Research Statement

Michael Schweinberger

Since the pioneering work of 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 connections and outcomes without independent replications is an underresearched area.** My research focuses on statistical learning in these challenging scenarios and attempts to bridge the gap between statistical theory and the social sciences, by providing interpretable models for dependent data along with statistical theory, with a view to studying non-causal or causal relationships among interdependent predictors and outcomes under network interference.

Selected research accomplishments

Statistical learning from discrete and dependent connections and outcomes without independent replications. To learn how the interconnected world of the twenty-first century affects individual and collective outcomes of interest, one needs to learn from connections and outcomes. More often than not, such data are discrete and dependent, and independent replications may be unavailable. In such scenarios, it is natural to base statistical learning on interpretable graphical models that possess conditional independence properties by construction and admit exponential-family representations of conditional or joint distributions. Such models can be viewed as extensions of regression models for dependent connections and outcomes and are widely used in practice, implemented in more than twenty `R` packages and downloaded more than three 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 connections and outcomes [see, e.g., 13, 2, 7, 4, 16].

In a decade-long and continuing sequence of single- and first-authored publications starting in 2011 (e.g., JASA [20], JRSSB [27], Annals of Statistics [33], Bernoulli [22], Statistical Science [28], arXiv:2012.07167 [38], arXiv:2410.07555 [9]), I have taken steps to address these concerns. Among other things, I have demonstrated that the absence of desirable properties of the models considered by [13, 2, 20, 7, 4, 16] can be overcome by leveraging additional structure (observed or unobserved) [27, 33, 38, 26]. In addition, I have shown that graphical models with $p \to \infty$ parameters can be learned based on a single observation of discrete and dependent data, without sacrificing computational scalability and theoretical guarantees [38, 33]. 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 restrictive assumptions that limit the scope of the theoretical results to classic models in physics, e.g., Ising models with $p = 1$ or $p = 2$ parameters [e.g., 19, 5, 3, 11]. By contrast, my research focuses on large classes of discrete graphical models with $p \to \infty$ parameters, which come with the benefit of theoretical guarantees in single-observation scenarios and help study how the interconnected world affects individual and collective outcomes of interest.

I developed the first stochastic block models with dependent edges [27, 22, 29, 1, 29, 10, 8]. Stochastic block models are widely use for learning from network data who is close to whom. Stochastic block models with dependent edges within communities, first introduced in my 2015 publication [27], combine the advantages of stochastic block models (capturing who is close to whom) and generalized linear models for dependent connections and outcomes (capturing local dependencies among connections and attributes).

I developed one of the first two latent space models and the first statistical approach to hierarchical community detection [31]. Latent space models are popular alternatives to stochastic block models for learning from network data who is close to whom.

I made key contributions to the first widely used temporal network models and the first joint probability models of connections and outcomes [e.g., 32, 34, 21, 17, 23, 36]. These models have been used in hundreds of publications in the social and health sciences for learning whether similar behavior among connected individuals (e.g., substance abuse among friends) is due to (a) the influence of friends, (b) the tendency to select similar others as friends, or (c) both.

To gain insight into the interconnected world of the twenty-first century, I have helped data scientists design stochastic models that do justice to the complexity of real-phenomena, e.g., hate speech on social media [9], mental health [15], substance abuse [35], disaster response [30], epidemics [25], air pollution [24], online trust networks [39], product recommendation [1], online educational assessments [15, 14], soccer games [12], terrorist networks [27], brain networks [28], financial networks [21, 23], socio-economic segregation [18], and vulnerability in software networks [8].

Selected directions of future research

Causal inference under interference. At the heart of science is the question of cause and effect. I intend to work on causal inference under interference. Interference arises when the outcomes of units are affected by the treatments or outcomes of other units. The resulting phenomenon is known as spillover: Treating a subset of units may affect the outcomes of untreated units, in addition to the outcomes of treated units. Understanding spillover is imperative in real-world applications, ranging from economics and the social sciences to medicine. My research team focuses on two challenging questions without answers:

- **Black boxes:** How can the indirect causal effect be characterized as an explicit mathematical function of model parameters, when outcomes are dependent due to both treatment and outcome spillover?
- **External validity:** How can conclusions based on a sample of outcomes be generalized to the population of interest, when the outcomes are dependent due to both treatment and outcome spillover and, therefore, *what we observe* depends on *what we do not observe*?

Scalable joint probability models of discrete and dependent connections and outcomes, capturing non-causal and causal relationships. Joint probability models of discrete and dependent connections and outcomes help answer questions about non-causal and causal relationships among attributes under network interference. I intend to work on a joint probability modeling framework for discrete and dependent connections and outcomes, 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 regression models, 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.

