Lecture 6 (supplemental): Stochastic Block Models

Aaron Clauset
@aarontlausset
Assistant Professor of Computer Science
University of Colorado Boulder
External Faculty, Santa Fe Institute

© 2017 Aaron Clauset
what is structure?
what is structure?

- makes data different from noise
  - makes a network different from a random graph
what is structure?

• makes data different from noise
  ▶ makes a network different from a random graph

• helps us compress the data
  ▶ describe the network succinctly
  ▶ capture most relevant patterns
what is structure?

- makes data different from noise
  - makes a network different from a random graph
- helps us compress the data
  - describe the network succinctly
  - capture most relevant patterns
- helps us generalize,
  from data we’ve seen to data we haven’t seen:
  1. from one part of network to another
  2. from one network to others of same type
  3. from small scale to large scale (coarse-grained structure)
  4. from past to future (dynamics)
• imagine graph $G$ is drawn from an ensemble or generative model: a probability distribution $\Pr(G|\theta)$ with parameters $\theta$

• $\theta$ can be continuous or discrete; represents structure of graph
statistical inference

- imagine graph $G$ is drawn from an ensemble or generative model: a probability distribution $\Pr(G \mid \theta)$ with parameters $\theta$
- $\theta$ can be continuous or discrete; represents structure of graph
- inference (MLE): given $G$, find $\theta$ that maximizes $\Pr(G \mid \theta)$
- inference (Bayes): compute or sample from posterior distribution $\Pr(\theta \mid G)$
statistical inference

- imagine graph $G$ is drawn from an ensemble or generative model: a probability distribution $\Pr(G \mid \theta)$ with parameters $\theta$
- $\theta$ can be continuous or discrete; represents structure of graph
- inference (MLE): given $G$, find $\theta$ that maximizes $\Pr(G \mid \theta)$
- inference (Bayes): compute or sample from posterior distribution $\Pr(\theta \mid G)$

- if $\theta$ is partly known, constrain inference and determine the rest
- if $G$ is partly known, infer $\theta$ and use $\Pr(G \mid \theta)$ to generate the rest
- if model is good fit (application dependent), we can generate synthetic graphs structurally similar to $G$
- if part of $G$ has low probability under model, flag as possible anomaly
statistical inference

- imagine graph $G$ is drawn from an ensemble or generative model: a probability distribution $\Pr(G|\theta)$ with parameters $\theta$
- $\theta$ can be continuous or discrete; represents structure of graph
- inference (MLE): given $G$, find $\theta$ that maximizes $\Pr(G|\theta)$
- inference (Bayes): compute or sample from posterior distribution $\Pr(\theta|G)$
- if $\theta$ is partly known, constrain inference and determine the rest
- if $G$ is partly known, infer $\theta$ and use $\Pr(G|\theta)$ to generate the rest
- if model is good fit (application dependent), we can generate synthetic graphs structurally similar to $G$
- if part of $G$ has low probability under model, flag as possible anomaly

statistical inference = principled approach to learning from data
combines tools from statistics, machine learning, information theory, and statistical physics
quantifies uncertainty
separates the model from the learning
statistical inference: key ideas

• interpretability
  model parameters have meaning for scientific questions

• auxiliary information
  node & edge attributes, temporal dynamics (beyond static binary graphs)

• scalability
  fast algorithms for fitting models to big data (methods from physics, machine learning)

• model selection
  which model is better? is this model bad? how many communities?

