Statistical Challenges in Network Modeling

Stephen E. Fienberg

Department of Statistics

Machine Learning Department

Cylab

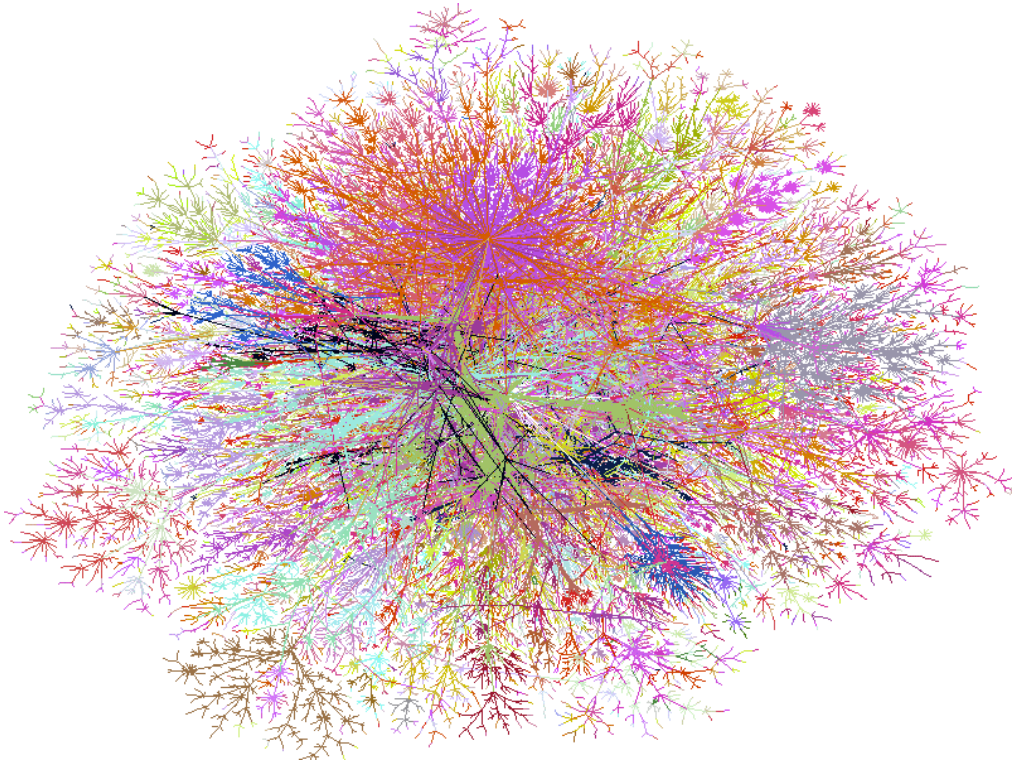
Carnegie Mellon University

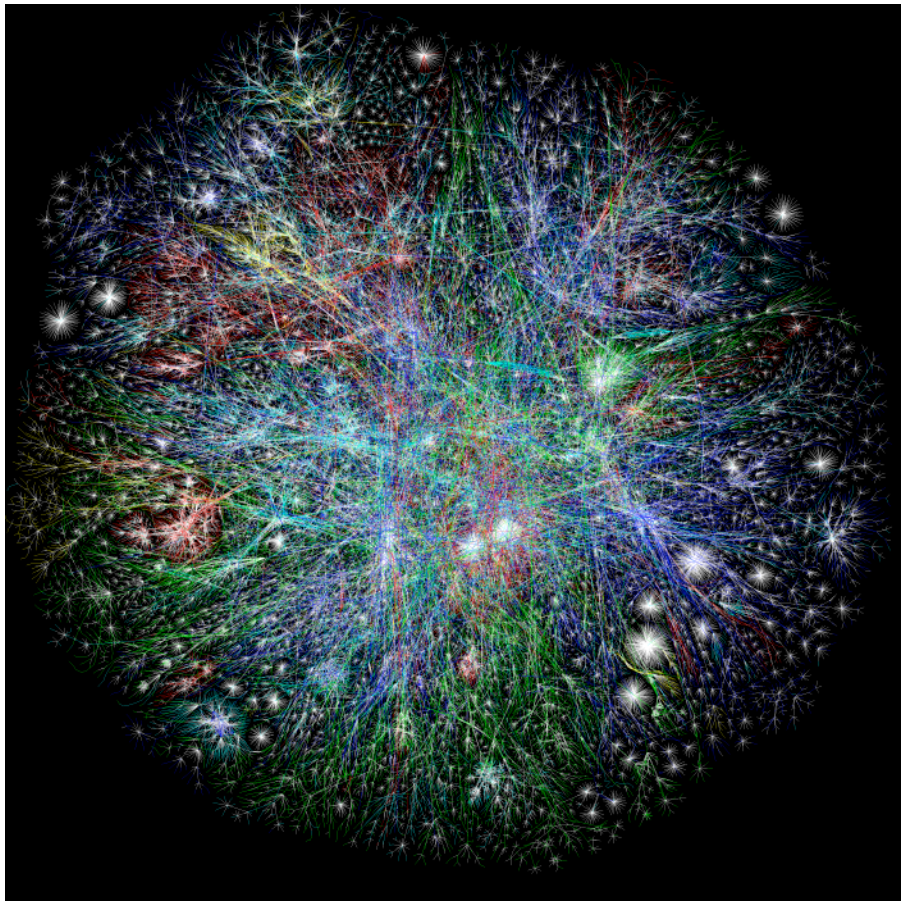
Pittsburgh, PA USA

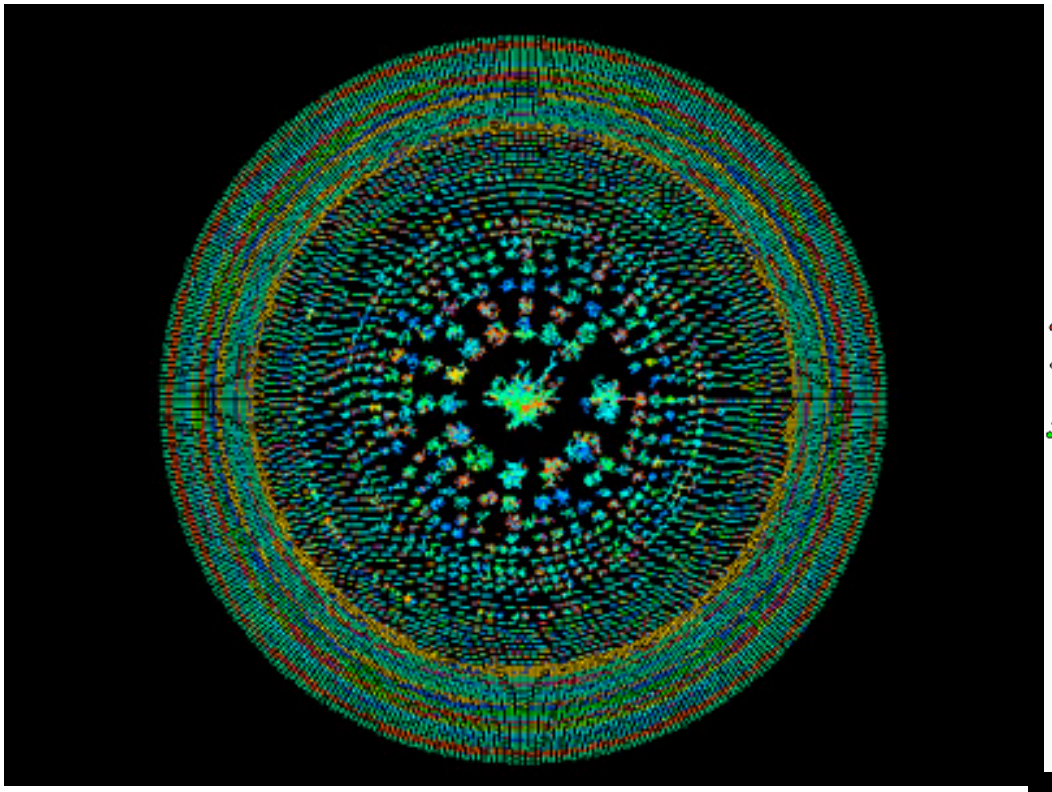
A Statistician's Apology

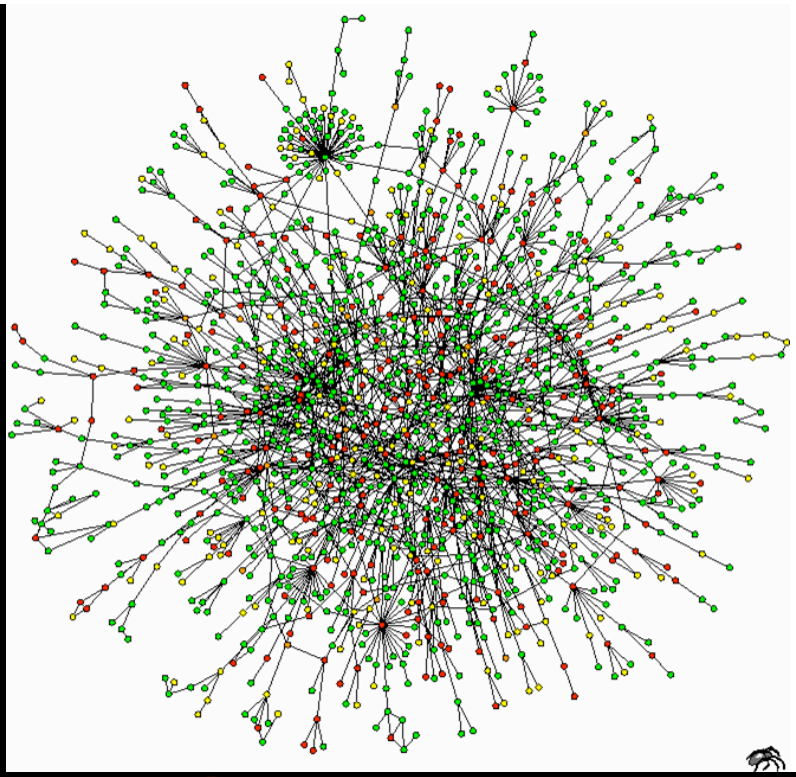
- **Lots of researchers do "network analysis."**
  - **Some approaches are generative; some merely descriptive.**
