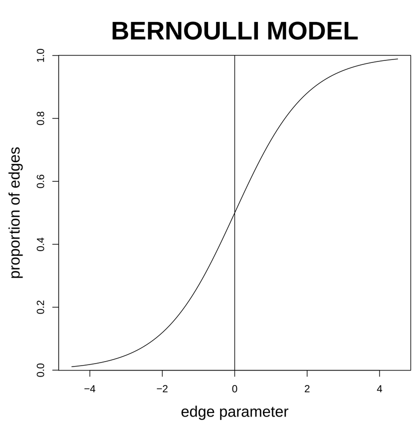
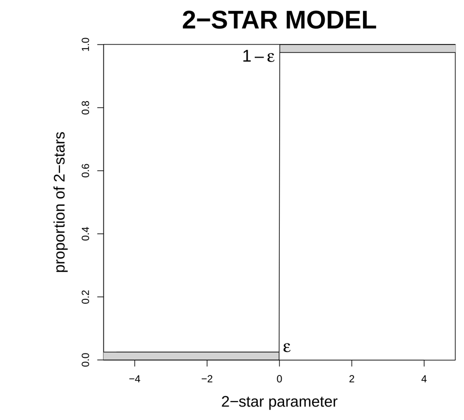


Viable and non-viable models of large networks, simulation and inference

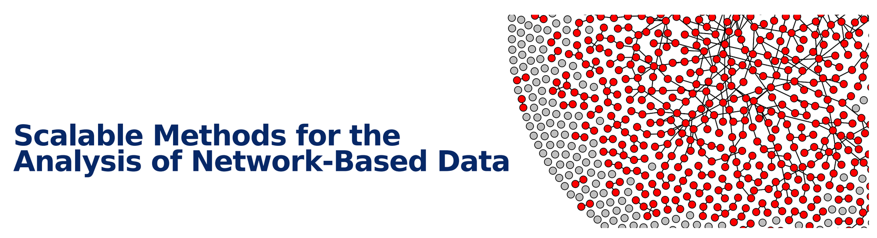
Michael Schweinberger

michael schweinberger at stat psu edu

Supported by ONR grant N00014-8-1-1015



Viable and non-viable models



Simulating large networks and learning the structure of large networks is based on models. Some models of large networks are viable, others are not (1, 2, 3, 4, 5, 6).

Contributions

- Introduce notion of instability of models.
- Discuss characteristics of unstable models.
- Show impact of instability on simulation.
- Show impact of instability on learning.
- Detect unstable models.

Instability

Model: discrete exponential family $\{P_\theta, \theta \in \Theta\}$ with probability mass function of the form

$$p_\theta(y_N) = \frac{\exp[q_\theta(y_N)]}{\sum_{x_N} \exp[q_\theta(x_N)]},$$

where

- y_N: network with $N \in O(n^2)$ possible edges between n nodes.
- $p_\theta(y_N)$: probability mass of network y_N.
- $q_\theta(y_N) = \langle \eta(\theta), g(y_N) \rangle$: inner product of vector of natural parameters $\eta(\theta)$ and vector of statistics $g(y_N)$.
- $I_N(\theta) = \min_{y_N}[q_\theta(y_N)] = 0$ (without loss).
- $S_N(\theta) = \max_{y_N}[q_\theta(y_N)]$.

Definition. A discrete exponential family distribution $P_\theta, \theta \in \Theta$, is stable if there exist constants $C > 0$ and $N_C > 0$ such that

$$S_N(\theta) \le CN \; \forall \; N > N_C,$$

and unstable if, for any $C > 0$, however large, there exists $N_C > 0$ such that

$$S_N(\theta) > CN \; \forall \; N > N_C.$$

A discrete exponential family $\{P_\theta, \theta \in \Theta\}$ is stable if all $\theta \in \Theta$ mapping to $\eta(\theta) \neq 0$ give rise to stable distributions P_θ, and unstable if all $\theta \in \Theta$ mapping to $\eta(\theta) \neq 0$ give rise to unstable distributions P_θ.

Example: model with number of 2-stars implies $S_N(\theta) = |\eta(\theta)| \, N \, (n-2) \in O(n^3)$ and is therefore unstable.

Characteristic I: sensitivity

Let

$$\Lambda(x_N, y_N; \theta) = \log \frac{p_\theta(y_N)}{p_\theta(x_N)}, \; x_N \sim y_N$$

be the log odds of $p_\theta(y_N)$ relative to $p_\theta(x_N)$, where $x_N \sim y_N$ means that networks x_N and y_N match in all but one edge.

