# Graphs for Machine Learning: Useful Metaphor or Statistical Reality?

Stephen E. Fienberg

Department of Statistics, Machine Learning Department,
Cylab, and i-Lab
Carnegie Mellon University

Mining and Learning with Graphs Workshop 2010 (MLG-2010)
KDD-2010
July 24, 2010

# Graph as Metaphors

- Representation in terms of graphs $G = \{V, E\}$, is a useful metaphor that allows us to exploit the mathematical language of graph theory and some relatively simple results.
- Graphs often provide a powerful representation for the interpretation of models.

# Characterizing Statistical Approaches to Studying Graphs

1. Specify model, or a sequence or a class of models.
2. Ask about a “sampling scheme”—mechanism for data generation.
    - Sometimes 1. and 2. go together.
3. Set up likelihood function (and specify priors).
4. Estimate and assess fit, e.g., using asymptotics.
    - Here is where algorithms fit in, e.g., IPF, EM, MCMC, variational methods.
    - Check on validity of assumptions.

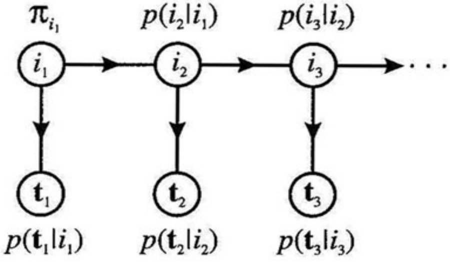
# Graphical Representations of Statistical Models

| | Variables | Individuals |
|---|---|---|
| Directed | a | b |
| Undirected | c | d |

- a—HMMs, state-space models, Bayes nets, causal models (DAGs), recursive partitioning models
- b—social networks, trees, citation and email networks
- c—covariance selection models, log-linear models
- d—relational networks, co-authorship networks

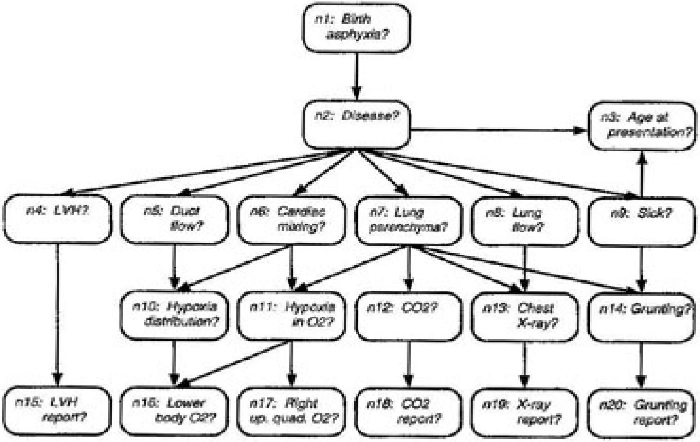
Note that a and c refer to probability models, while b and d are used to describe observed data.

# a—HMMs, State-Space Models



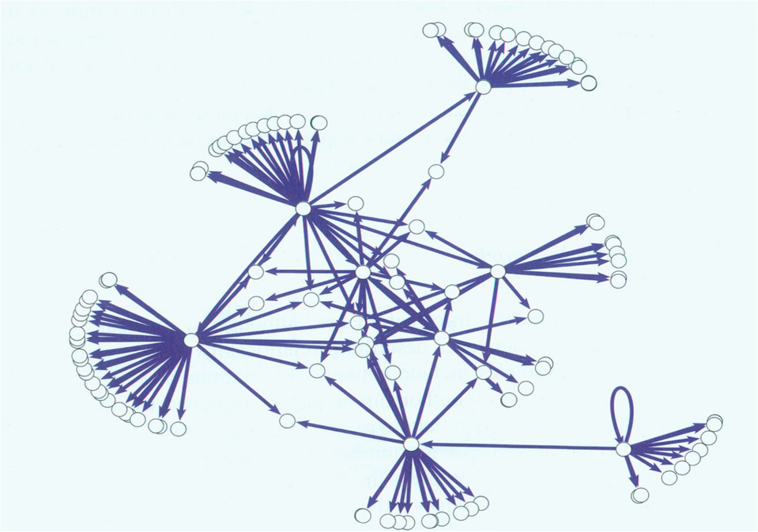
# a—Causal Models, DAGs

- CHILD network (blue babies) (Cowell et al.,1999)