  - **Some insights based for actual networks are truly innovative.**
  - **Many ideas and methods are regularly reinvented.**
- **Today I'll give a broad outline for a systematic approach to dynamic network problems rooted in my statistical perspective:**
  - **Many pieces exist in work of others.**
  - **Many of challenges are relevant to other systematic approaches.**

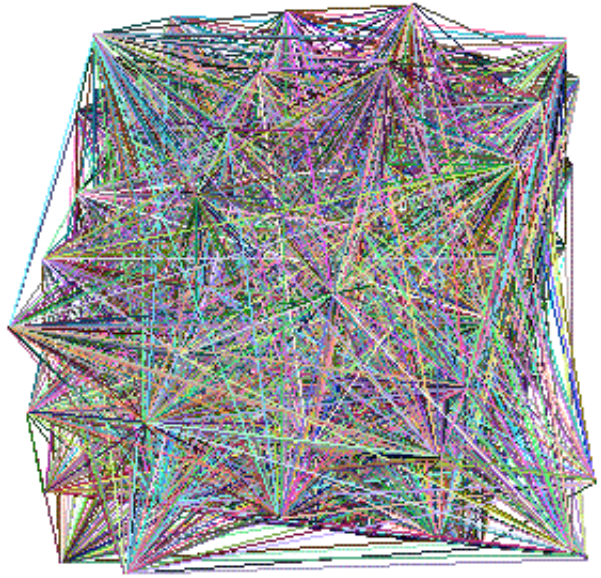
Making Pretty Pictures — Visualizing Networks — Is Easy

