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 (X_i, Y_i) of population members i and scenarios in which $n \geq 2$ independent observations $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{ind}}{\sim} \mathbb{P}$ from the same source $\mathbb{P}$ are available. 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 attributes (X_i, Y_i) and connections $Z_{i,j}$ is an underresearched area.**

My research seeks to bridge the gap between statistical theory and economics, the social and health sciences, and other fields, by providing interpretable models for dependent data along with statistical theory, with a view to studying non-causal or causal relationships among dependent attributes (X_i, Y_i) under network interference $Z_{i,j}$.

Selected research accomplishments, past and present

Statistical learning from discrete and dependent attributes (X_i, Y_i) under network interference $Z_{i,j}$. To learn how the interconnected world of the twenty-first century affects individual and collective outcomes of interest, one needs to learn from data on attributes (X_i, Y_i) and connections $Z_{i,j}$. More often than not, such data are discrete and dependent. In such scenarios, it is natural to base statistical learning on interpretable models that possess conditional independence properties and admit exponential-family representations of conditional or joint distributions. Such models can be viewed as extensions of linear and generalized linear models for dependent attributes (X_i, Y_i) and connections $Z_{i,j}$. That said, some of the world's leading probabilists and statisticians have expressed concern about the probabilistic behavior of simplistic versions of such models for dependent connections $Z_{i,j}$ and whether statistical learning is possible based on a single observation of dependent connections $Z_{i,j}$ [e.g., 14, 2, 8, 4, 20].

In a decade-long and continuing sequence of lead-authored publications starting in 2011 (e.g., JASA [25], JRSSB [33], Annals of Statistics [39], Bernoulli [28], Statistical Science [34], arXiv:2012.07167 [45], arXiv:2410.07555 [10]), 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 [14, 2, 25, 8, 4, 20] can be overcome by leveraging additional structure (observed or unobserved) [33, 39, 45, 32]. In addition, I have shown that models for discrete and dependent data 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 [45, 39]. By comparison, the small body of existing statistical theory for models of discrete and dependent data 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., 24, 5, 3, 12]. By contrast, my research focuses on large classes of models of discrete and dependent data with $p \to \infty$ parameters, which come with the benefit of theoretical guarantees and help study how the interconnected and interdependent world affects individual and collective outcomes of interest.

I developed the first stochastic block models with dependent edges [33, 28, 35, 1, 35, 11, 9]. Stochastic block models are widely used for learning from network data who is close to whom. Stochastic block models with dependent edges within communities, first introduced in my 2015 publication [33], combine the advantages of stochastic block models (capturing who is close to whom) and regression models for dependent connections and outcomes (capturing local dependencies among connections). My research team has developed scalable computational-statistical methods [1, 46, 35], implemented in `R` packages `hergm` [35], `lighthergm` [6], and `bigergm` [11], along with statistical theory [28]. The Japanese company Sansan Inc. applied these methods to professional networks with ∼240,000 members [6].

I developed one of the first latent space models and the first statistical approach to hierarchical community detection [37]. Latent space models are popular alternatives to stochastic block models for learning from network data who is close to whom [16, 40, 27]. The ultrametric latent space models I introduced in [37] have intrinsic hierarchical structure and can be used for hierarchical community detection. I published them one year after the Euclidean latent space models of Hoff et al. [15], seven years before the hyperbolic latent space models with intrinsic hierarchical structure of Krioukov et al. [19], and nineteen years before the hierarchical community detection method of [21] [see, e.g., 40, 27].

I made core contributions to the first widely used temporal network models and the first joint probability models of connections and outcomes [e.g., 38, 41, 26, 22, 29, 43]. 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. My contributions include likelihood-based inference [41], uncertainty quantification [38], statistical tests [26, 22], latent variable models [29], and statistical software [43].

To gain insight into the interconnected and interdependent world of the twenty-first century, I design stochastic models of real-world phenomena, e.g., hate speech on social media [10], mental health [18], epidemics [31], air pollution [30], disaster response [36], terrorist networks [33], systemic risk in software networks [9], online trust networks [46], offline and online educational assessments [18, 17], product recommendation [1], financial networks [26, 29], soccer games [13], brain networks [34], substance abuse [42], socio-economic segregation [23], and other real-world phenomena.

