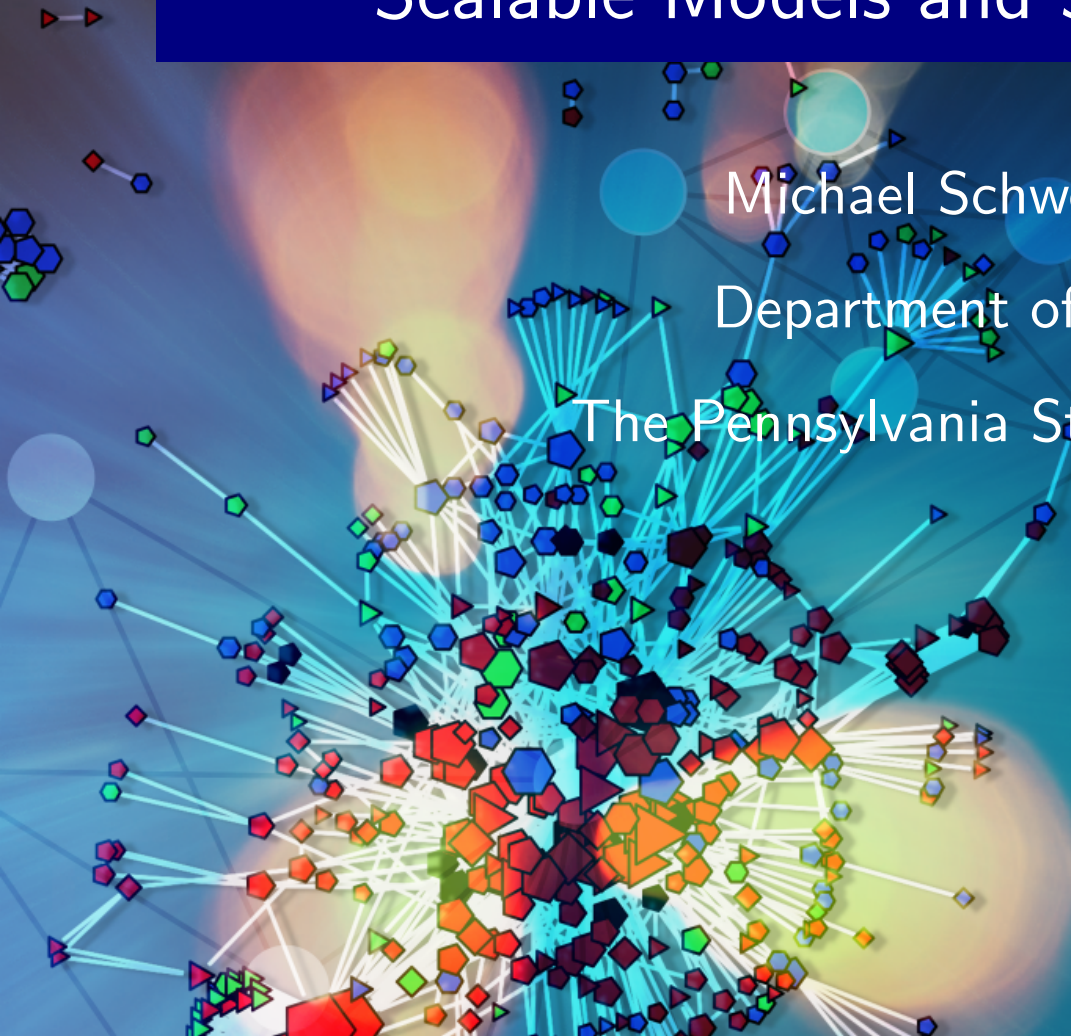
Large Networks: Scalable Models and Scalable Methods

Michael Schweinberger

Department of Statistics

The Pennsylvania State University



Outline

Structure

Motivation

Scalable models

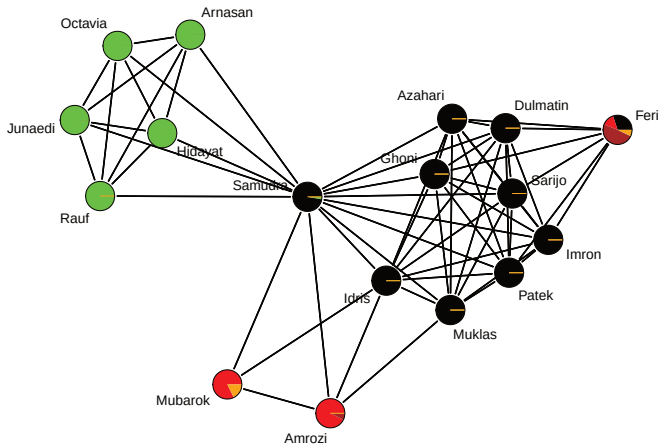
- Advantage: theoretical guarantees
- Advantage: scalable methods

Scalable methods

Software

Motivation

Before the age of data science: small networks with dozens of members (e.g., terrorist network behind Bali bombing in 2002):



Age of data science: large networks with thousands or millions of members (e.g., hate speech on Twitter (X)):

What do we know about large networks?