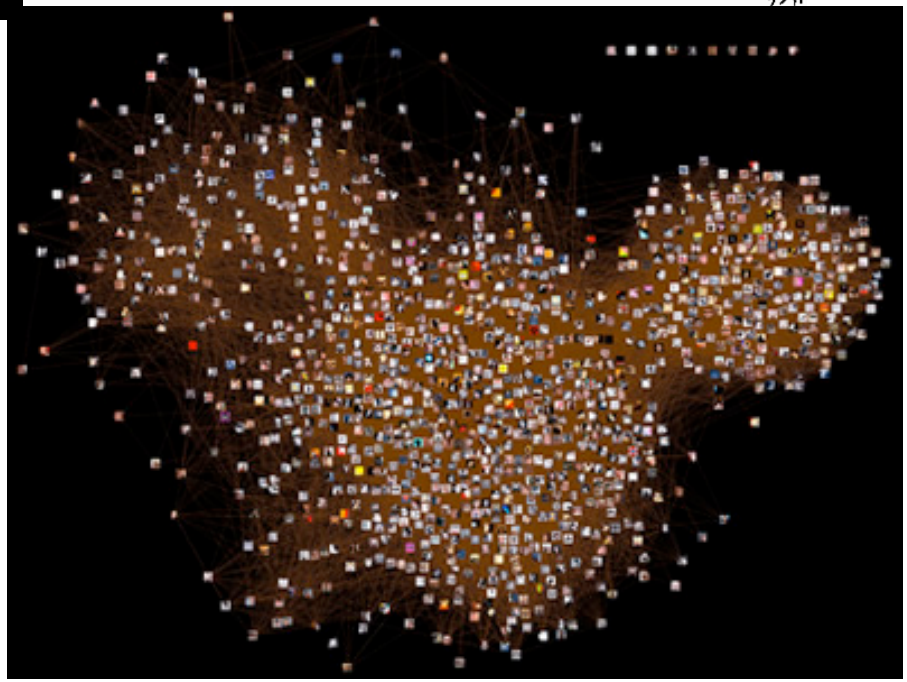




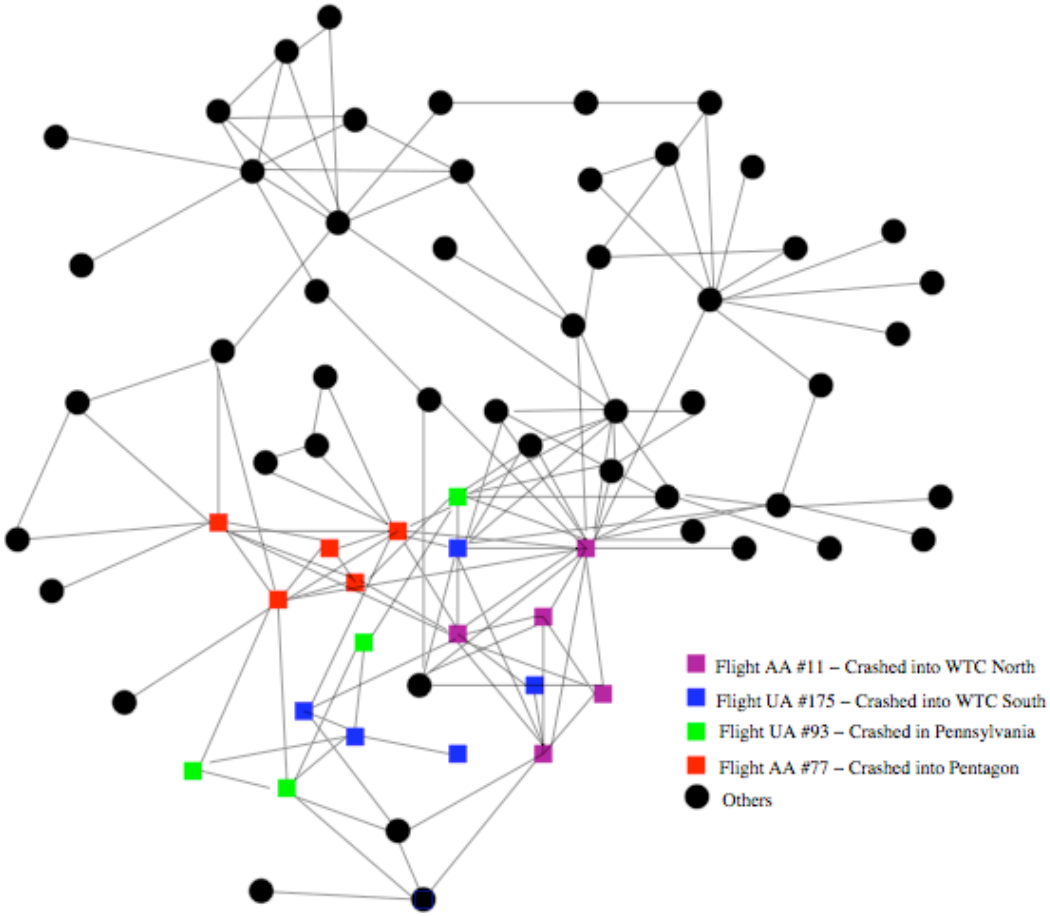








9/11 Terrorists



Lots of Probabilistic/Statistical Models

- **Types of models:**
  - **Descriptive vs. Generative.**
  - **Static vs. Dynamic.**
- **Origin of social network models in 1930s, integrated with graph representation in 1950s.**
- **Erdos-Renyi random graph models.**
  - **Generalized random graph models.**
  - **Stochastic process reinterpretations.**
- **Sociometric models such as p_1 and ERGMs.**
- **Machine learning / latent-variable models:**
  - **Stochasitic block models for mixed membership.**

Applications Galore

- Small world studies
- Social networks:
  - Sampson's monks
  - Classroom friendship
  - My Space, Facebook
- Organization theory
  - Branch banks
- Homeland security
- Politics
  - Voting behavior
  - Bill co-sponsorship
- Public health
  - Needle sharing
  - Spread of AIDS
  - Obesity
- Computer science:
  - Email networks (Enron)
  - Internet
  - WWW routing systems
- Biology:
  - Protein-protein interactions
  - Zebras

Oodles of Data

- **Networks with multiple relationships, and multiple attributes/covariates at each node.**
- **Dynamics of edge (link) and node creation and disappearance.**
- **P38 Dynamic Network Evolution**

But Doing Careful Statistical Analysis is Difficult

- **Claims for network behavior are often based on casual empiricism:**
  - **Power laws are everywhere, yet nowhere once we look closely at the data.**
- **Inferential issues usually buried:**
  - **Algorithms, simulations, and "experiments" are not substitutes for formal statistical representation and theory.**

Framework for Networks Evolving over Time

- **Our representation for a network will be a graph: $G_t=\{N_t;E_t\}$.**
  - **Nodes and edges can be created and can die.**
  - **Edges can be directed or undirected.**
  - **Data are available to be observed beginning at time t_0.**
- **There exists stochastic process, evolving over time which, combined with initial conditions, describes the network structure and evolution.**
  - **May involve more than dyadic relationships.**

Markovity, Heterogeneity & Non-stationarity

- **Continuous time Markov process.**
- **Model intensity matrix:** $q(x, \bar{x})$

  $P(X(t+\varepsilon) = \bar{x} \mid X(t) = x) \approx \varepsilon q(x, \bar{x})$ **if** $x \neq \bar{x}$.

  – **"Rate" parameters may vary across nodes and time.**
- **Model may depend on "characteristics" of the nodes (attributes) as well as "characteristics" of their connections.**
  – **Fixed or time-varying.**
  – **Characteristics may be concomitants or they may be outcomes.**
- **Stationarity vs. non-stationarity?**

Forms of Network Data

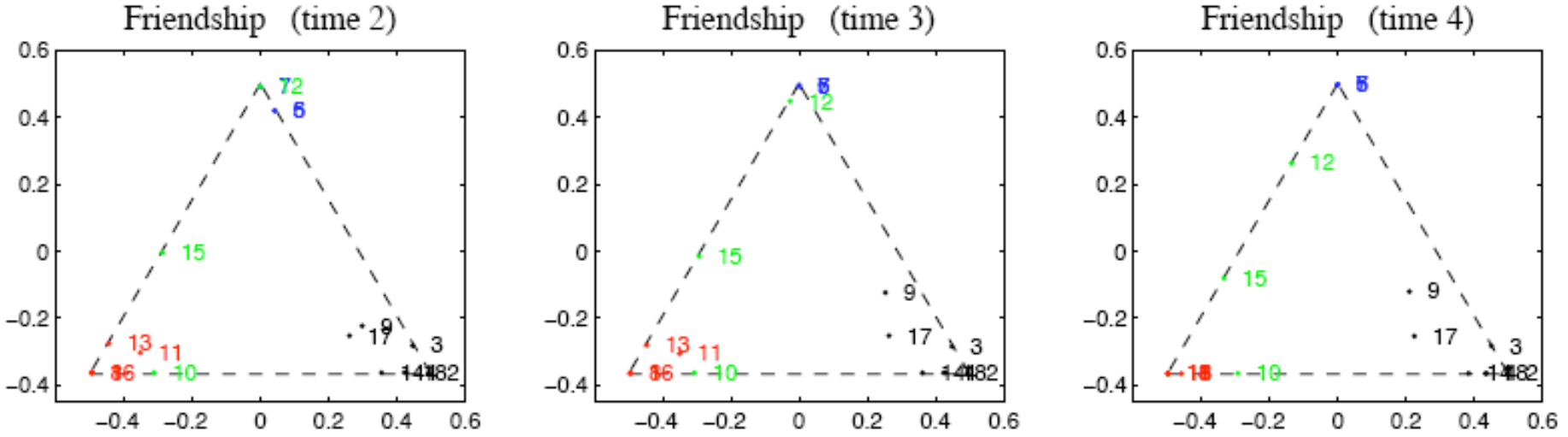
1. **Observe formation (or removal) of each edge with a time stamp indicating when this occurs.**
   - **Can see how entire network or subnetwork changes with each transaction.**
2. **Observe status of network or sub-network at *T* epochs.**
   - **Represent snapshots of network and correspond to information on incidence of links and information on relationships.**
3. **Observe the cumulative effect of the stochastic process at one or more time points.**
   - **"Prevalence" approach.**

Example 1: Enron E-mail Database

- **Attributes nodes (including organization chart!) and full text on all e-mail messages.**
- **Multiple addressees and cc's. Thus observations produce structure different from dyadic edges.**
- **Messages contain time stamps, so we are in situation 3.**
- **Question: Who was party to fraudulent transactions and when?**

Example 2: Monks in a Monastery (Airoldi, et al.)