Selected directions of future research

Causal inference under interference. At the heart of science is the question of cause and effect. I am interested in causal inference for attributes (X_i, Y_i) under interference $Z_{i,j}$. Interference arises when the outcome Y_i of a given unit i is affected by the treatments X_j or outcomes Y_j of other units j connected to unit i ($Z_{i,j} = 1$), in addition to the treatment X_i assigned to unit i. 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. Two forms of spillover can be distinguished: treatment spillover (X_i and X_j affect Y_i) and outcome spillover (X_i and Y_j affect Y_i). Most research has focused on treatment spillover, which implies that the outcomes $Y_1, \ldots, Y_n$ are *independent conditional on the treatments* $X_1, \ldots, X_n$. I am interested in both treatment and outcome spillover, which implies that the outcomes $Y_1, \ldots, Y_n$

are *dependent conditional on the treatments* $X_1, \ldots, X_n$. Understanding treatment and outcome spillover is imperative in real-world applications in the life sciences and social sciences. My research team intends to answer the following two questions, among other questions:

(a) Black boxes. How can the indirect causal effect be characterized as an explicit mathematical function of the effect of treatment, the effect of treatment spillover, and the effect of outcome spillover, when outcomes are dependent conditional on treatments? While answers exist in the special case when there is treatment spillover but there is no outcome spillover, no answers exist when outcomes are dependent due to outcome spillover.

(b) 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?

Both (a) and (b) have in common that outcomes are dependent conditional on treatments, and no answers are known. The advances of my research team during decade has made it possible to obtain answers in these challenging scenarios.

Any question about statistical procedures of attribute data (X_i, Y_i) can be asked about dependent attributes (X_i, Y_i) and connections $Z_{i,j}$. Many of these questions have few if any questions. My research team intends to answer them.

(a) Model selection. As a case in point, there are countless models of discrete and dependent data, but model selection procedures are scarce and lack either computational scalability or theoretical guarantees or both. My research team plans to work on a scalable approach to model selection in dependent-data problems with intractable likelihood functions. We intend to explore two directions of research, one based on pseudo-likelihood Dantzig selectors and the other one based on pseudo-likelihood Bayesian procedures.

(b) Uncertainty quantification. 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 and statistical tests on firm mathematical grounds, Berry-Esseen-type bounds for bounding the error of normality approximations for dependent data are needed. That said, 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 [44, Theorem 2.5] or strong mixing conditions [7, Theorem 3.27, p. 34]. These restrictions are too strong in a wide range of applications. My research team intends to develop Berry-Esseen-type bounds under weaker restrictions on dependence.

Stochastic models of network-space-time data. 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 and interdependent world, expanding my work on stochastic models of network-space and network-time data to network-space-time data.

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] Dahbura, J. N. M., Komatsu, S., Nishida, T., and Mele, A. (2021), “A structural model of business card exchange networks,” *arXiv:2105.12704*, 1–33.

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

[8] 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.

[9] Fritz, C., Georg, C.-P., Mele, A., and Schweinberger, M. (2024), “A Strategic Model of Software Dependency Networks,” in *25th ACM (Association for Computing Machinery) Conference on Economics and Computation (EC ’24)*.

[10] 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*.

[11] 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.

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

[13] 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.

[14] 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.

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

[16] Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012), "Computational statistical methods for social network models," *Journal of Computational and Graphical Statistics*, 21, 856–882.

[17] 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.

[18] 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.

[19] Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguna, M. (2010), "Hyperbolic geometry of complex networks," *Physical Review E*, 82.

[20] 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.

[21] 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.

[22] 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.

[23] Nandy, S., Holan, S. H., and Schweinberger, M. (2024), "A socio-demographic latent space approach to spatial data when geography is important but not all-important," *arXiv:2304.03331*.

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

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

[26] — (2012), "Statistical modeling of digraph panel data: goodness-of-fit," *British Journal of Mathematical and Statistical Psychology*, 65, 263–281.

[27] — (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.

[28] — (2020), "Consistent structure estimation of exponential-family random graph models with block structure," *Bernoulli*, 26, 1205–1233.

[29] — (2020), “Statistical inference for continuous-time Markov processes with block structure based on discrete-time network data,” *Statistica Neerlandica*, 74, 342–362.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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*.

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