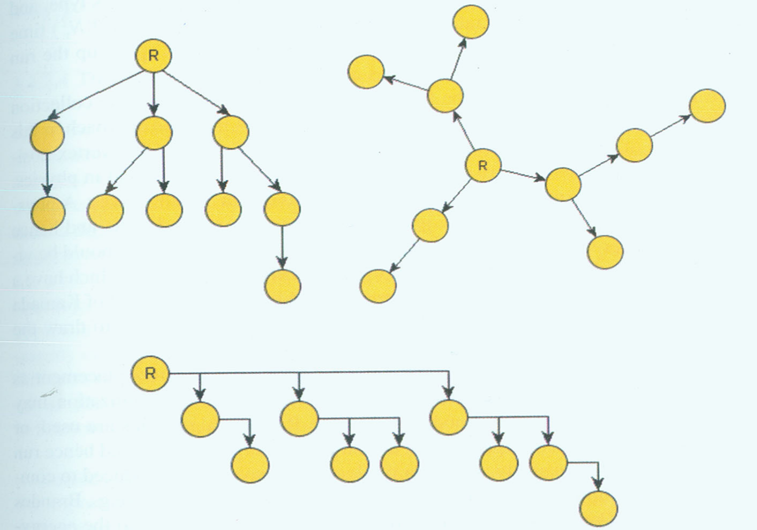
# b—Social Networks

- AIDS blog network (Kolaczyk, 2009)



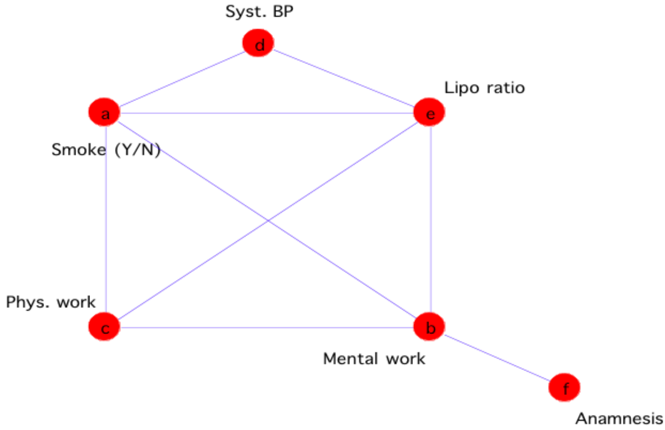
# b—Trees

- Ancestral Trees (Kolaczyk, 2009)



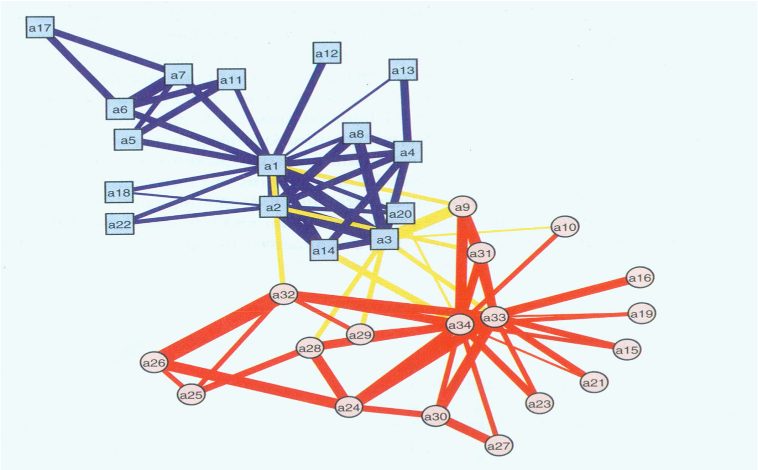
# c—Log-linear Models

- Prognostic factors for coronary heart disease for Czech autoworkers—$2^6$ table (Edwards and Hrvanek, 1985)



# d—Relational Networks

- Zachary's "karate club" network (Kolaczyk, 2009)



# d—Framingham “obesity” network

- Christakis and Fowler (2007)