- **18 noviates observed over two years.**
  - **Network data gather a 4 time points; friendship relationship among noviates measured at 3 successive times.**



Dynamic Mixed Membership Stochastic Blockmodel

- **Data:**

  $X_t(n,m) \quad n,m = 1,2,...,N = 18; t = 1,2,3.$

- **Combine MMSB for observed relations with a simple state-space model for evolution of latent aspects:**

$$P(\vec{\pi}_0(n) \mid \theta) \sim f \circ Gaussian(\vec{0}, A)$$

$$P(\vec{\pi}_t(n) \mid \vec{\pi}_{t-1}(n), \theta) \sim f \circ [Gaussian(\vec{0}, A) + f^{-1} \circ \vec{\pi}_{t-1}(n)]$$

$$P(X_t(n,m) \mid \Pi_t, \theta) \sim Bernoulli(\theta) \sim (\vec{\pi}_t(n)' B \vec{\pi}_t(m))$$

Example 3: The Framingham "Obesity" Study

- **Original Framingham "sample" cohort with offspring cohort of N_0=5124 individuals measured beginning in 1971 for T=7 epochs centered at 1971,1981, 1985, 1989, 1992, 1997, 1999.**
- **Link information on family members and one "close friend." Total number of individuals on whom we have obesity measures is N=12,067.**
- **NEJM, July 2007.**

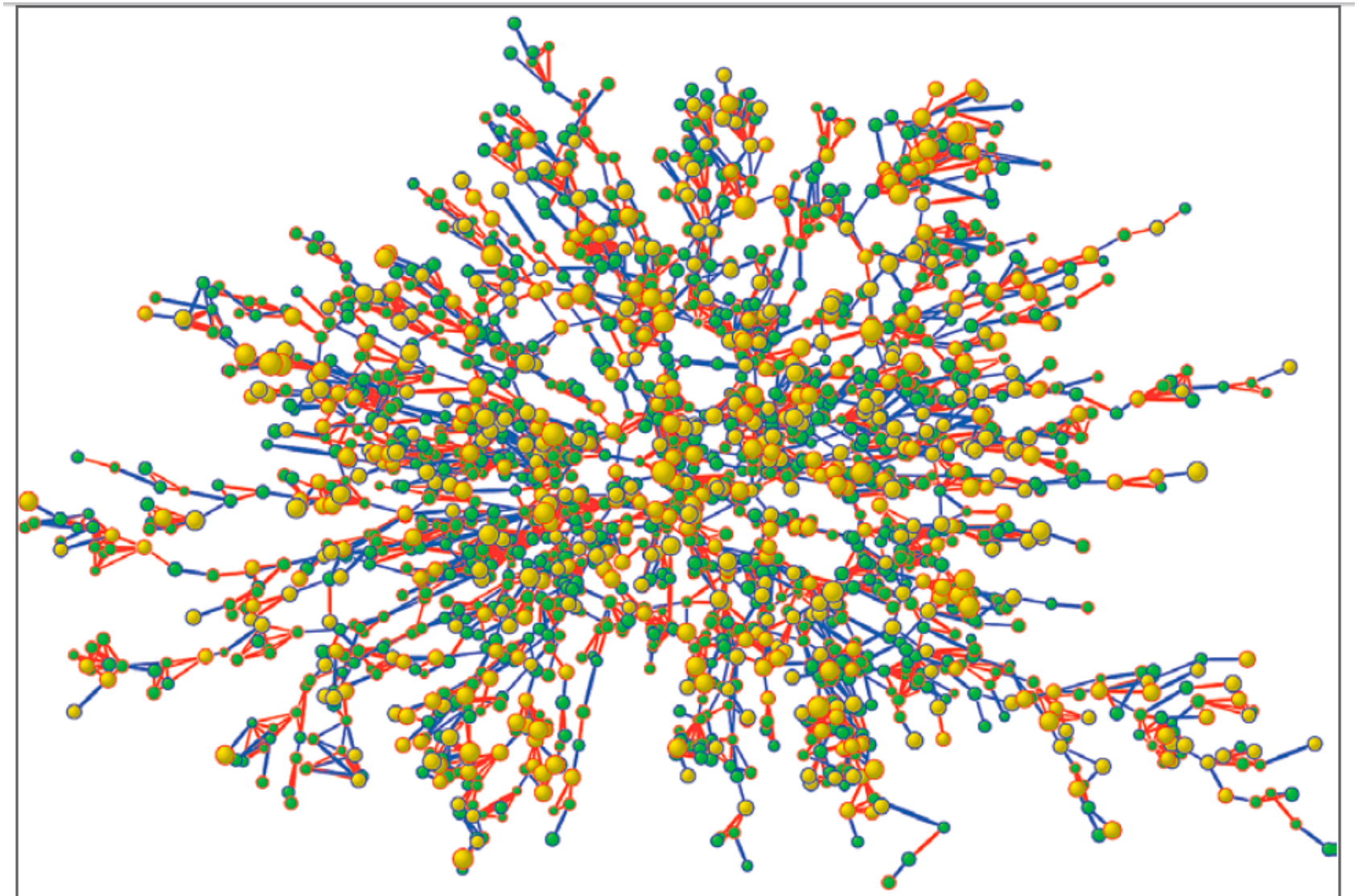
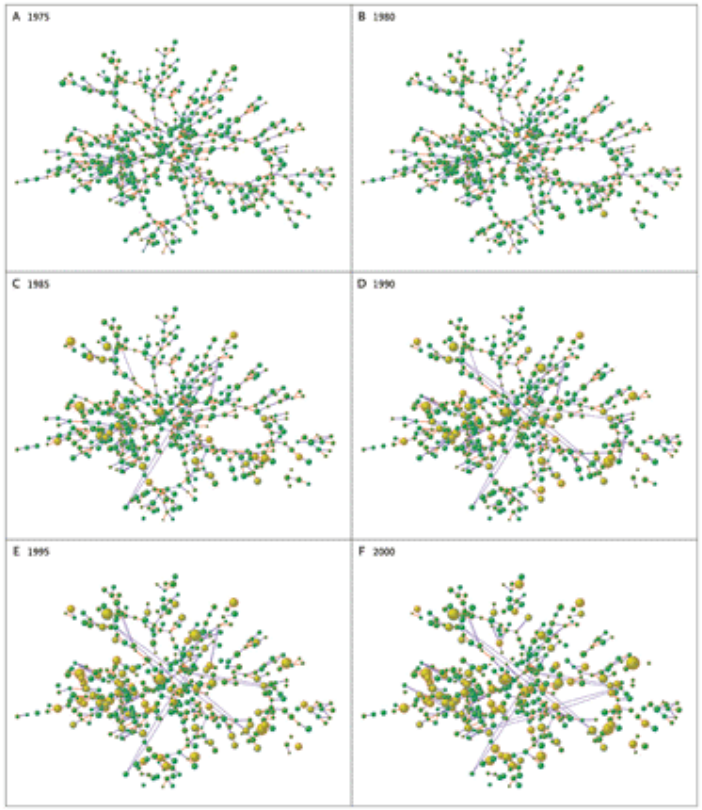


Figure 1. Largest Connected Subcomponent of the Social Network in the Framingham Heart Study in the Year 2000.