• partial or noisy data
  extrapolation, interpolation, hidden data, missing data

• anomaly detection
  low probability events under generative model
• define a parametric probability distribution over networks \( \Pr(G \mid \theta) \)

• **generation**: given \( \theta \), draw \( G \) from this distribution

• **inference**: given \( G \), choose \( \theta \) that makes \( G \) likely

\[
\Pr(G \mid \theta) \quad G = (V, E)
\]
generative models for complex networks

**general form**

$$\Pr(G \mid \theta) = \prod_{ij} \Pr(A_{ij} \mid \theta)$$

assumptions about “structure” go into $\Pr(A_{ij} \mid \theta)$

consistency $\lim_{n \to \infty} \Pr(\hat{\theta} \neq \theta) = 0$

requires that edges be conditionally independent [Shalizi, Rinaldo 2011]

two general classes of these models
generative models for complex networks

stochastic block models

\( k \) types of vertices, \( \Pr(A_{ij} \mid M, z) \) depends only on node types \( z_i, z_j \)

originally invented by sociologists [Holland, Laskey, Leinhardt 1983]

many, many flavors, including

- simple assortative SBM [Hofman & Wiggins 2008]
- mixed-membership SBM [Airoldi et al. 2008]
- hierarchical SBM [Clauset et al. 2006, 2008, Peixoto 2014]
- fractal SBM [Leskovec et al. 2005]
- infinite relational model [Kemp et al. 2006]
- degree-corrected SBM [Karrer & Newman 2011]
- SBM + topic models [Ball et al. 2011]
- SBM + vertex covariates [Mariadassou et al. 2010, Newman & Clauset 2016]
- bipartite SBM [Larremore et al. 2014]
- multilayer SBM [Peixoto 2015, Valles-Catata et al. 2016]

and many others
latent space models

- nodes live in a latent space, \( \Pr(A_{ij} \mid f(x_i, x_j)) \) depends only on vertex-vertex proximity

originally invented by statisticians \([\text{Hoff, Raftery, Handcock } 2002]\)

many, many flavors, including

- logistic function on vertex features \([\text{Hoff et al. } 2002]\)
- social status / ranking \([\text{Ball, Newman } 2013]\)
- nonparametric metadata relations \([\text{Kim et al. } 2012]\)
- multiplicative attribute graphs \([\text{Kim & Leskovec } 2010]\)
- nonparametric latent feature model \([\text{Miller et al. } 2009]\)
- infinite multiple memberships \([\text{Morup et al. } 2011]\)
- ecological niche model \([\text{Williams et al. } 2010]\)
- hyperbolic latent spaces \([\text{Boguna et al. } 2010]\)

and many others
opportunities and challenges

- **richly annotated data**
  - edge weights, node attributes, time, etc.
  - = new classes of generative models

- **generalize from** \( n = 1 \) **to ensemble**
  - useful for modeling checking, simulating other processes, etc.

- **many familiar techniques**
  - frequentist and Bayesian frameworks
  - makes probabilistic statements about observations, models
  - predicting missing links \( \approx \) leave-k-out cross validation
  - approximate inference techniques (EM, VB, BP, etc.)
  - sampling techniques (MCMC, Gibbs, etc.)

- **learn from partial or noisy data**
  - extrapolation, interpolation, hidden data, missing data
opportunities and challenges

- only two classes of models
  - stochastic block models (categorical latent variables)
  - latent space models (ordinal / continuous latent variables)
- bootstrap / resampling for network data
  - critical missing piece
  - depends on what is independent in the data
- model comparison
  - naive AIC, BIC, marginalization, LRT can be wrong for networks
  - what is goal of modeling: realistic representation or accurate prediction?
- model assessment / checking?
  - how do we know a model has done well? what do we check?
- what is v-fold cross-validation for networks?
  - Omit $n^2/v$ edges? Omit $n/v$ nodes? What?
the stochastic block model

- each vertex $i$ has type $z_i \in \{1, \ldots, k\}$ ($k$ vertex types or groups)
- stochastic block matrix $M$ of group-level connection probabilities
- probability that $i, j$ are connected $= M_{z_i, z_j}$

*community = vertices with same pattern of inter-community connections*
the stochastic block model

