inference, such as survival analysis (Cox regression), cohort designs and propensity score methods.

I also plan to extend my prior work on epidemic processes to unobserved contact links using branching processes or dynamic random graph models. In collaboration with fellow participants of the "Graph Limits and Processes on Networks" program at the Simons Institute, I also hope to further our theoretical understanding of the long-term behavior of pandemics and adaptive networks, in the context of pharmeceutical and non-pharmaceutical interventions.

[b] Study protocol for empirical evaluation on real-world health databases is available at `https://suchard-group.github.io/Better/Protocol.html`.

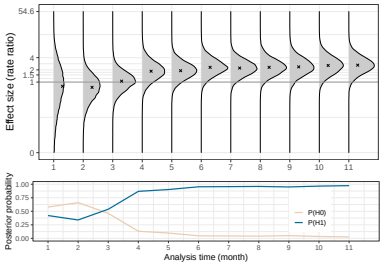

Figure 7: Example of how Bayesian posterior inference integrates sequential data evidence through distributional information on parameter of interest (top panel), which offers highly interpretable data-driven probabilities of hypotheses (bottom panel).

References

[1] A. E. Aiello, A. M. Simanek, M. C. Eisenberg, A. R. Walsh, B. Davis, E. Volz, C. Cheng, J. J. Rainey, A. Uzicanin, H. Gao, et al. Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial. *Epidemics*, 15:38–55, 2016.

[2] F. Bu, A. E. Aiello, J. Xu, and A. Volfovsky. Likelihood-based inference for partially observed epidemics on dynamic networks. *Journal of the American Statistical Association*, pages 1–17, 2020.

[3] F. Bu, A. E. Aiello, A. Volfovsky, and J. Xu. Likelihood-based inference for partially observed stochastic epidemics with individual heterogeneity. *manuscript submitted for review*, 2021+.

[4] O. Ratmann, J. Kagaayi, M. Hall, T. Golubchick, G. Kigozi, X. Xi, C. Wymant, G. Nakigozi, L. Abeler-Dörner, D. Bonsall, et al. Quantifying hiv transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in rakai, uganda. *The Lancet HIV*, 7(3):e173–e183, 2020.

[5] F. Bu, J. Xu, and O. Ratmann. Inferring hiv transmission patterns from viral deep-sequence data via latent spatial poisson processes. *Under embargo for data consortium review; full manuscript available upon request*, 2022.

[6] L. Abeler-Dörner, M. K. Grabowski, A. Rambaut, D. Pillay, and C. Fraser. Pangea-hiv 2: phylogenetics and networks for generalised epidemics in africa. *Current Opinion in HIV and AIDS*, 14(3):173, 2019.

[7] P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical association*, 97(460):1090–1098, 2002.

[8] M. Schweinberger and T. A. Snijders. Settings in social networks: A measurement model. *Sociological Methodology*, 33(1):307–341, 2003.

[9] P. Hoff. Additive and multiplicative effects network models. *Statistical Science*, 36(1):34–50, 2021.

[10] R. Asencio, F. Bu, L. Tucker, G. Varela, J. Moody, and A. Volfovsky. Latent network position and emergent phenomena: A multi-team system case study. *Under revisions*, 2022+.

[11] D. K. Sewell and Y. Chen. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657, 2015.

[12] F. Bu, S. Xu, K. Heller, and A. Volfovsky. Smogs: Social network metrics of game success. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2406–2414. PMLR, 2019.

[13] W. Zhang, F. Bu, D. Owens-Oas, K. Heller, and X. Zhu. Who started it? identifying root sources in textual conversation threads. *arXiv preprint arXiv:1809.03648*, 2018.