Research Statement

Michael Schweinberger

My research is concerned with statistical learning from discrete and dependent data without independent replications. Applications include disaster response [20], terrorist networks [17], air pollution [15], economic networks [13], online trust networks [27], brain networks [18], epidemics [16], education [9], and sports [7]. My research has been supported by NSF award DMS-1812119 (sole PI), NSF award DMS-1513644 (sole PI), DoD award ARO W911NF-21-1-0335 (lead PI), and NWO award Rubicon-44606029 (sole PI).

Selected research accomplishments 2011–present

1. **Statistical learning from discrete and dependent network and attribute data without independent replications (2011–present).** In scenarios in which the data at hand is discrete and dependent and replications are unavailable (arising in, e.g., the social sciences), it is natural to learn from data using models with conditional independence properties: graphical models in exponential-family form, with conditional distributions parameterized by generalized linear models. That being said, questions have been raised by probabilists and statisticians, starting with Handcock [8], Bhamidi et al. [1], Schweinberger [12], Fienberg [4], Chatterjee and Diaconis [3], Lauritzen et al. [10], and others. In a decade-long sequence of first-authored publications (e.g., Annals of Statistics [23], JASA [12], JRSSB [17], Bernoulli [14], Statistical Science [18], arXiv:2012.07167 [26]), I have contributed constructive answers to these questions. Among other things, I have introduced novel models with desirable properties and scalable statistical methods with theoretical guarantees. My most important contribution has been to demonstrate that the absence of desirable properties in [8, 1, 12, 4, 3, 10] can be overcome by endowing models with additional structure (observed or unobserved) and that graphical models with $p \to \infty$ parameters can be learned from a single observation from discrete and dependent data. The closest theoretical work on discrete graphical models in single-observation scenarios is Chatterjee [2] and Ghosal and Mukherjee [6], concerned with classic models in physics with $p = 1$ or $p = 2$ parameters. By contrast, my work has introduced novel discrete graphical models with $p \to \infty$ parameters in single-observation scenarios, with theoretical guarantees in scenarios with $p \to \infty$ parameters. Applications are discrete and dependent network and attribute data arising in the social sciences and other fields.

2. **The first stochastic block models with dependent edges within communities (2015–present)** [17, 14, 19]. These stochastic block models with dependent edges predate Yuan and Qu [28] by 6 years and the statistical theory in [14] predates Yuan and Qu [28] by one year. In contrast to Yuan and Qu [28], the statistical theory in [14] does not require independent replications from the same source.

Selected research accomplishments 2003–2011

3. **The first widely used temporal network models and the first joint probability models of network and attribute data (2007–2012).** Such models are of vital importance, because—more often than not—networks are of secondary interest. Of primary interest is to understand and predict how networks affect health-related, economic, social, and other outcomes. These models preceded the models of Fosdick and Hoff [5] by 8 years. These methods have been applied in hundreds, if not thousands of scientific publications. My contributions include likelihood-based inference [25], uncertainty quantification [22], and statistical tests [13].

4. **The first joint probability models of network and attribute data (2007).** Such models are of vital importance, because—more often than not—networks are of secondary interest. Of primary interest is to understand and predict how networks affect health-related, economic, social, and other outcomes. These models preceded the models of Fosdick and Hoff [5] by 8 years.

5. **One of the first two latent space models and the first statistical approach to hierarchical community detection (2003)** [21]. These methods preceded Bickel's work [11] on hierarchical community detection by 19 years: see Smith et al. [24].

Selected directions of future research

Stochastic processes involving networks, space, and time: Many real-world processes involve networks, space, and time: e.g., infectious diseases spread by way of contact, contacts depend on geographical distance, and contacts change over time. While there are existing stochastic processes indexed by networks, space, and time, many of them make either simplifying assumptions or have unknown probabilistic and statistical properties. One of my directions of future research is to design stochastic processes indexed by networks, space, and time that do justice to the complexity of network-mediated phenomena and develop scalable statistical methods for learning them from data, leveraging my decade-long research on the basics of learning from discrete and dependent data without independent replications.

Scalable selection of models of discrete and dependent data without independent replications: Developing scalable model selection procedures with theoretical guarantees is non-trivial when the likelihood function is intractable, the number of parameters is large, and the data consists of a single observation of dependent random variables. Such scenarios arise in the statistical analysis of discrete and dependent data, including network, spatial, and temporal data. As a case in point, there are many models of dependent network data, but model selection procedures are scarce and lack either computational scalability or theoretical guarantees or both. I am working on a scalable approach to model selection in dependent-data problems with intractable likelihood functions based on regularized pseudo- and composite-likelihood methods, with theoretical guarantees.

Quantifying uncertainty based on discrete and dependent data without independent replications: In applications of statistics, it is important to provide a disclaimer, acknowledging that statistical conclusions based on data are subject to error. In scenarios when the number of parameters is unbounded and a single observation of discrete and dependent random variables is available, it is not obvious how to quantify the uncertainty about statistical conclusions, because the small- and large-sample distributions of many statistical quantities are unknown. A natural approach to capturing uncertainty is a Bayesian approach. I intend to elaborate on scalable Bayesian approaches to uncertainty quantification for discrete and dependent data without independent replications and with intractable likelihood functions based on factorized objective functions (e.g., pseudo- and composite-likelihood functions), with theoretical guarantees.