Each circle (node) represents one person in the data set. There are 2200 persons in this subcomponent of the social network. Circles with red borders denote women, and circles with blue borders denote men. The size of each circle is proportional to the person's body-mass index. The interior color of the circles indicates the person's obesity status: yellow denotes an obese person (body-mass index, ≥30) and green denotes a nonobese person. The colors of the ties between the nodes indicate the relationship between them: purple denotes a friendship or marital tie and orange denotes a familial tie.

Animation

- http://content.nejm.org/content/vol357/issue4/images/data/370/DC2/NEJM_Christakis_370v1.swf



Obesity Statistical Issues

- **Sampling?**
- **What is a cluster? How do they arise in context of dynamic models?**
- **Embeddabilty?**

Example 4: Social Network of Zebras

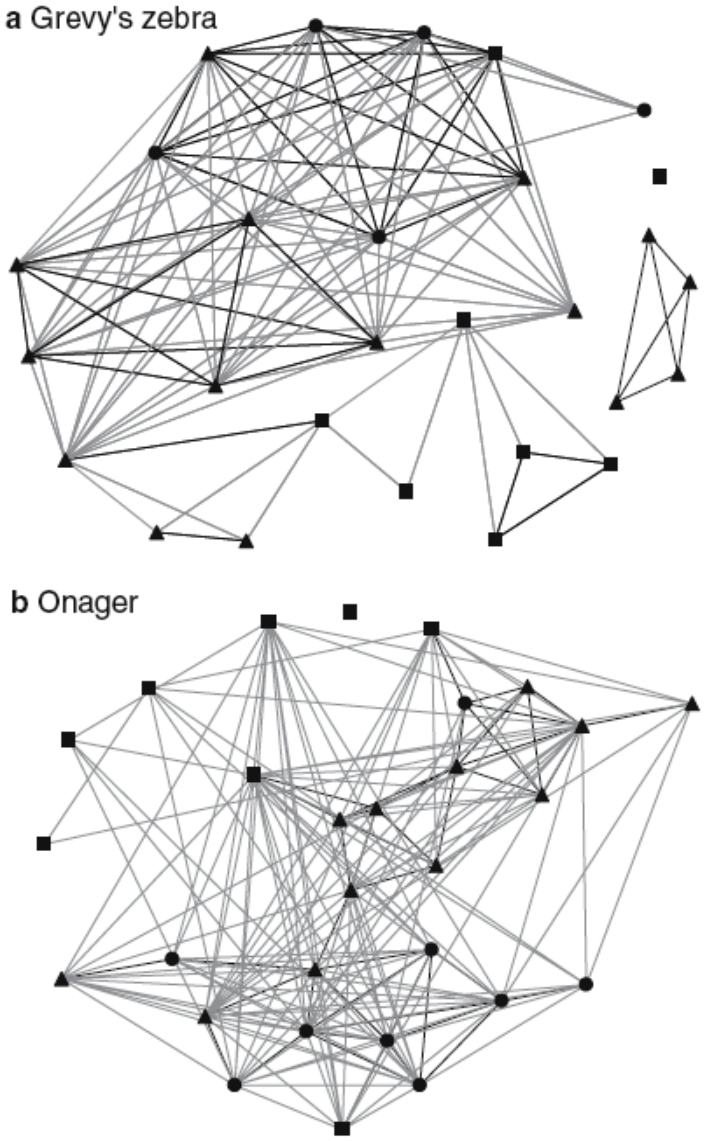
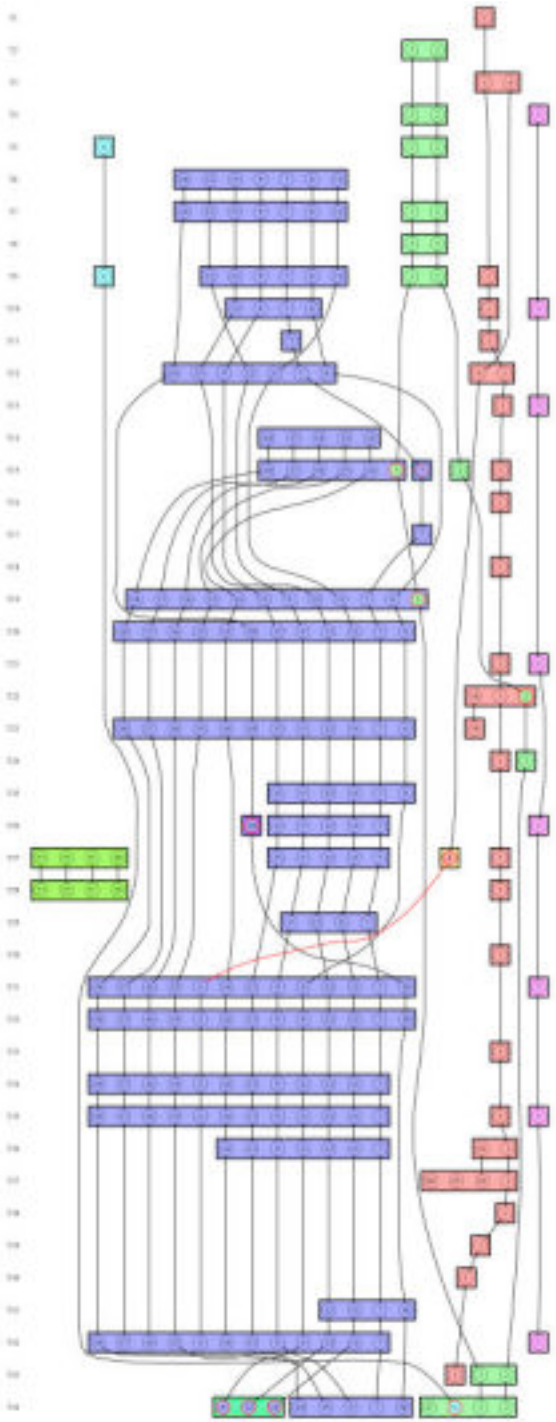


Fig. 3a–b Observed networks for **a** Grevy's zebra (28 individuals) and **b** Onagers (29 individuals). Individuals are vertices, with reproductive status indicated by shape: males (*squares*), lactating females (*circles*), and nonlactating females (*triangles*). *Thin gray lines* join individuals observed together at least once (nonzero network). *Thick black lines* represent statistically significant associations (preferred network)

Dynamical Representation

- What is the stochastic model for group formation and change?
- Groups of females and shifting males who are mating?