Theorem 1. If a discrete exponential family distribution P_θ, $\theta \in \Theta$, is unstable, then there exist no constants $C > 0$ and $N_C > 0$ such that

$$|\Lambda(x_N, y_N; \theta)| \le C \; \forall \; x_N \sim y_N \; \forall \; N > N_C. \square$$

$\Longrightarrow$ smallest possible changes, changes of one edge, may result in extremely large log odds.

Example: model with number of 2-stars implies $|\Lambda(x_N, y_N; \theta)| \le 2\,|\eta(\theta)|\,(n-2) \in O(n)$.

Characteristic II: degeneracy

Let $\mathcal{M}_{\epsilon,N}$ be the set of networks y_N with $q_\theta(y_N) > (1-\epsilon) S_N(\theta)$.

Theorem 2. If a discrete exponential family distribution P_θ, $\theta \in \Theta$, is unstable, then it is degenerate in the sense that, for any $0 < \epsilon < 1$, however small,

$$P_\theta(Y_N \in \mathcal{M}_{\epsilon,N}) \longrightarrow 1 \text{ as } N \longrightarrow \infty. \square$$

$\Longrightarrow$ almost all probability tends to be concentrated on the modes of the probability mass function $p_\theta(y_N)$ provided N is large.

$\Longrightarrow$ effective support of probability mass function $p_\theta(y_N)$, the subset of networks y_N with non-negligible probability mass, is reduced.

$\Longrightarrow$ in general, model cannot represent observed networks, because modes of probability mass function $p_\theta(y_N)$ do not resemble observed networks.

Impact of instability on simulation

Gibbs samplers: sample edges y_{ij} between nodes i and j from full conditional distributions of the form

$$Y_{ij} \mid y_{-ij} \sim \text{Bernoulli}(\pi_{ij}(y_{-ij}; \theta)),$$

where y_{-ij} denotes the collection of edges y_N excluding y_{ij}, and the log odds of $\pi_{ij}(y_{-ij}; \theta)$ is given by

$$\log \frac{\pi_{ij}(y_{-ij}; \theta)}{1 - \pi_{ij}(y_{-ij}; \theta)} = \Lambda(\{y_{-ij}, y_{ij} = 0\}, \{y_{-ij}, y_{ij} = 1\}; \theta).$$

Metropolis-Hastings algorithms: move from x_N to y_N, generated from probability mass function f with support $\{y_N : y_N \sim x_N\}$, with probability

$$\alpha(x_N, y_N; \theta) = \min\left\{1, \exp[\Lambda(x_N, y_N; \theta)] \frac{f(x_N \mid y_N)}{f(y_N \mid x_N)}\right\}.$$

$\Rightarrow$ due to the excessive sensitivity of the stationary distribution, convergence to, and sampling from, unstable distributions may require an extremely large number of iterations.

$\Rightarrow$ due to the degeneracy of stationary the distribution, multiple starting points may be required, because algorithms tend to trapped at modes of the probability mass function $p_\theta(y_N)$.

$\Rightarrow$ problematic behavior of simulation algorithms tends to be rooted in the stationary distribution: some simulation algorithms may outperform others, but all tend to suffer from the excessive sensitivity and degeneracy of the stationary distribution.

Impact of instability on learning

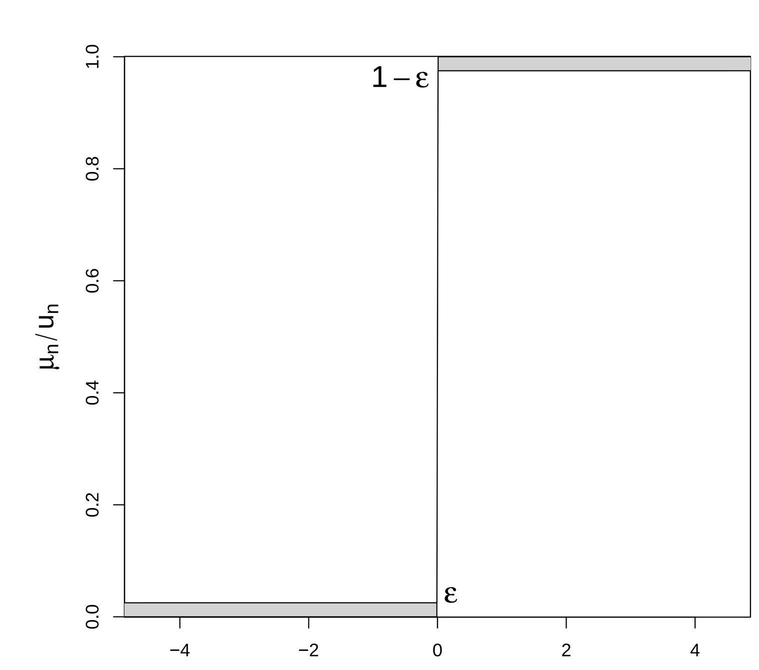
Simple example: unstable, one-parameter exponential family with natural parameter $\eta(\theta) = \theta$ and mean-value parameter $\mu_N(\theta) = E_\theta[g(Y_N)]$, e.g., model with number of 2-stars. Let $L_N = \min_{y_N}[g(y_N)] = 0$ (without loss) and $U_N = \max_{y_N}[g(y_N)]$.

