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 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 world of the twenty-first century affects individual and collective outcomes of interest 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 are unavailable. In such scenarios, it is natural to base statistical learning on graphical models that possess conditional independence properties by construction and admit exponential-family representations of joint distributions and generalized linear model-representations of conditional distributions. Such models can be viewed as extensions of generalized linear models for dependent network and attribute data and are widely used in practice, implemented in more than twenty `R` packages and downloaded more than two 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 [see, e.g., 10, 2, 7, 4, 15].

In a decade-long sequence of single- and first-authored publications starting in 2011 (e.g., Annals of Statistics [33], JASA [20], JRSSB [27], Bernoulli [23], Statistical Science [28], arXiv:2012.07167 [38]), 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 [10, 2, 20, 7, 4, 15] can be overcome by leveraging additional structure (observed or unobserved). 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, 8]. 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 of the twenty-first century affects individual and collective outcomes of interest.

I developed the first stochastic block models with dependent edges within communities [27, 23, 29, 1]. 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 [27], combine the advantages of stochastic block models (capturing who is close to whom) and generalized linear models for dependent network and attribute data (capturing local dependencies among connections and attributes). My research team has developed scalable computational-statistical methods [1, 39, 29], implemented in `R` packages `hergm` [29] and `lighthergm` [6]. Dahbura et al. [6, Sansan Inc., Japan] applied them to professional networks with ∼670,000 business connections among ∼240,000 members [6].

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. The ultrametric latent space models I introduced [31] have intrinsic hierarchical structure and can be used for hierarchical community detection. I published them one year after the Euclidean latent space models by Hoff et al. [11], seven years before the hyperbolic latent space models with intrinsic hierarchical structure by [14], and nineteen years before the hierarchical community detection method by [16] [see, e.g., 34, 22].

I made key contributions to the first widely used temporal network models and the first joint probability models of network and attribute data [e.g., 32, 35, 21]. These models are widely used in the social and health sciences for learning whether similar behavior among connected individuals (e.g., substance abuse among friends) is due to (1) the influence of friends, (2) the propensity of individuals to select similar others as friends, or (3) both. My contributions include likelihood-based inference [35], uncertainty quantification [32], statistical tests [21, 17], latent variable models [24], and software [37].

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., air pollution [25], disaster response [30], epidemics [26], mental health [13], substance abuse [36], online trust networks [39], product recommendation [1], online educational assessments [13, 12], soccer games [9], terrorist networks [27], brain networks [28], ownership networks [21, 24], and spatial segregation [18].

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 generalized linear 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.

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 (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. 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 intractable likelihood functions using pseudo-likelihoods, 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. I plan 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] Dahbura, J. N. M., Komatsu, S., Nishida, T., and Mele, A. (2021), "A structural model of business card exchange networks," Available at https://arxiv.org/abs/2105.12704.

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

[9] 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*, submitted to *Statistica Sinica*.

[10] 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.

[11] 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.

[12] 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.

[13] 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*, revised and resubmitted to *The Annals of Applied Statistics*.

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

[15] 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.

[16] 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.

[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 *Journal of the American Statistical Association*.

[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] — (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.

[23] — (2020), "Consistent structure estimation of exponential-family random graph models with block structure," *Bernoulli*, 26, 1205–1233.

[24] — (2020), "Statistical inference for continuous-time Markov processes with block structure based on discrete-time network data," *Statistica Neerlandica*, 74, 342–362.

[25] 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.

[26] 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.

[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] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Stewart, J. R., and Schweinberger, M. (2022), "Pseudo-likelihood-based M-estimators for random graphs with dependent edges and parameter vectors of increasing dimension," *https://arxiv.org/abs/2012.07167*, invited major revision by *The Annals of Statistics*.

[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.