Example 5: Links on the Web

Challenging Classes of Problems

- **Data integration.**
- **Computability.**
- **Asymptotics (Assessing goodness of fit).**
- **Sampling.**
- **Embeddability.**
- **Prediction.**
- **Privacy/Confidentiality.**

Data Integration

- **Data arising from multiple sources, with uncertainty associated with node identification.**
  - **Record linkage (De-duping); entity resolution.**
- **How do we link this problem to estimation of dynamical models?**

Computability

- **Algorithms that scale:**
  - **R package on exponential random graph models and latent models for networks: http://csde.washington.edu/statnet/index.shtml**
  - **Siena package from Tom Snidjers.**
- **Approximations:**
  - **Variational methods.**
- **Bayes vs. frequentist.**

Inference and Asymptotics

- ***n*** **nodes,** ***N*** **edges (links) and** ***r*** **relations,** ***p*** **attributes on nodes, etc.**
  - **Analogy with Rasch model?**
- **What is unit of statistical analysis?**
  - **Nodes, dyads, larger groups. etc.**
- **Elaborate models but little knowledge of how they fit the data, especially dynamically. This is why we need asymptotics.**
  - **Some work by Handcock and colleagues**

ERGMs and Identification

- **Some ERGMs are not hierarchical and this means interpretation is problematic; also get strange degeneracies.**
- **Evaluating ERGM likelihoods:**

number nodes	number of edges	number of graphs
7	21	$2^{21} = 2,097,152$
8	28	$2^{28} = 268,435,456$
9	36	$2^{36} = 68,719,476,736$
10	45	$2^{45} = 35,184,372,088,832$

ERGM

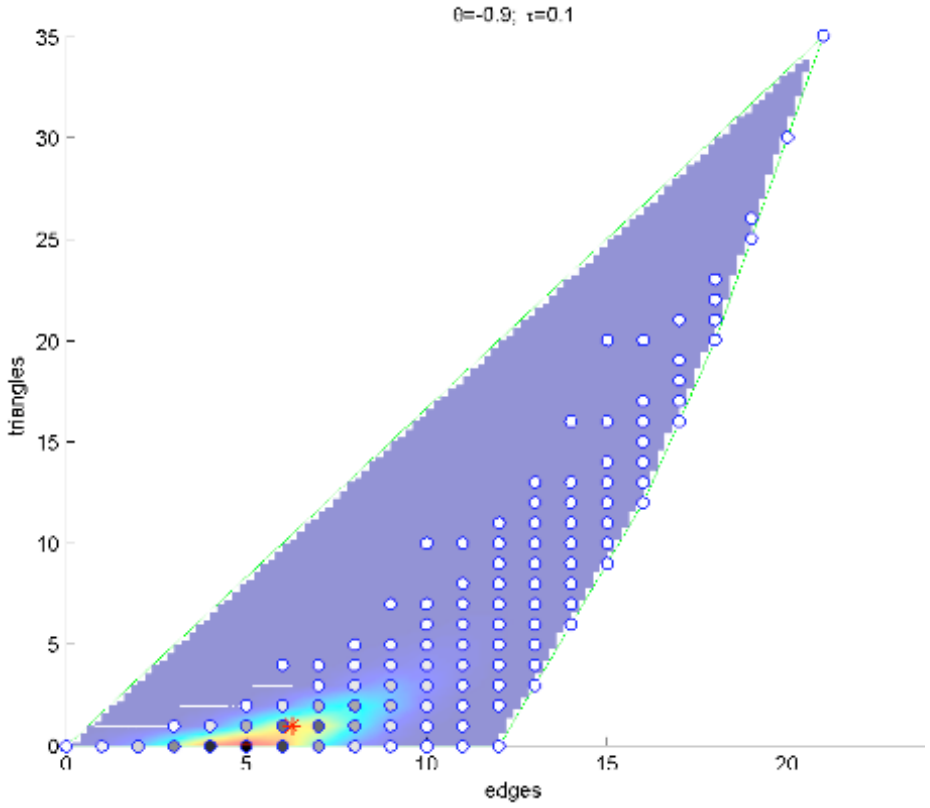
A p^* model is a probabillity model defined on a directed or undirected graph whose density takes an exponential form like

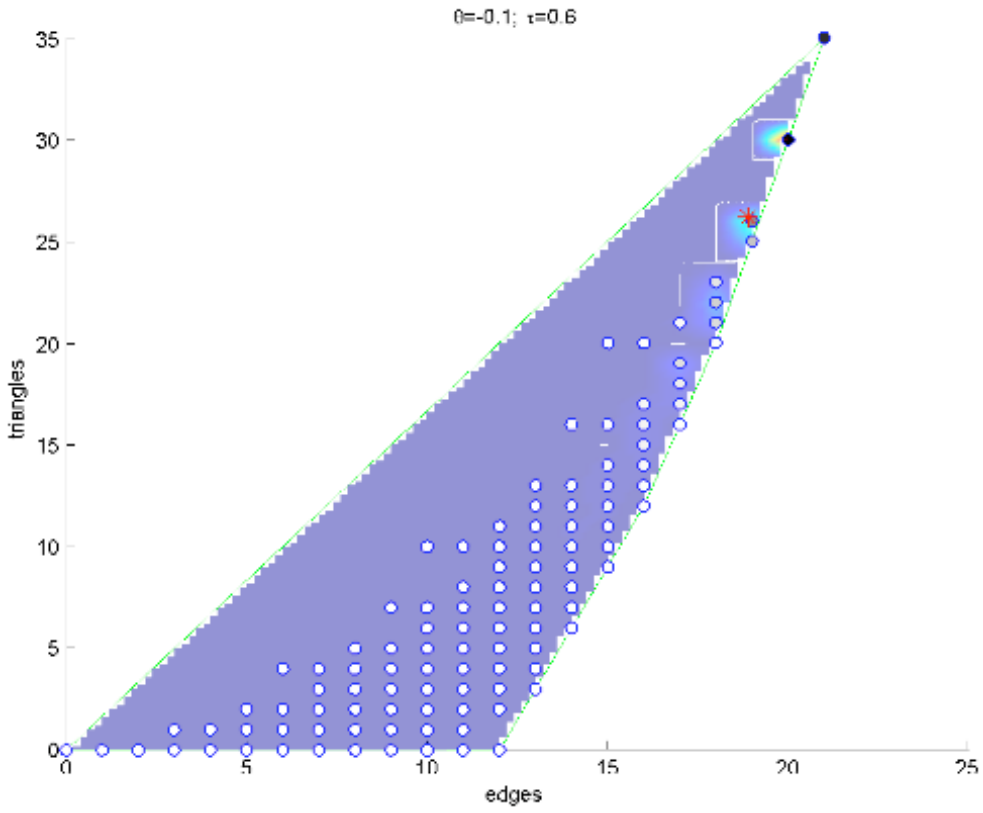
$$p(x) \propto \exp\{\theta E(x) + \sigma_2 S_2(x) + \sigma_3 S_3(x) + \tau T(x)\}$$

The sufficient statistics $E(x)$, $S_2(x)$, $S_3(x)$ and $T(x)$ usually reflect the local configurative patterns in the graph, for example, number of edge, number of 2-star, number of 3-stars and number of triangles.