Corollary. If a one-parameter exponential family $\{P_\theta, \theta \in \Theta\}$ is unstable, then, for any $\theta < 0$, however small,

$$\frac{\mu_N(\theta)}{U_N} \longrightarrow 0 \text{ as } N \longrightarrow \infty$$

and, for any $\theta > 0$, however small,

$$\frac{\mu_N(\theta)}{U_N} \longrightarrow 1 \text{ as } N \longrightarrow \infty. \square$$



$\Longrightarrow$ mean-value parameter $\mu_N(\theta)$ is close to infininum (all $\theta < 0$) or supremum (all $\theta > 0$).

$\Longrightarrow$ mean-value parameter $\mu_N(\theta)$ is extremely sensitive to changes of θ around 0.

Maximum likelihood estimate of θ is the root of the estimating function

$$\nabla_\theta \log p_\theta(y_N) = g(y_N) - E_\theta[g(Y_N)] = g(y_N) - \mu_N(\theta),$$

where $\nabla_\theta \log p_\theta(y_N)$: gradient of $p_\theta(y_N)$ with respect to θ:

$\Longrightarrow$ finding the root of the estimating function $\nabla_\theta \log p_\theta(y_N)$ amounts to finding the value of θ such that the expected value of the statistic $\mu_N(\theta) = E_\theta[g(Y_N)]$ matches the observed value of the statistic $g(y_N)$.

$\Longrightarrow$ unless $g(y_N)$ is close to the boundary of the $\mu_N(\theta)$-space, the root of the estimating function $\nabla_\theta \log p_\theta(y_N)$ tends to be close to 0 in the θ-space.

$\Longrightarrow$ since $\mu_N(\theta)$ is extremely sensitive to changes of θ around 0, estimating function $\nabla_\theta \log p_\theta(y_N)$ is extremely sensitive to changes of θ around 0.

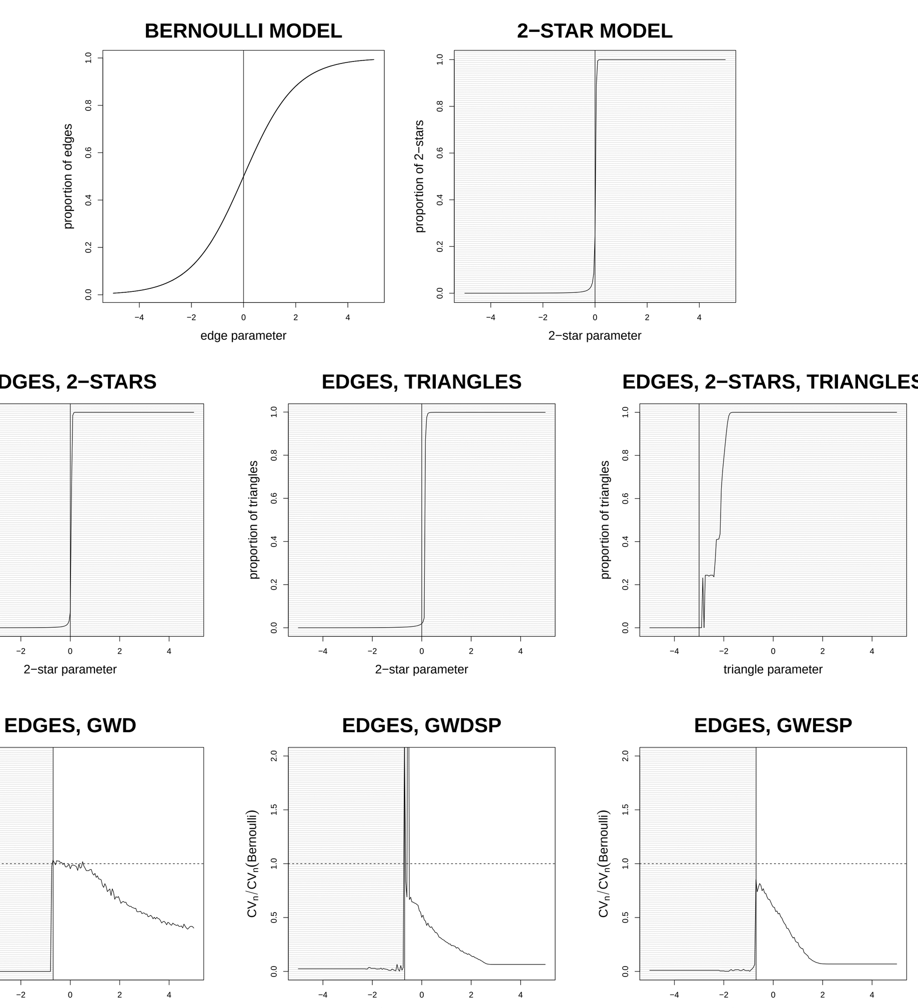
$\Longrightarrow$ root-finding algorithms (e.g., Newton-Raphson, Fisher scoring, Robbins-Monro, Geyer-Thompson) tend to make small steps in the θ-space around 0 and large steps in the $\mu_N(\theta)$-space and struggle to converge.