Online educational assessment data: In collaboration with Minjeong Jeon (Graduate School of Education & Information Studies, University of California, Los Angeles), I am working on educational assessment data, including online educational assessment data. Among other things, we are developing statistical interaction and learning progression maps based on latent space models, with a view to providing educators with visual tools for monitoring student progress and detecting disadvantaged groups of students who need more support than other students, with applications to traditional and online educational assessments.

References

[1] Bhamidi, S., Bresler, G., and Sly, A. (2011), "Mixing time of exponential random graphs," *The Annals of Applied Probability*, 21, 2146–2170.

[2] Chatterjee, S. (2007), "Estimation in spin glasses: A first step," *The Annals of Statistics*, 35, 1931–1946.

[3] Chatterjee, S., and Diaconis, P. (2013), "Estimating and understanding exponential random graph models," *The Annals of Statistics*, 41, 2428–2461.

[4] Fienberg, S. E. (2012), "A brief history of statistical models for network analysis and open challenges," *Journal of Computational and Graphical Statistics*, 21, 825–839.

[5] Fosdick, B. K., and Hoff, P. D. (2015), "Testing and modeling dependencies between a network and nodal attributes," *Journal of the American Statistical Association*, 110, 1047–1056.

[6] Ghosal, P., and Mukherjee, S. (2020), "Joint estimation of parameters in Ising model," *The Annals of Statistics*, 48, 785–810.

[7] Grieshop, N., Feng, Y., Hu, G., and Schweinberger, M. (2023), "A continuous-time stochastic process for high-resolution network data in sports," *arXiv:2023.01318*, 1–29.

[8] Handcock, M. S. (2003), "Assessing degeneracy in statistical models of social networks," Tech. rep., Center for Statistics and the Social Sciences, University of Washington, www.csss.washington.edu/Papers.

[9] Jeon, M., Jin, I. H., Schweinberger, M., and Baugh, S. (2021), "Mapping unobserved item-respondent interactions: A latent space item response model with interaction map," *Psychometrika*, 86, 378–403.

[10] Lauritzen, S., Rinaldo, A., and Sadeghi, K. (2018), "Random networks, graphical models and exchangeability," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 481–508.

[11] Li, T., Lei, L., Bhattacharyya, S., van den Berge, K., Sarkar, P., Bickel, P. J., and Levina, E. (2020), "Hierarchical community detection by recursive partitioning," *Journal of the American Statistical Association*, 117, 951–968.

[12] Schweinberger, M. (2011), "Instability, sensitivity, and degeneracy of discrete exponential families," *Journal of the American Statistical Association*, 106, 1361–1370.

[13] Schweinberger, M. (2012), "Statistical modeling of digraph panel data: goodness-of-fit," *British Journal of Mathematical and Statistical Psychology*, 65, 263–281.

[14] Schweinberger, M. (2020), "Consistent structure estimation of exponential-family random graph models with block structure," *Bernoulli*, 26, 1205–1233.

[15] Schweinberger, M., Babkin, S., and Ensor, K. B. (2017), "High-dimensional multivariate time series with additional structure," *Journal of Computational and Graphical Statistics*, 26, 610–622.

[16] Schweinberger, M., Bomiriya, R. P., and Babkin, S. (2022), "A semiparametric Bayesian approach to epidemics, with application to the spread of the coronavirus MERS in South Korea in 2015," *Journal of Nonparametric Statistics*, 34, 628–662.

[17] Schweinberger, M., and Handcock, M. S. (2015), "Local dependence in random graph models: characterization, properties and statistical inference," *Journal of the Royal Statistical Society, Series B*, 77, 647–676.

[18] Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. R. (2020), "Exponential-family models of random graphs: Inference in finite, super, and infinite population scenarios," *Statistical Science*, 35, 627–662.

[19] Schweinberger, M., and Luna, P. (2018), "HERGM: Hierarchical exponential-family random graph models," *Journal of Statistical Software*, 85, 1–39.

[20] Schweinberger, M., Petrescu-Prahova, M., and Vu, D. Q. (2014), "Disaster response on September 11, 2001 through the lens of statistical network analysis," *Social Networks*, 37, 42–55.

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

[22] Schweinberger, M., and Snijders, T. A. B. (2007), "Markov models for digraph panel data: Monte Carlo-based derivative estimation," *Computational Statistics and Data Analysis*, 51, 4465—4483.

[23] Schweinberger, M., and Stewart, J. R. (2020), "Concentration and consistency results for canonical and curved exponential-family models of random graphs," *The Annals of Statistics*, 48, 374–396.

[24] Smith, A. L., Asta, D. M., and Calder, C. A. (2019), "The geometry of continuous latent space models for network data," *Statistical Science*, 34, 428–453.

[25] Snijders, T. A. B., Koskinen, J., and Schweinberger, M. (2010), "Maximum likelihood estimation for social network dynamics," *The Annals of Applied Statistics*, 4, 567–588.

[26] Stewart, J. R., and Schweinberger, M. (2021), "Pseudo-likelihood-based M-estimators for random graphs with dependent edges and parameter vectors of increasing dimension," Tech. rep., Department of Statistics, Florida State University, https://arxiv.org/abs/2012.07167.

[27] Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013), "Model-based clustering of large networks," *The Annals of Applied Statistics*, 7, 1010–1039.

[28] Yuan, Y., and Qu, A. (2021), "Community detection with dependent connectivity," *The Annals of Statistics*, 49, 2378 – 2428.