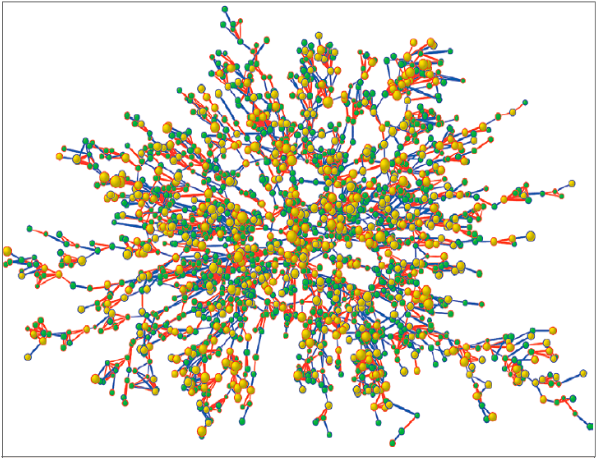
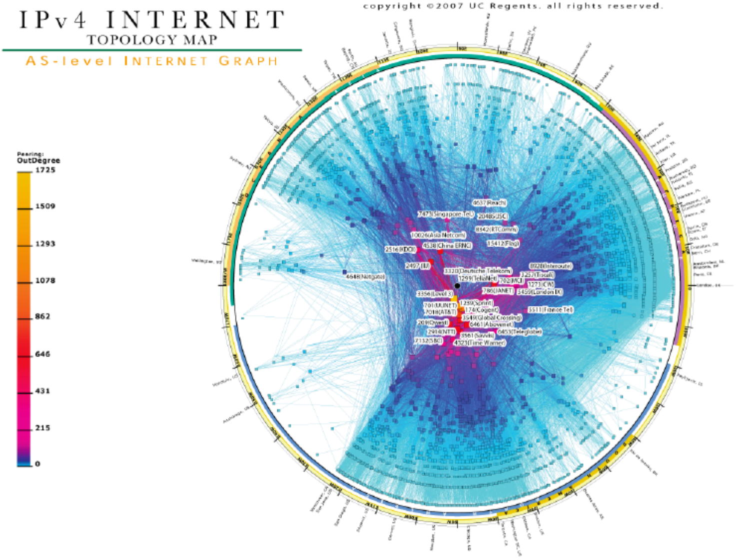


**Figure 1. Largest Connected Subcomponent of the Social Network in the Framingham Heart Study in the Year 2000.**

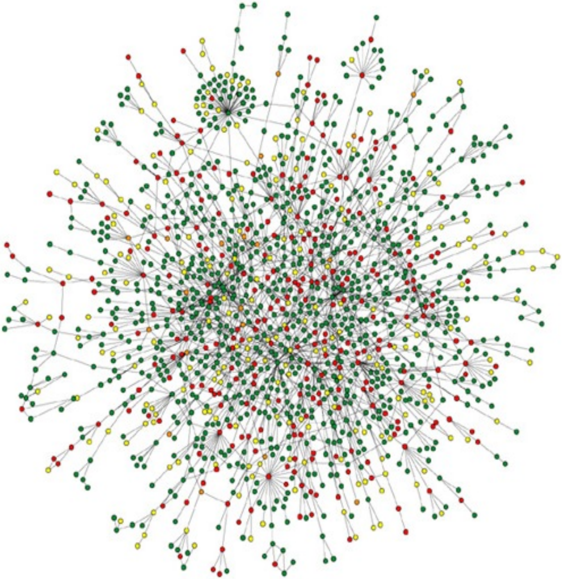
Each circle (node) represents one person in the data set. There are 2200 persons in this subcomponent of the social network. Circles with red borders denote women, and circles with blue borders denote men. The size of each circle is proportional to the person's body-mass index. The interior color of 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.

# d—Internet Topology



# d—Yeast Protein-Protein Interaction

- Airoldi et al. (2008)



# Graphical Models for Variables

- The following Markov conditions are equivalent:
  - Pairwise Markov Property: For all nonadjacent pairs of vertices, $i$ and $j$, $i \perp j \mid K \setminus \{i, j\}$.
  - Global Markov Property: For all triples of disjoint subsets of $K$, whenever $a$ and $b$ are separated by $c$ in the graph, $a \perp b \mid c$.
  - Local Markov Property: For every vertex $i$, if $c$ is the boundary set of $i$, i.e., $c = bd(i)$, and $b = K \setminus \{i \cup c\}$ , then $i \perp b \mid c$.
- All discrete graphical models are log-linear.
- The Gaussian graphical model selection problem involves estimating the zero-pattern of the inverse covariance or concentration matrix.
- For DAGs, we continue to use Markov properties but also exploit partial ordering of variables.
- Always assume individuals or units are independent r.v.'s.

# Models for Individuals/Units in Networks

- Graph describes *observed* adjacency matrix.
  - Usually use 1 for presence of an edge, and 0 for absence.
  - Can also have weights in place of 1's.
- Except for Erdös-Rényi-Gilbert model, where occurrence of edges corresponds to iid Bernoulli r.v.'s, units are *dependent.*
- Simplest generalization of E-R-G model assumes that dyads are independent—e.g., the $p_1$ model of Holland and Leinhardt, which has additional parameters for reciprocation in directed networks.
- Exponential Random Graph Models (ERGMs) that include "star" and "triangle" motifs no longer have dyadic independence.
- Can have multiple relationships measure on same individuals/units.