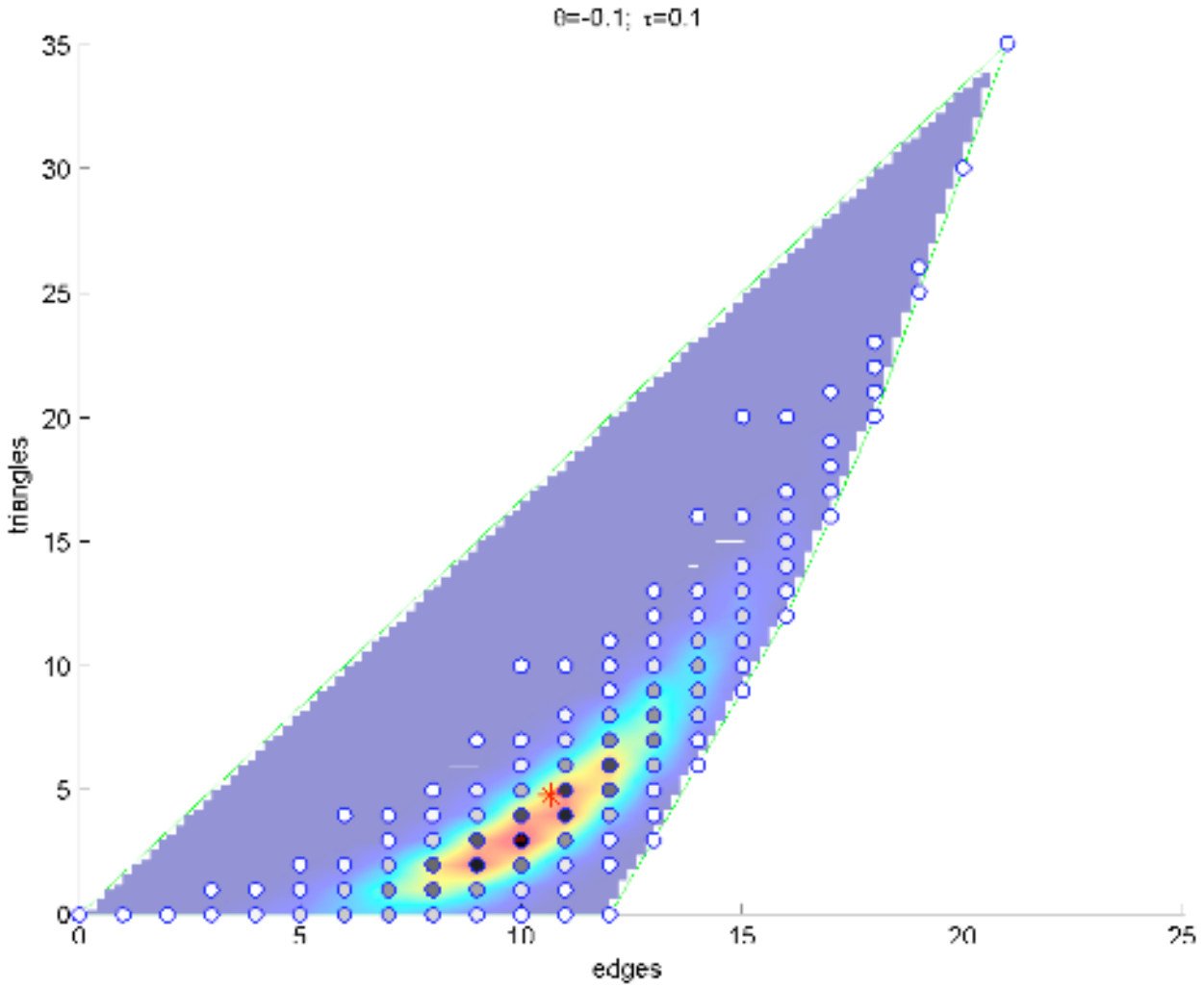
- **When effects associated with stars are zero, we often get near-degenerate behavior.**