- ** assortative**: edges within groups
- ** disassortative**: edges between groups
- ** ordered**: linear group hierarchy
- ** core-periphery**: dense core, sparse periphery
the stochastic block model

likelihood function

the probability of $G$ given labeling $z$ and block matrix $M$

$$\Pr(G \mid z, M) = \prod_{(i, j) \in E} M_{z_i, z_j} \prod_{(i, j) \not\in E} (1 - M_{z_i, z_j})$$

edge / non-edge probability
the stochastic block model

likelihood function

the probability of $G$ given labeling $z$ and block matrix $M$

$$\Pr(G \mid z, M) = \prod_{(i,j) \in E} M_{z_i, z_j} \prod_{(i,j) \notin E} (1 - M_{z_i, z_j})$$

$$= \prod_{rs} M_{r,s}^{e_{r,s}} (1 - M_{r,s})^{n_r n_s - e_{r,s}}$$

(Bernoulli edges)

Bernoulli random graph with parameter $M_{r,s}$
the stochastic block model

the most general SBM

\[
\Pr(A | z, \theta) = \prod_{i,j} f(A_{ij} | \theta_{\mathcal{R}(z_i,z_j)})
\]

\(A_{ij}\) : value of adjacency
\(\mathcal{R}\) : partition of adjacencies
\(f\) : probability function
\(\theta_{a,*}\) : pattern for \(a\)-type adjacencies

Binomial = simple graphs
Poisson = multi-graphs
Normal = weighted graphs
etc.

\[
\begin{array}{cccc}
\theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\
\theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} \\
\theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} \\
\theta_{41} & \theta_{42} & \theta_{43} & \theta_{44}
\end{array}
\]
the stochastic block model

degree-corrected SBM ($f = \text{Poisson}$)
the stochastic block model

degree-corrected SBM ($f = \text{Poisson}$)

key assumption \( \Pr(i \to j) = \theta_i \theta_j \omega_{z_i, z_j} \)

stochastic block matrix \( \omega_{r,s} \)

(degree) propensity of node \( \theta_i \)

likelihood:

\[
\Pr(A | z, \theta, \omega) = \prod_{i<j} \frac{(\theta_i \theta_j \omega_{z_r, z_j})^{A_{ij}}}{A_{ij}!} \exp \left( -\theta_i \theta_j \omega_{z_r, z_j} \right)
\]

where \( \hat{\theta}_i = \frac{k_i}{\sum_j k_j \delta_{z_i, z_j}} \)

\( \hat{\omega}_{rs} = m_{rs} = \sum_{ij} A_{ij} \delta_{z_i, r} \delta_{z_j, s} \)

fraction of \( i \)'s group's stubs on \( i \)

total number of edges between \( r \) and \( s \)
the stochastic block model

comparing SBM vs. DC-SBM : Zachary karate club

the stochastic block model

comparing SBM vs. DC-SBM: Zachary karate club

SBM
leader/follower division

DC-SBM
social group division

the stochastic block model

comparing SBM vs. DC-SBM: Zachary karate club

SBM
leader/follower division

DC-SBM
social group division

comparing SBM vs. DC-SBM: Zachary karate club

Peel, Larremore, Clauset, forthcoming (2016)
the stochastic block model

comparing SBM vs. DC-SBM: Zachary karate club

SBM
leader/follower division

DC-SBM
social group division

Peel, Larremore, Clauset, forthcoming (2016)
extending the SBM

many variants! we’ll cover three:

• bipartite community structure
• weighted community structure
• hierarchical community structure
bipartite networks

many networks are bipartite

- scientists and papers (co-authorship networks)
- actors and movies (co-appearance networks)
- words and documents (topic modeling)
- plants and pollinators
- genes and genomes
- etc.

bipartite networks

many networks are bipartite

• scientists and papers (co-authorship networks)
• actors and movies (co-appearance networks)
• words and documents (topic modeling)
• plants and pollinators
• genes and genomes
• etc.

most analyses focus on one-mode projections which discard information

bipartite networks

bipartite stochastic block model (biSBM)

• exactly the SBM, but model knows network is bipartite

• if \( \text{type}(z_i) = \text{type}(z_j) \)

  then require \( M_{z_i, z_j} = 0 \)