Any question about statistical procedures of attribute data can be asked about network and attribute data—many of them do not have answers. 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 intend to work on a scalable approach to model selection in dependent-data problems with intractable likelihood functions (using, e.g., pseudo-likelihood Dantzig selectors).

Quantifying uncertainty. In applications, 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 uncertainty, because the small- and large-sample distributions of many statistical quantities are unknown. To place uncertainty quantification on sound mathematical grounds, Berry-Esseen-type bounds for bounding the error of normality approximations for dependent data are needed. Few Berry-Esseen-type bounds for dependent data exist. Worse, all existing Berry-Esseen-type bounds impose strong restrictions on dependence, such as strong forms of local dependence [37, Theorem 2.5] or strong mixing conditions [6, Theorem 3.27, p. 34]. These restrictions are too strong in many applications. My research team intends to develop Berry-Esseen-type bounds under weaker restrictions on dependence.

Stochastic processes involving networks, space, and time. Many real-world processes involve networks, space, and time. I intend to help data scientists design stochastic processes involving networks, space, and time that do justice to the complexity of an interconnected world.

References

[1] Babkin, S., Stewart, J. R., Long, X., and Schweinberger, M. (2020), "Large-scale estimation of random graph models with local dependence," *Computational Statistics & Data Analysis*, 152, 1–19.

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

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

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

[5] 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.

[6] Dedecker, J., Doukhan, P., Lang, G., Leon, J. R., Louhichi, S., and Prieur, C. (eds.) (2007), *Weak Dependence: With Examples and Applications*, Springer-Verlag.

[7] 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.

[8] Fritz, C., Georg, C.-P., Mele, A., and Schweinberger, M. (2024), "A Strategic Model of Software Dependency Networks," in *The 25th ACM Conference on Economics and Computation (EC '24)*.

[9] Fritz, C., Schweinberger, M., Bhadra, S., and Hunter, D. R. (2024), "A regression framework for studying relationships among attributes under network interference," *arXiv:2410.07555*.

[10] Fritz, C., Schweinberger, M., Komatsu, S., Martínez Dahbura, J. N., Nishida, T., and Mele, A. (2024), *bigergm: Fit, Simulate, and Diagnose Hierarchical Exponential-Family Models for Big Networks*, R package version 1.2.1.

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

[12] Grieshop, N., Feng, Y., Hu, G., and Schweinberger, M. (2024), "A continuous-time stochastic process for high-resolution network data in sports," *Statistica Sinica*, to appear.

[13] 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.

[14] 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.

[15] Jeon, M., and Schweinberger, M. (2024), "A latent process model for monitoring progress towards hard-to-measure targets, with applications to mental health and online educational assessments," *The Annals of Applied Statistics*, 18, 2123–2146.

[16] 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.

[17] Lospinoso, J., Schweinberger, M., Snijders, T. A. B., and Ripley, R. (2011), "Assessing and accounting for time heterogeneity in stochastic actor oriented models," *Advances in Data Analysis and Classification*, 5, 147–176.

[18] 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*, submitted to *The Annals of Applied Statistics*.

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

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

[21] — (2012), "Statistical modeling of digraph panel data: goodness-of-fit," *British Journal of Mathematical and Statistical Psychology*, 65, 263–281.

[22] — (2020), "Consistent structure estimation of exponential-family random graph models with block structure," *Bernoulli*, 26, 1205–1233.

[23] — (2020), "Statistical inference for continuous-time Markov processes with block structure based on discrete-time network data," *Statistica Neerlandica*, 74, 342–362.

[24] 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.

[25] 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.

[26] Schweinberger, M., and Fritz, C. (2023), "Invited discussion of "A tale of two datasets: Representativeness and generalisability of inference for samples of networks" by P.N. Krivitsky, P. Coletti, and N. Hens," *Journal of the American Statistical Association*, 118, 2225—-2227.

[27] 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.

[28] 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.

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

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] Snijders, T. A. B., Steglich, C. E. G., and Schweinberger, M. (2007), "Modeling the co-evolution of networks and behavior," in *Longitudinal models in the behavioral and related sciences*, eds. van Montfort, K., Oud, H., and Satorra, A., Lawrence Erlbaum, pp. 41–71.

[36] Snijders, T. A. B., Steglich, C. E. G., Schweinberger, M., and Huisman, M. (2010), *Manual for Siena 3.0*, Department of Statistics, University of Oxford, UK.

[37] Stewart, J. R. (2024), "Rates of convergence and normal approximations for estimators of local dependence random graph models," *arXiv:2404.11464*.

[38] Stewart, J. R., and Schweinberger, M. (2023), "Pseudo-likelihood-based M-estimators for random graphs with dependent edges and parameter vectors of increasing dimension," *arXiv:2012.07167*.

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