Research Statement

Michael Schweinberger

Since the pioneering work of K. Pearson, R.A. Fisher, C.R. Rao, J. Neyman and others, the bulk of statistical research has focused on attributes of individual population members and scenarios in which replication is possible. In more recent times, a mounting body of evidence has revealed that the world of the twenty-first century is interconnected and interdependent, underscored by recent events that started out as local problems and turned into global crises (e.g., pandemics, political and military conflicts, economic and financial crises). More often than not, such events are unique and cannot be replicated, and the data at hand are discrete and dependent. Despite the fact that the interconnected world of the twenty-first century affects the welfare of billions of people around the world, **statistical learning with theoretical guarantees from discrete and dependent network and attribute data without independent replications is an underresearched area.** My research focuses on statistical learning in these challenging scenarios.

Selected research accomplishments

Statistical learning from discrete and dependent network and attribute data without independent replications. Understanding and predicting how the interconnected and interdependent world of the twenty-first century operates and affects the welfare of billions of people around the world requires statistical procedures for learning from network and attribute data. More often than not, network and attribute data are discrete and dependent, and independent replications of networks are unavailable. In such scenarios, it is natural to base statistical learning on models with conditional independence properties, which facilitate theoretical guarantees for statistical procedures in single-observation scenarios. In other words, it is natural to base statistical learning on graphical models which possess conditional independence properties by construction, and admit exponential-family representations of joint distributions and GLM-representations of conditional distributions. Such models can be viewed as generalizations of GLMs for dependent network and attribute data and are widely used in practice, implemented in more than 20 `R` packages and downloaded more than 2.5 million times from the `RStudio` CRAN server alone. Having said that, some of the world's leading probabilists and statisticians have expressed concern about the probabilistic behavior of such models and whether statistical learning is possible based on a single observation of discrete and dependent network and attribute data [e.g., 8, 1, 5, 3, 11]. In a decade-long sequence of single- and first-authored publications starting in 2011 (e.g., Annals of Statistics [28], JASA [15], JRSSB [22], Bernoulli [18], Statistical Science [23], arXiv:2012.07167 [31]), I have taken steps to address these concerns and demonstrate how well-behaved models can be built, and that statistical learning based on a single observation of discrete and dependent network and attribute data is possible without sacrificing computational scalability and theoretical guarantees. My most important contribution has been to demonstrate that the absence of desirable properties in [8, 1, 15, 5, 3, 11] 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, without sacrificing computational scalability and theoretical guarantees. By comparison, the small body of existing statistical theory for discrete graphical models in single-observation scenarios assumes that the number of parameters p is fixed and makes other assumptions that limit the scope of the theoretical results to classic models in physics, e.g., Gibbs measures and Ising models with $p = 1$ or $p = 2$ parameters [e.g., 14, 4, 2, 6]. By contrast, my research focuses on large classes of discrete graphical models with $p \to \infty$ parameters, with theoretical guarantees in single-observation scenarios and applications in the social sciences, the health sciences, and other fields.

The design of stochastic models for complex real-phenomena along with statistical methods: e.g., air pollution [20], disaster response [25], epidemics [21], online trust networks [32], sport games (e.g., soccer games) [7], terrorist networks [22], brain networks [23], financial ownership networks [19], mental health [10], online educational assessments [10, 9], and geographical disparities in terms of income [13].

The first stochastic block models with dependent edges within communities [22, 18, 24]. These stochastic block models with dependent edges predate Yuan and Qu [33]. In contrast to Yuan and

Qu [33], [22, 18, 24] do not rely on independent replications.

The first widely used temporal network models and the first joint probability models of network and attribute data [e.g., 27, 30, 16]. These models have been applied in hundreds, if not thousands of scientific publications. My contributions include likelihood-based inference [30], uncertainty quantification [27], and statistical tests [16].

One of the first two latent space models and the first statistical approach to hierarchical community detection [26]. These methods preceded Bickel's work [12] on hierarchical community detection by 19 years [29, 17].

Selected directions of future research

Scalable joint probability models of discrete and dependent network and attribute data, capturing non-causal and causal relationships. Joint probability models of discrete and dependent network and attribute data help answer questions about non-causal and causal relationships among attributes under network interference. I am working on a joint probability modeling framework for discrete and dependent network and attribute data, which is (a) flexible, in the sense that it can capture a wide range of attribute-attribute, attribute-connection, and connection-connection dependencies; (b) interpretable, in that it builds on the proven statistical platform of GLMs, facilitating interpretation and dissemination; and (c) scalable, in the sense that it allows large populations to be more heterogeneous than small populations and can capture interesting forms of dependence among attributes and connections in large populations. These joint probability models provide a statistical platform for studying non-causal and causal relationships among attributes of population members under network interference, including spill-over effects.

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.

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.

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] Comets, F. (1992), "On consistency of a class of estimators for exponential families of Markov random fields on the lattice," *The Annals of Statistics*, 20, 455–468.

[5] 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.

[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), "Statistical Models for Social Networks: Inference and Degeneracy," in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, eds. Breiger, R., Carley, K., and Pattison, P., Washington, D.C.: National Academies Press, pp. 1–12.

[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] Jeon, M., and Schweinberger, M. (2023), "A latent process model for monitoring progress towards hard-to-measure targets, with applications to mental health and online educational assessments," *arXiv:2305.09804*.

[11] 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.

[12] Li, T., Lei, L., Bhattacharyya, S., van den Berge, K., Sarkar, P., Bickel, P. J., and Levina, E. (2022), "Hierarchical community detection by recursive partitioning," *Journal of the American Statistical Association*, 117, 951–968.

[13] Nandy, S., Holan, S. H., and Schweinberger, M. (2023), "A socio-demographic latent space approach to spatial data when geography is important but not all-important," *arXiv:2304.03331*.

[14] Pickard, D. K. (1987), "Inference for discrete Markov fields: The simplest non-trivial case," *Journal of the American Statististical Association*, 82, 90–96.

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

[16] — (2012), "Statistical modeling of digraph panel data: goodness-of-fit," *British Journal of Mathematical and Statistical Psychology*, 65, 263–281.

[17] — (2019), "Random graphs," in *Wiley StatsRef: Statistics Reference Online*, eds. Everitt, B., Molenberghs, G., Piegorsch, W., Ruggeri, F., Davidian, M., and Kenett, R., Wiley, pp. 1–11.

[18] — (2020), "Consistent structure estimation of exponential-family random graph models with block structure," *Bernoulli*, 26, 1205–1233.

[19] — (2020), "Statistical inference for continuous-time Markov processes with block structure based on discrete-time network data," *Statistica Neerlandica*, 74, 342–362.

[20] 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.

[21] 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.

[22] 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 (Statistical Methodology)*, 77, 647–676.

[23] 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.

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

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Stewart, J. R., and Schweinberger, M. (2022), "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.

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

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