The Exchangeable Graph Model

Edoardo Airoldi

Department of Computer Science
Lewis-Sigler Institute for Integrative Genomics
Princeton University, USA

Santa Fe Institute, December 4th 2008, Santa Fe, NM

Overview

- Statistical elements of graph data analysis
- Data is collection of measurements on pairs
  - Binary case: a graph, denoted $G = (N,Y)$
  - General case: square matrices, same N units
- This talk
  - A model that resolves measurements on pairs into node-specific binary strings via exchangeability

Agenda

- Background
- The exchangeable graph model
- Applications

Background: p* or ERG models

$$Pr\ (Y=y|\Theta=\theta) = \exp\{\ \Sigma_k\ \theta_k S_k(y) + A(\theta)\ \}$$

where $S_k(y)$ counts specific structure k, such as

- edges $S_1(y) = \Sigma_{1\leq i\leq j\leq n}\ y_{ij}$
- triangles $S_3(y) = \Sigma_{1\leq i\leq j\leq h\leq n}\ y_{ij}\ y_{ih}\ y_{jh}.$

Frank & Strauss (JASA, 1986), Snijders et al. (Soc. Met., 2004), Hanneke & Xing (LNCS, 2007)

Background: latent space models

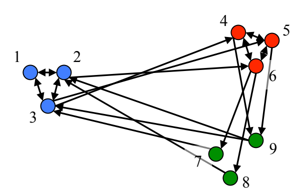
$$\text{log-odds } (Y_{ij}=1|\theta_i,\theta_j,\alpha) = \alpha - |\theta_i - \theta_j|$$

where θ_i is a point in $\mathbb{R}^k$, for all nodes i in N.

Close points in $\mathbb{R}^k$ are likely to be connected.

Hoff et at. (JASA, 2002), Sarkar & Moore (NIPS, 2006), Handcock et al. (JRSS/A, 2007, with disc.)

Background: blockmodels



$$\theta_i \sim \text{Dirichlet } (\alpha), \text{ for all nodes i in N}$$
$$y_{ij}|\theta_i,\theta_j \sim \text{Bernoulli } (\theta_i` B\ \theta_j), \text{ for all pairs (i,j)}$$

where θ_i is a point in the K-simplex, and B is K×K.

Nodes in the same block share similar connectivity.

Loraine & White (J. Math. Soc., 1971), Nowicki & Snijders (JASA, 2001), Airoldi et al. (Link-KDD 2005, JMLR, 2008), Newman & Leicht (PNAS 2007), Hofman & Wiggins (Phys. Rev Lett. 2008)

Remarks

ERG summarizes G using exp model on motif counts. Cannot offer node-specific predictions.

LSM projects Y onto a $\mathbb{R}^k$. MCMC does not scale, hard identifiability problem, no clustering effect.

MMB fractionally, sparsely maps nodes to blocks with similar connectivity, as per B. nVEM scales.

Desiderata: node attributes, sparsity, tractability

Agenda

- Background
- The exchangeable graph model
- Applications

The exchangeable graph model

b_i ~ Uniform (vertex set of hypercube), for all nodes i

$y_{ij}|b_i,b_j$ ~ Bernoulli ($q(b_i,b_j)$), for all pairs (i,j)

where b_i is a binary string, K-bit long.

A step-up in complexity from,

y_{ij} ~ Bernoulli (p), for all pairs (i,j)

Erdos & Renyi (1959), Gilbert (1959)

Specifications

- Number of bits captures complexity, $K < |N|$
- Function $q(b_i,b_j)$ is asymmetric in the arguments, e.g., consider $q = b_i`b_j/|b_i|$, where $|b_i| \neq |b_j|$
- How to control sparsity of the bit-strings?

  Concentrate density in the corners of the hypercube, h_i, then sample IID bits $b_i|h_i$. Write $\text{Bit}^K(\alpha)$

Some results

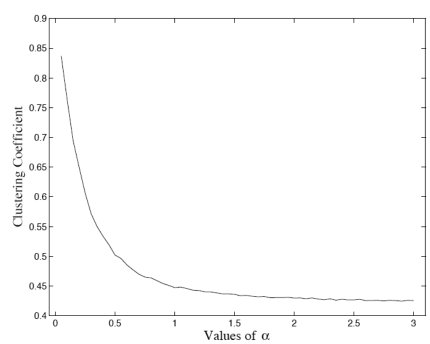
- Two sources of variability:
  - Pr (edge) decreases with K
    More complexity reduces chances of an edge
  - Pr (edge) increases with $1/\alpha$
    Concentration of density improve chances of an edge
- Emergence of the giant component in *(K,α)*

  (i) phase transition, (ii) structure in the giant comp.

Santa Fe Institute 11

Structure in the giant component

- As negative correlation among bits in node-specific strings increases, clear structure emerges



Santa Fe Institute Edo Airoldi

Inference on the bit-strings

- Model

$$b_i \sim \text{Bit}^K(\alpha), \text{ for all nodes } i$$

$$y_{ij}|b_i,b_j \sim \text{Bernoulli} \left(q(b_i,b_j) \right), \text{ for all pairs } (i,j)$$

- Likelihood

$$\mathbb{L}(Y=y|\alpha) = \int db_{1:N} \left(\Pi_{ij} \text{ Bern } (y_{ij}|b_i,b_j) \ \Pi_i \text{ Bit}^K(b_i|\alpha) \right)$$

- Variational EM algorithm; approximate E-Step

Agenda

- Background
- The exchangeable graph model
- Applications

Assessing graph complexity

Make inference on bit strings

- How many bits to explain observed connectivity with high probability?
- How much information is retained at different bit-lengths? For instance, compute information profile for $K \leq N$, or entropy histogram for a given k.

Performing model comparison

Consider statistical models of paired measurements

- Degree distribution, suite of properties, likelihood
- Alternative:
  1. given graph G, fit models $p_{1:M}(G|\Theta)$
  2. sample B graphs from $p_{1:M}(G|\Theta_{EST})$
  3. compute M distributions on EHs and IPs
  4. compare models:
     - complexity of models' supports
     - similarity of graph and model complexity

P-value of size of modules' overlap

In general

- Distribution $H^K(\alpha)$ specifies multiple membership
- EGM gives model-based probability of overlap size, e.g. via empirical null *(Efron, JASA, 2003)*

Genomics example

- Size of common neighborhood is used to infer gene duplication and loss, given evolutionary tree

Concluding remarks

1. In theory

   New tool to explore variability of graph connectivity

2. In practice

   i. Likelihood-based approach to size of neighborhoods

   ii. Information-based approach to goodness-of-fit and model comparison for models of paired measurements