- Large populations are **heterogeneous.**
- **Some population members are closer than others.**
- Large population networks are **sparse.**
- **Connections** depend on other connections, but **do not depend on all other connections.**

In large populations, populations members are not aware of the attributes and connections of most other population members and therefore cannot be directly affected by them.

How can we study large networks?

We need

- scalable models
- scalable methods.

The bulk of the literature focuses on *scalable methods*, while ignoring the importance of *scalable models.*

The importance of scalable models

Lessons from model degeneracy and other undesirable properties of classic models (e.g., Frank and Strauss 1986):

- **Model degeneracy stems from a lack of structure:** Without additional structure, it is challenging to build scalable models that are well-behaved in small and large networks.
- **Additional structure** is important in practice and theory, because it **helps construct scalable models.**
- **The statistical analysis of large networks requires scalable models, in addition to scalable methods.**

Schweinberger (Journal of the American Statistical Association, Theory & Methods, 2011)

Structure

Motivation

Scalable models

- Advantage: theoretical guarantees
- Advantage: scalable methods

Scalable methods

Software

Notation

Define

$$X_{i,j} := \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases}$$

Extensions to directed and weighted networks are possible.

Scalable models with local dependence

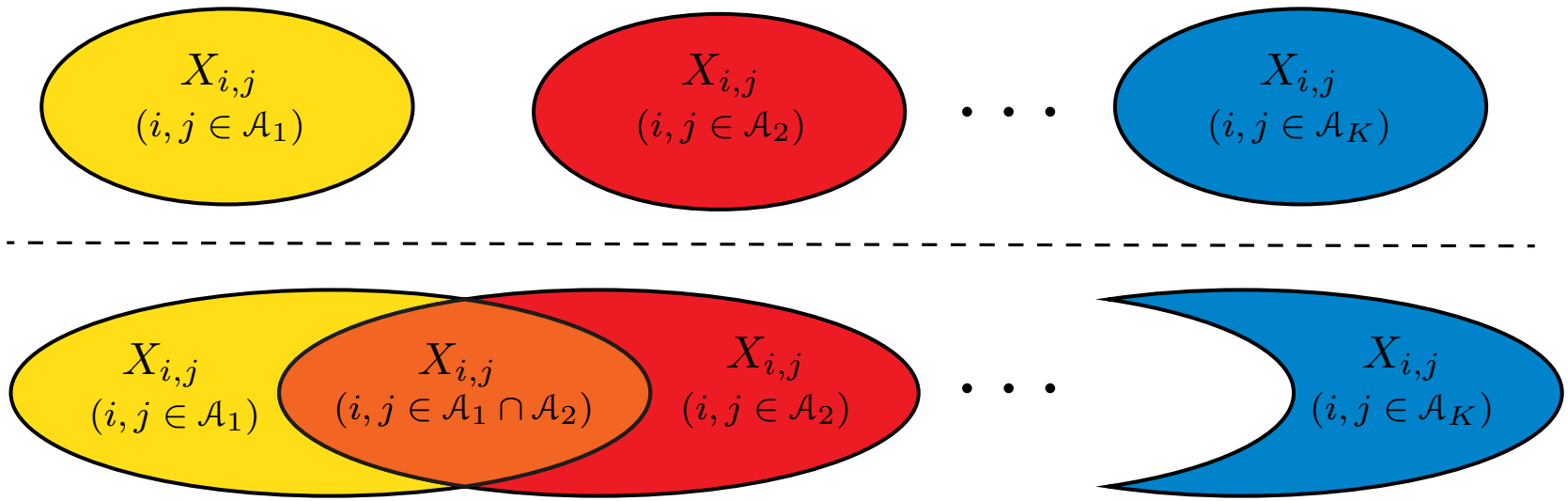
A probability law $\mathbb{P}$ governing connection indicators $X_{i,j}$ induces local dependence if, for each pair of members $\{i, j\}$, the conditional probability of the event $\{X_{i,j} = 1\}$ depends on the connections among a subset of other members:

$$\mathbb{P}(X_{i,j} = 1 \mid \text{universe}) \quad = \quad \mathbb{P}(X_{i,j} = 1 \mid \text{subset of universe})$$

Schweinberger and Handcock (Journal of the Royal Statistical Society, Series B, 2015)

Scalable models with local dependence

The construction of models with local dependence is facilitated by **additional structure,** e.g., $K \geq 2$ subpopulations $\mathcal{A}_1, \ldots, \mathcal{A}_K$:



- Observed: e.g., multilevel networks.
- Unobserved: learned from data.