Near-Degenerate 7-node

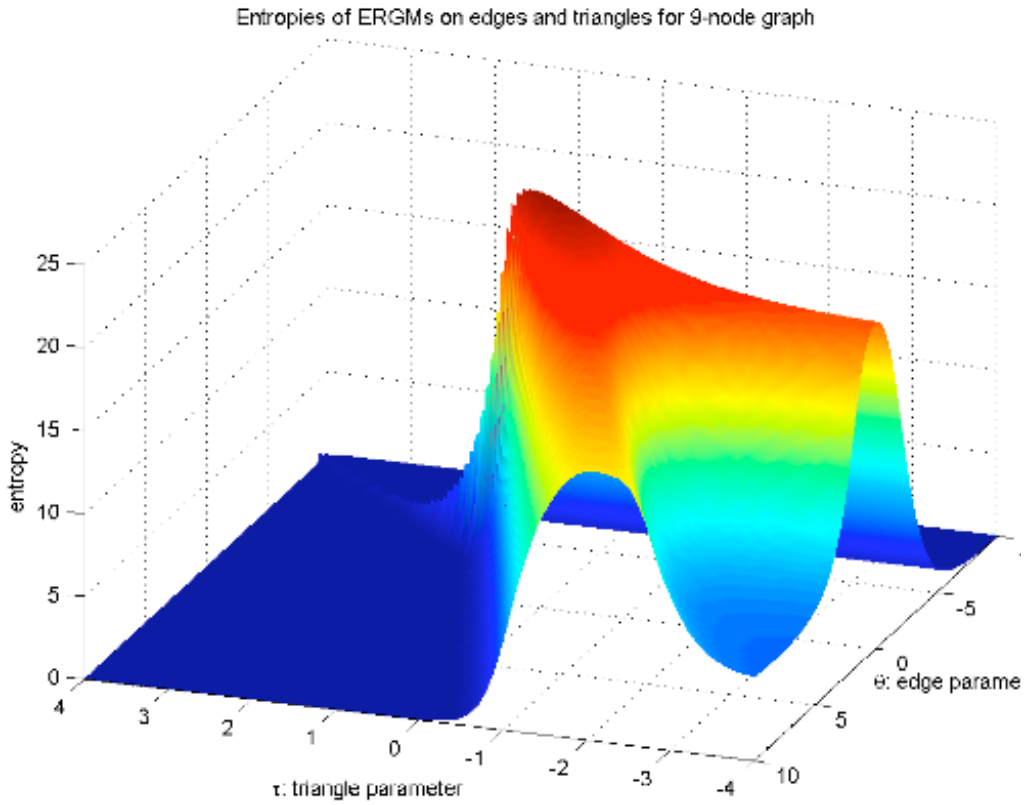


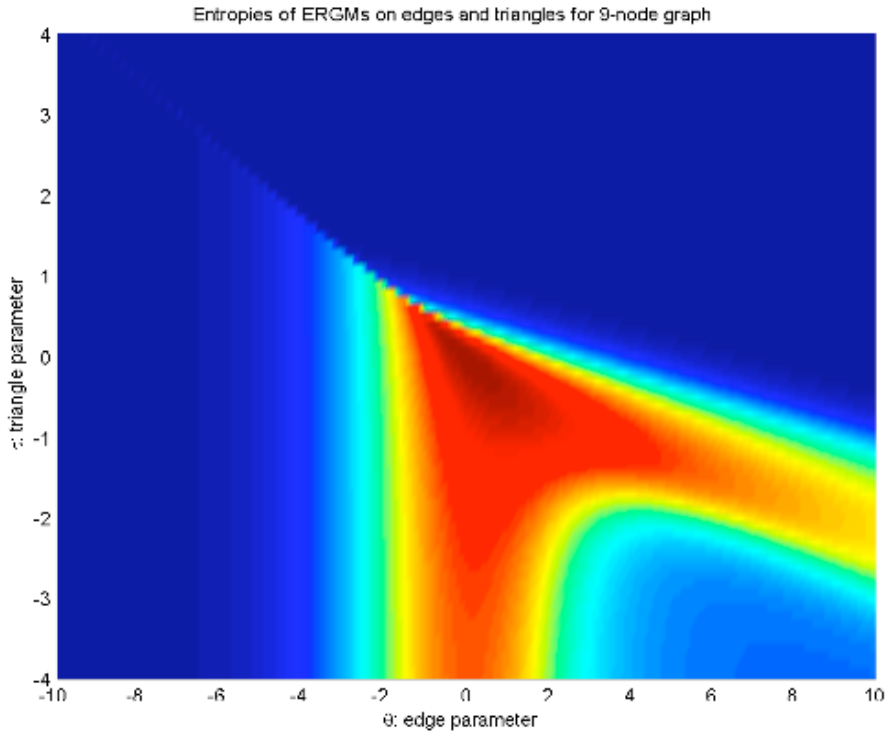


Non-degenerate 7-node

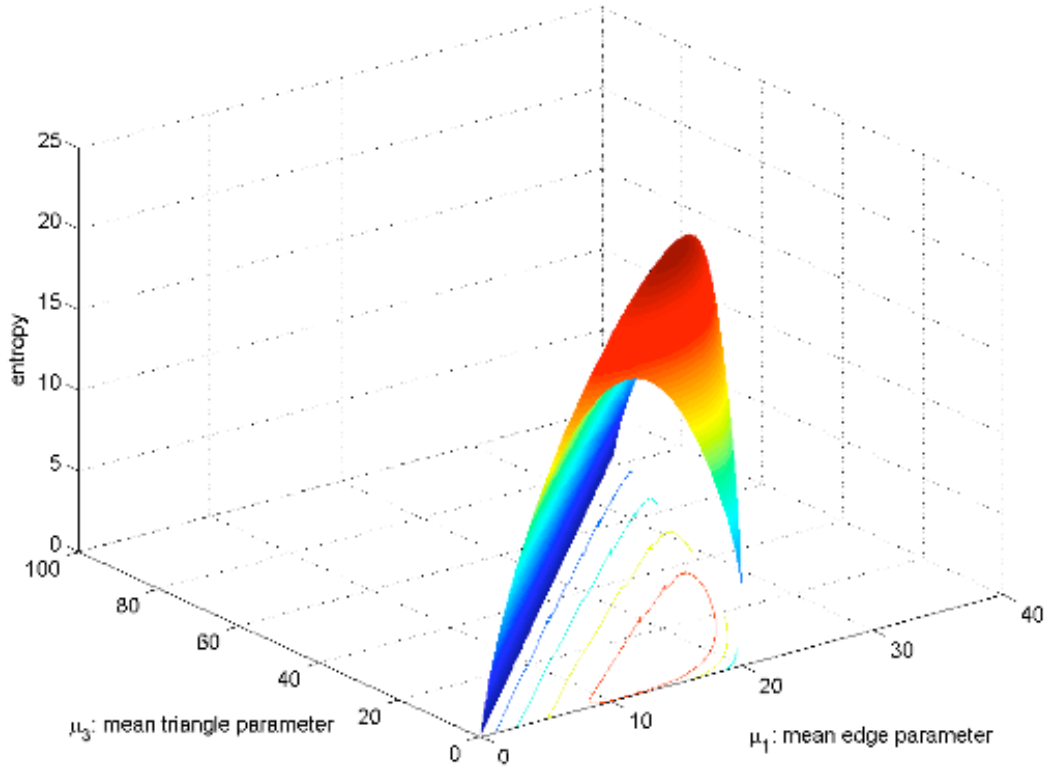


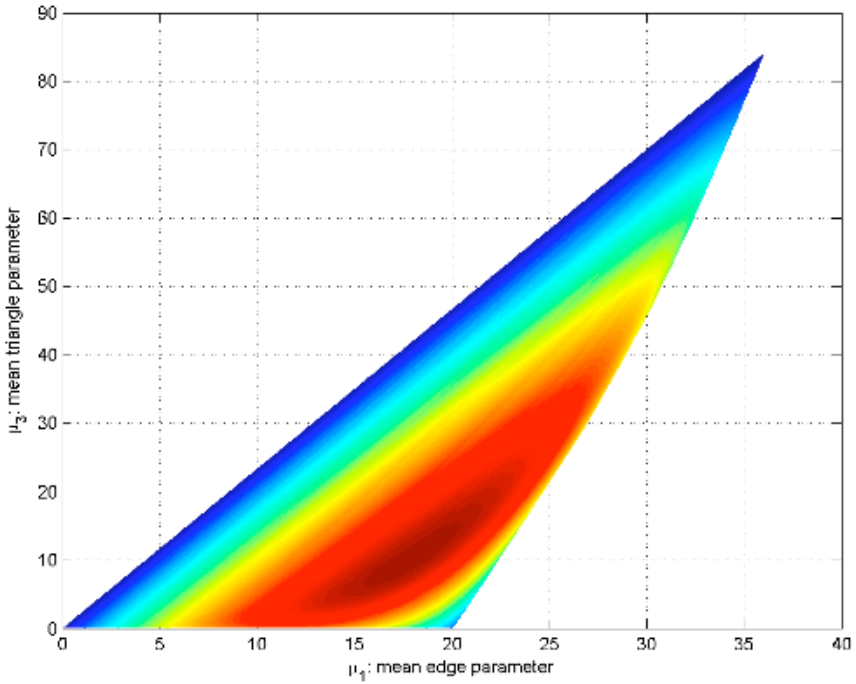
Entropies for 9-nodes





Entropies for 9-Nodes (Means)





Sampling

- **What to sample and how?**
  - **Nodes, edges, relations?**
  - **Adaptive sampling designs, link trace sampling, snowball sampling.**
    - **Framingham design?**
- **Does sampling effect ability to generalize?**
  - **Design-based inference vs. model-based inference.**
  - **Work of Stumpf and colleagues.**

Embeddability

- **Want dynamic network models are explicitly or implicitly expressible in terms of stochastic processes.**
- **Suppose we collect data at *T* epochs as in Framingham study.**
- **Can we estimate parameters of models from observed data?**
- **What might be role of discrete time Markov chain approach for ERGM of Hanneke and Xing?**

Snijders' Approach

- **Stochastic model leads to set of observed network characteristics:**
  - **Simulation models to get at intensity matrices.**
  - **Approximate ERGMs and their transitions.**
  - **Likelihood methods still in infancy.**
- **Looking at degree distributions is not even close to an approximation.**
- **Degeneracy problems similar to those for ERGMs arise, especially for cross-sectional data.**

Prediction

- **For dynamic network settings, and data generated over time there are a series of forecasting problems.**
- **How should we evaluate alternative predictions from different models?**

Privacy Protection & Facebook Fiasco

- **Facebook and other networking sites do little to protect privacy.**
- **Is protect privacy in social network settings an oxymoron?**
  - **"Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography," Backstrom, Dwork, Kleinberg**
- **Do anonymized attributes and links reveal all? If not can we have our cake and eat it too, i.e., protect individuals (nodes) but reveal structure?**

Summary

- **Stochastic model perspective for dynamic networks.**
- **Links to existing approaches:**
  - **ERGM.**
  - **Statistical physics models.**
  - **MMSBM.**
- **Challenging statistical issues.**

Thanks

- **Collaborators:**
  - **A. Rinaldo and Yi Zhou (p* plots)**
  - **Edo Airoldi and David Krackardt (longitudinal Monk model and analyses)**
  - **David Blei, Lise Getoor, Anna Goldenberg, Eric Xing, Alice Zheng**
- **Mark Hancock and Tom Snijders**
- **Stan Wasserman for introducing me to these problems 30 years ago; doing the first continuous time Markov models.**