# Holland and Leinhardt's $p_1$ model

- $n$ nodes, random occurrence of directed edges.
- Describe the probability of an edge occurring between nodes $i$ and $j$:
  - $\log P_{ij}(0,0) = \lambda_{ij}$
  - $\log P_{ij}(1,0) = \lambda_{ij} + \alpha_i + \beta_j + \theta$
  - $\log P_{ij}(0,1) = \lambda_{ij} + \alpha_j + \beta_i + \theta$
  - $\log P_{ij}(1,1) = \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{ij}$
- 3 common forms:
  - $\rho_{ij} = 0$ (no reciprocal effect)
  - $\rho_{ij} = \rho$ (constant reciprocation factor)
  - $\rho_{ij} = \rho + \rho_i + \rho_j$ (edge-dependent reciprocation)

# Inference for Graphical Models for Variables

- For discrete r.v.'s, we use maximum likelihood estimation with asymptotics as $n \to \infty$ with $p$ fixed.
- For Gaussian r.v.'s, we use standard normal theory.
- Identifiability issues arise when $p > n$. Role for "regularization" penalties, e.g., LASSO.
- Hierarchical Bayesian models allow for smoothing:
  - Dirichlet process prior is often useful.
  - Mixtures of DP models tend to destroy graphical interpretation (i.e., conditional independence).

# Inference for Models for Individuals/Units in Networks

- Relevant asymptotics has number of nodes, $n \to \infty$.
- When there are node-specific parameters, asymptotics are far more complex.
- Maximum likelihood approaches available for ERGMs.
- For blockmodels, with constant structure within blocks, there is asymptotic theory.
  - Related literature on "community formation" and "modularity."
- Degeneracy problems arise for more general ERGMs.
- No suitable inference approaches for most "statistical physics" style models.
  - Fitting "power laws" to degree distributions is especially problematic!

# Estimation for $p_1$

- The likelihood function for the $p_1$ model is clearly of exponential family form.
- For the constant reciprocation version, we have

$$\log p_1(x) \propto x_{++}\theta + \sum_i x_{i+}\alpha_i + \sum_j x_{+j}\beta_j + \sum_{ij} x_{ij}x_{ji}\rho \quad (1)$$

- Get MLEs using iterative proportional fitting—method scales.
- Holland-Leinhardt explored goodness of fit of model empirically by comparing $\rho_{ij} = 0$ vs. $\rho_{ij} = \rho$.
  - Standard asymptotics (normality and $\chi^2$ tests) aren't applicable; no. parameters increases with no. of nodes.
- Fienberg and Wasserman used the edge-dependent reciprocation model to test $\rho_{ij} = \rho$.
- See Goldenberg et al. (2010) review of these and related models.

# Exponential Random Graph Models

- Let $X$ be a $n \times n$ adjacency matrix or a 0-1 vector of length $\binom{n}{2}$ or a point in $\{0, 1\}^n$).
- Identify a set of *network statistics*

$$t = (t_1(X), \ldots, t_k(X)) \in \mathbb{R}^k$$

  and construct a distribution such that $t$ is a vector of *sufficient statistics*.
- This leads to an *exponential family* model of the form:

$$P_\theta(X = x) = h(x) \exp\{\theta \cdot t - \psi(\theta)\},$$

  where
  - $\theta \in \Theta \subseteq \mathbb{R}^k$ is the natural parameter space;
  - $\psi(\theta)$ is a normalizing constant (often intractable);
  - $h(\cdot)$ depends on $x$ only.

# Estimation for Exponential Random Graph Models

- Pseudo-estimation using independent logistic regressions, one per node.
- MCMC.
- Problem of degeneracy or near degeneracy.
  - MLEs don't exist—maximize on the boundary.
  - Likelihood function is not well-behaved and most observable configurations are near the boundary.

