Hierarchical Block Models

Lecture 19
14 November 2013
CSCI 5352, Network Analysis and Models

Prof. Aaron Clauset
University of Colorado, Boulder
no structure
no structure | modular structure
---|---
![Image](image1.png) | ![Image](image2.png)

one scale
no structure  modular structure  hierarchical structure

one scale  multi-scale
how can we measure a network’s hierarchy?

step 1: network data

step 3: hierarchy
One Approach

model-based inference

1. describe how to generate hierarchies (a model)
2. estimate / learn model from data (algorithms)
3. test fitted model(s)
4. extract predictions, insight
A Model of Hierarchy
A Model of Hierarchy

$D$, $\{p_r\}$

probability $p_r$

assortative modules
"inhomogeneous" random graph

\[
\Pr(i, j \text{ connected}) = p_r
\]

\[
= P(\text{lowest common ancestor of } i, j)
\]
Hierarchical Random Graph

- explicit model = explicit assumptions
- flexible (2n parameters)
- captures structure at all scales
- mixtures of assortativity, disassortativity
- decomposition into set of random bipartite graphs
- learnable directly from data
Learning From Data

a direct approach

- **likelihood function** \( \mathcal{L} = \Pr( \text{data} | \text{model} ) \) 
  (\( \mathcal{L} \) scores **quality** of model)

- sample all good models
  via Markov chain Monte Carlo*
  over all dendrograms

- **technical details in**

* other sampling or optimization methods possible
Likelihood Function

\[ L(D, \{p_r\}) = \prod_{r} p_r^{E_r} (1 - p_r)^{L_r R_r - E_r} \]

- \( L_r \) = number of nodes in left subtree
- \( R_r \) = number of nodes in right subtree
- \( E_r \) = number of edges with \( r \) as lowest common ancestor
Example
Bad Dendrogram

\[ \mathcal{L}(\mathcal{D}, \{p_r\}) = \prod_r p_r^{E_r} (1 - p_r)^{L_r R_r - E_r} \]
Bad Dendrogram

\[ \mathcal{L}(D, \{p_r\}) = \prod_r p_r^{E_r} (1 - p_r)^{L_r R_r - E_r} \]

\[ \mathcal{L} = \left[ \left( \frac{1}{3} \right)^1 \left( \frac{2}{3} \right)^2 \right] \cdot \left[ \left( \frac{1}{4} \right)^2 \left( \frac{3}{4} \right)^6 \right] \]

\[ \mathcal{L} = 0.0016 \]
Good Dendrogram

\[ \mathcal{L}(\mathcal{D}, \{p_r\}) = \prod_r p_r^{E_r} (1 - p_r)^{L_r R_r - E_r} \]
Good Dendrogram

\[ \mathcal{L}(D, \{p_r\}) = \prod_r p_r^{E_r} (1 - p_r)^{L_r R_r - E_r} \]

\[ \mathcal{L} = \left[ \left( \frac{1}{9} \right)^1 \left( \frac{8}{9} \right)^8 \right] \]

\[ \mathcal{L} = 0.0433 \]
Markov chain Monte Carlo (MCMC)

Given $D$, choose random internal node

Choose random reconfiguration of subtrees

Recompute probabilities $\{p_r\}$ and likelihood $\mathcal{L}$

Sampling states according to their likelihood

three subtree configurations
(up to relabeling)
Some applications
Two Case Studies

NCAA Schedule 2000
\[ n = 115 \quad m = 613 \]

Zachary’s Karate Club
\[ n = 34 \quad m = 78 \]
Mixing Times

MCMC mixes relatively quickly

Equilibrium in $\sim O(n^2)$ steps
Hierarchies

point estimate

consensus hierarchy
Edge Annotations

Average likelihood of edge existing

- For each edge \((i, j)\) in \(G\), compute average associated parameter \(\langle \theta_r \rangle_{(i,j)}\) over sampled models

- \(\langle \theta_r \rangle_{(i,j)}\) is edge annotation (weight)
Vertex Annotations

Group-affiliation strengths

• If each vertex has known group label
• Ask, how often does vertex $i$ appear in a subtree with majority of its fellows?
• Frequency is vertex annotation (strength)
Edge, Note Annotations
From Graph to Ensemble
From Graph to Ensemble

- Given graph $G$
- run MCMC to equilibrium
- then, for each sampled $\mathcal{D}$, draw a resampled graph $G'$ from ensemble

A test: do resampled graphs look like original?
Degree Distribution
Clustering Coefficient

![Graph showing fraction of graphs with clustering coefficient vs. clustering coefficient](image)

- **Original**
- **Resampled**

The graph plots the fraction of graphs with a clustering coefficient against the clustering coefficient, showing a peak at a certain value.
**Distance Distribution**

![Graph showing distance distribution](image)

- **Fraction of vertex-pairs at distance $d$**
- **Distance, $d$**

- **Original** curve
- **Resampled** curve
Missing Links

many networks partially known, noisy
  • social nets, food webs, protein interactions, etc.

can hierarchies predict their missing links?

previous approaches
  • Liben-Nowell & Kleinberg (2003)
  • Goldberg & Roth (2003)
  • Szilágyi et al. (2005)
  • many more now
Accuracy is Hard

- remove $k$ edges from $G$
- how easy to guess a missing link?

\[
p_{\text{guess}} \approx \frac{k}{n^2 - m + k} = O(n^{-2})
\]

$n = 75$
$m = 113$

\[
p_{\text{guess}} = \frac{k}{2662 + k}
\]
Given incomplete graph \( G \)
run MCMC to equilibrium
then, over sampled \( \mathcal{D} \), compute average \( \langle p_r \rangle \) for links \((i, j) \notin G\)
predict links with high \( \langle p_r \rangle \) values are missing

Test via leave-\( k \)-out cross-validation
perfect accuracy: AUC = 1
no better than chance: AUC = 1/2
Scoring the Predictions

ROC curve

AUC = area under curve

\[ \text{AUC} = \Pr(\text{distinguish } + \text{ from } - ) \]
Performance 1

Grassland species network

- Pure chance
- Common neighbors
- Jaccard coeff.
- Degree product
- Shortest paths
- Hierarchical structure

AUC
Fraction of edges observed, k/m

hierarchy
simple predictors
pure chance
Performance 2

Terrorist association network

- Pure chance
- Common neighbors
- Jaccard coefficient
- Degree product
- Shortest paths
- Hierarchical structure

AUC vs Fraction of edges observed

T. pallidum metabolic network

- Pure chance
- Common neighbors
- Jaccard coefficient
- Degree product
- Shortest paths
- Hierarchical structure

AUC vs Fraction of edges observed
Some Final Thoughts

• what processes create these hierarchical structures?
• scaling up the running time from $O(n^2)$?
• active learning
• generalization to weighted, directed edges
• generalization to non-Poisson distributions