$\Longrightarrow$ problematic behavior of learning algorithms tends to be rooted in the model: some learning algorithms may outperform others, but all tend to suffer from the instability of the model.

Detection of unstable models

Detection of unstable models: exponential families with Markov dependence (e.g., number of 2-stars, triangles) and curved exponential families (e.g., GWD, GWDSP, GWESP).

Simulation: undirected graphs with $n = 32$ nodes and $N = 496$ possible edges. Shaded regions indicate unstable regions.



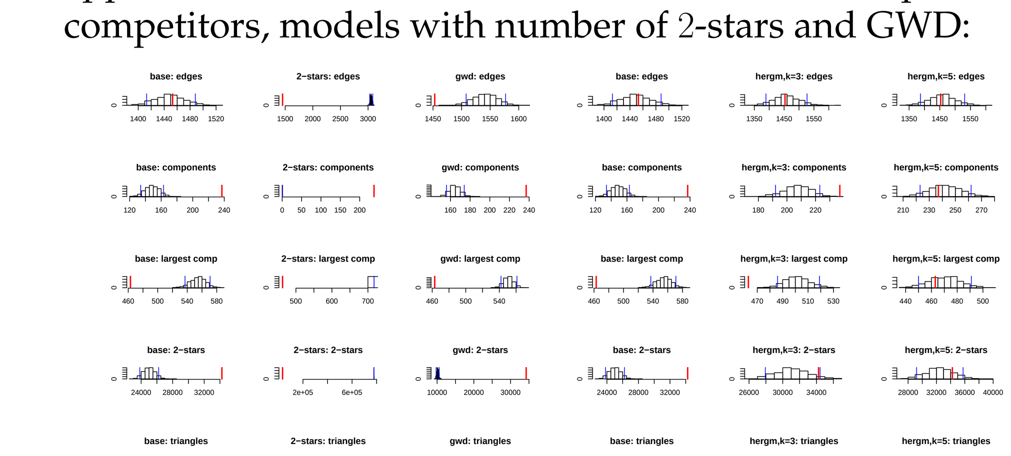
Discussion

- **Unstable models:** problematic due to excessive sensitivity and degeneracy and its impact on simulation and learning: penalties recommended.
- **Super-stable model:** model with bounded log odds $\Lambda(x_N, y_N; \theta)$.
- **Example 1:** Bernoulli model.
- **Example 2:** Ising model, exploiting spatial structure to bound log odds.
- **Example 3:** hierarchical degree model, exploiting latent structure to bound log odds (work with Duy Quang Vu and Miruna Petrescu-Prahova, funded by ONR grant N00014-8-1-1015):

$$p_\theta(y_N) = \frac{\exp\left[\sum_{i=1}^n \eta_i(\theta) g_i(y_N)\right]}{\sum_{x_N} \exp\left[\sum_{i=1}^n \eta_i(\theta) g_i(x_N)\right]}$$

motivated by maximum entropy principle, where

- $g_1(y_N), \ldots, g_n(y_N)$: degrees of nodes $1, \ldots, n$.
- $\eta_1(\theta), \ldots, \eta_n(\theta)$: degree parameters of nodes $1, \ldots, n$, which are functions of degree parameters $\theta_1, \ldots, \theta_k$ of latent classes $1, \ldots, k$.
- applied to 9/11 communication network and compared to competitors, models with number of 2-stars and GWD:



References

[1] D. Strauss. On a general class of models for interaction. *SIAM Review*, 28: 513–527, 1986.

[2] J. Jonasson. The random triangle model. *Journal of Applied Probability*, 36: 852–876, 1999.

[3] T. A. B. Snijders. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3:1–40, 2002.

[4] M. S. Handcock. Assessing degeneracy in statistical models of social networks, 2003. Center for Statistics and the Social Sciences, University of Washington. Available from: http://www.csss.washington.edu/Papers.

[5] A. Rinaldo, S. E. Fienberg, and Y. Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, 3:446–484, 2009.

[6] J. H. Koskinen, G. L. Robins, and P. E. Pattison. Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology*, 7(3):366–384, 2010.