# Connecting the Two Graphical Approaches

- There is link between them, not just a common metaphor.
- Frank and Strauss (1986) introduce a pairwise Markov property for individual-level undirected network models.
- $X_{ij}$ and $X_{i'j'}$ are conditionally independent given the other r.v.'s $X_{kl}$ iff they do not share a vertex.
- Set up conditional independence graph, $G^* = \{V^*, E^*\}$, where edges from original graph, $G = \{V, E\}$, are nodes.
- Distribution of $X$ satisfies the pairwise Markov property iff

$$Pr\{X = x\} \propto \exp\{\sum_{k=1}^{N_V - 1} \theta_k S_k(x) + \theta_\tau T(x)\}$$

  where $S_k(x)$ is no. of $k$-stars and $T(x)$ is no. triangles.
- Many other ERGMs don't have this property, e.g., those with alternating $k$-stars and alternating triangles.

# Roles for Latent Variables

- For graphical models for variables:
  - Natural for many models, e.g., HMMs.
  - Arise naturally in Hierarchical Bayesian structures. Hyperparameters are latent quantities.
- For models for individuals/units in networks:
  - Random effects versions of node-specific models such as $p_1$.
  - Arise naturally in hierarchical Bayesian approaches, such as Mixed Membership Stochastic Blockmodels and latent space models.
- Can also use latent structure to infer network links from data on variables for individuals, e.g., as in relational topic models.

# Role of Time/Dynamics

- For graphical models for variables:
  - Time gives ordering to variables and assists in causal models.
  - Note distinction between position of underlying “latent” quantity over time and the actual manifest measurement associated with it, which is often measured retrospectively.
- Dynamic models for individual-based networks:
  - Continuous-time stochastic process models for event data, perhaps aggregated into chinks.
  - Discrete-time transition models, perhaps embedded into continuous time process.
  - Details in Xing presentation yesterday and in Neville and Smyth presentations this morning.

# Drawing Inferences From Subnetworks and Subgraphs

- Inferences from Subgraphs
  - Conditional independence structure allows for local message passing and inference from cliques and regular subgraphs when there are separator sets that isolate components.
  - Interpretation in terms of GLM regression coeficients always depends on the other variables in the model.
- Inferences from Subnetworks
  - Most properties observed in subnetworks don't generalize to full network, and vice versa, e.g., power laws for degree distributions.
  - Problem is dependencies among nodes and boundary effects for subnetworks.
  - Missing edges are generally not missing at random, except for some sampling settings, e.g., see Handcock and Gile (2010).

# Cross-Validation for Graph Mining

- Take a subset of the data for training and cross validate on remainder, 80% for training and 20% for validation. Possibly repeat.
- Works for iid setting involving graphical models for variables.
- DOESN'T WORK FOR NETWORKS OF INDIVIDUALS!!!!
  - Except perhaps for Markov graphs à la Frank and Strauss.
  - Recall discussion from Xing's talk re approximate results.

# Summary

- Two types of settings:

| | Variables | Individuals |
|---|---|---|
| Directed | a | b |
| Undirected | c | d |

- For a and c we use conditional independence ideas to model probabilistic relations among variables.
- for b and d we use graph to represent observed data.
- Statisticians and machine learning approach should involve the following sequence:
  1. data generation process
  2. statistical models
  3. algorithms (need to be appropriate for 1. and 2.)
  4. model assessment
  5. inferences, predictions, etc.

# Some Useful References


Airoldi, E M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008) Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, **9**, 1981–2014.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* MIT Press. Reprinted by Springer (2007).

Cowell, R. G., Dawid, A. P., Lauritzen and S. L., Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems.* Springer.

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

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi E. M. (2010) A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning*, **2** (2), 129–233.

Handcock, M. S. and Gile, K. J. (2010) Modeling Social Networks from Sampled Data. *Annals of Applied Statistics*, **4**, 5–25.

# Some Useful References

Hanneke, S., Fu, W. and Xing, E. P. (2010) Discrete Temporal Models of Social Networks. *Electronic J. Statistics*, **4**, 585–605.

Lauritzen, S. (1996) *Graphical Models*. Oxford Univ. Press.

Kolaczyk, E. D. (2009) *Statistical Analysis of Network Data: Methods and Models.* Springer.

Petrović, S., Rinaldo, A., and Fienberg, S. E. (2010) Algebraic Statistics for a Directed Random Graph Model with Reciprocation. *Proceedings of the Conference on Algebraic Methods in Statistics and Probability,* Contemporary Mathematics Series, AMS.

Rinaldo, A., Fienberg, S. E. and Yi Zhou, Y. (2009) On the Geometry of Discrete Exponential Families with Application to Exponential Random Graph Models. *Electronic J. Statistics*, **3**, 446–484. See also

Two-Part Special Section on Network Modeling, *Annals of Applied Statistics*, (2010) **4**, Numbers 1 and 2.