Scalable models with local dependence

Scalable models with local dependence leveraging additional structure: subpopulation structure

A probability law $\mathbb{P}$ governing connection indicators $X_{i,j}$ in a population consisting of $K \geq 2$ non-overlapping subpopulations $\mathcal{A}_1, \ldots, \mathcal{A}_K$ induces local dependence if

$$\mathbb{P}(\boldsymbol{X}) = \prod_{k=1}^{K} \mathbb{P}_{k,k}(\boldsymbol{X}_{\mathcal{A}_k,\mathcal{A}_k}) \prod_{l=k+1}^{K} \mathbb{P}_{k,l}(\boldsymbol{X}_{\mathcal{A}_k,\mathcal{A}_l})$$

where the dependence is restricted to connections within the same subpopulation.

The building blocks $\mathbb{P}_{k,k}$ and $\mathbb{P}_{k,l}$ can be parameterized by generalizations of logistic regression models (ERGMs), which have game-theoretic foundations: see Butts (2009) and Mele (2017).

Extensions

More exciting developments down the road:

- Models with overlapping subpopulations ("overlapping social circles"): Stewart and Schweinberger (Annals of Statistics, invited revision, 2023)
- Models for studying relationships among attributes under network interference in large populations: Fritz, Schweinberger, Bhadra, and Hunter (forthcoming).

Structure

Advantage: theoretical guarantees

Challenges:

- Network data are **non-standard data.**
- Network data are **dependent data.**
- **It may not be possible to observe independent replications from the same source,** so that classical LLNs and CLTs cannot be invoked to obtain theoretical guarantees.

Existing statistical theory *does not guarantee that estimators of parameters based on dependent network data are close to the truth.*

Advantage: theoretical guarantees

Local dependence facilitates some of the first theoretical guarantees for models with complex dependence:

- Recovery of non-overlapping subpopulations: Schweinberger (Bernoulli, 2020).
- Recovery of parameters given non-overlapping subpopulations: Schweinberger and Stewart (Annals of Statistics, 2020).
- Recovery of parameters (in general): Stewart and Schweinberger (Annals of Statistics, invited revision, 2023)
- Disclaimers: quantifying uncertainty about parameter estimators: Stewart (Bernoulli, invited revision, 2024).

The general mathematical results in Stewart and Schweinberger (2023) underscore the importance of additional structure for the purpose of controlling dependence ($\rightarrow$ local dependence).

Structure

Advantage: scalable methods

Local dependence facilitates local computing on subnetworks along with large-scale computing:

- Simulation.
- Estimation.

Structure

Motivation

Scalable models

- Advantage: theoretical guarantees
- Advantage: scalable methods

Scalable methods

Software

Software

If the non-overlapping subpopulations are unobserved, the model can be estimated in two steps:

- **Step 1:** Estimate the subpopulation structure by taking advantage of the fact that local dependence models are generalizations of stochastic block models and that stochastic block models can be estimated by scalable methods.
- **Step 2:** Estimate the parameters conditional on the estimated subpopulation structure using local computing on subnetwoks.

Babkin, Stewart, Long, and Schweinberger (Computational Statistics & Data Analysis, 2020)

Structure

Motivation

Scalable models

- Advantage: theoretical guarantees
- Advantage: scalable methods

Scalable methods

Software

Software

- `R` package **hergm** (2008–2024), created and maintained by Schweinberger and published in Schweinberger and Luna (Journal of Statistical Software, 2018).
- `R` package **lighthergm** based on `R` package **hergm** (2019–present), created and maintained by Sansan Inc. with the help of Mele and Schweinberger.
- `R` package **bigergm**, which is the `CRAN` version of `R` package **lighthergm** (2024–present), created and maintained by Fritz with the help of the `hergm` and `lighthergm` core development teams.

All of them implement the scalable methods of Babkin, Stewart, Long, and Schweinberger (2020), but **bigergm** is on `CRAN` and is the most advanced version.

References

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.

Butts, C. T. (2009), "A behavioral micro-foundation for cross-sectional network models," Tech. rep., University of California, Irvine.

Frank, O., and Strauss, D. (1986), "Markov graphs," *Journal of the American Statistical Association*, 81, 832–842.

Mele, A. (2017), "A structural model of dense network formation," *Econometrica*, 85, 825–850.

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

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

References

(cont.)

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.

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

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.

Stewart, J. R. (2024), "Rates of convergence and normal approximations for estimators of local dependence random graph models," *Bernoulli*, invited major revision. arXiv:2404.11464.

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.org/abs/2012.07167*, invited major revision by *The Annals of Statistics*.