• inference proceeds as before

bipartite networks

SBM can learn bipartite structure on its own

but often over fits or returns mixed-type groups

bipartite networks

bipartite stochastic block model (biSBM)

- how do we know it works well?

bipartite networks

bipartite stochastic block model (biSBM)

- how do we know it works well?
- synthetic networks with planted partitions

\[
\omega_{\text{planted}} = \begin{pmatrix}
\alpha & 0 & 0 & 0 \\
0 & \beta & 0 & 0 \\
0 & 0 & \gamma & 0 \\
0 & 0 & 0 & \delta \\
\end{pmatrix}
\]

**easy case:**
4x4 easy-to-see communities

**hard case:**
3x2 hard-to-see communities

**bipartite networks**

**bipartite stochastic block model (biSBM)**

- how do we know it works well?
- synthetic networks with *planted partitions*

\[
\omega_{\text{planted}} = \begin{pmatrix}
\cdots & \alpha & 0 & 0 & 0 \\
\cdots & 0 & \beta & 0 & 0 \\
\cdots & 0 & 0 & \gamma & 0 \\
\cdots & 0 & 0 & 0 & \delta \\
\alpha & 0 & 0 & 0 & \cdots \\
0 & \beta & 0 & 0 & \cdots \\
0 & 0 & \gamma & 0 & \cdots \\
0 & 0 & 0 & \delta & \cdots \\
\end{pmatrix}
\]

**easy case:**
4x4 easy-to-see communities

**hard case:**
3x2 hard-to-see communities

bipartite networks

bipartite stochastic block model (biSBM)

- how do we know it works well?

- synthetic networks with *planted partitions*

---

bipartite networks

bipartite stochastic block model (biSBM)

• always find pure-type communities

• more accurate than modeling one-mode projections (even weighted projections)

• finds communities in both modes

bipartite networks

example 1: malaria gene network

bipartite networks

example 2: Southern Women network

bipartite networks

example 3: IMDb

bipartite networks

other approaches

minimum description length (MDL) principle

*learns* that a network is bipartite

marginalize over bipartite SBM parameterizations
bipartite networks
weighted networks

most interactions are weighted

• frequency of interaction
• strength of interaction
• outcome of interaction
• etc.

• but! thresholding discards information and can obscure underlying structure
weighted networks

weighted networks

Aicher et al., J. Complex Networks 3, 221-228 (2015).
Valued Ties Tell Fewer Lies: Why Not To Dichotomize Network Edges With Thresholds*

Andrew C. Thomas† Joseph K. Blitzstein‡

• how will the results depend on the threshold?
• what impact does noise have, under threshold?
recall...

the most general SBM

\[ \Pr(A | z, \theta) = \prod_{i,j} f(A_{ij} | \theta \mathcal{R}(z_i, z_j)) \]

\( A_{ij} \) : value of adjacency

\( \mathcal{R} \) : partition of adjacencies

\( f \) : probability function

\( \theta_{a,*} \) : pattern for \( a \)-type adjacencies

<table>
<thead>
<tr>
<th>\theta_{11}</th>
<th>\theta_{12}</th>
<th>\theta_{13}</th>
<th>\theta_{14}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\theta_{21}</td>
<td>\theta_{22}</td>
<td>\theta_{23}</td>
<td>\theta_{24}</td>
</tr>
<tr>
<td>\theta_{31}</td>
<td>\theta_{32}</td>
<td>\theta_{33}</td>
<td>\theta_{34}</td>
</tr>
<tr>
<td>\theta_{41}</td>
<td>\theta_{42}</td>
<td>\theta_{43}</td>
<td>\theta_{44}</td>
</tr>
</tbody>
</table>

Binomial = simple graphs
Poisson = multi-graphs
Normal = weighted graphs etc.
weighted networks

weighted stochastic block model (WSBM)

- model edge existence and edge weight separately
- edge existence: SBM
- edge weights: exponential family distribution

log-likelihood:

\[
\ln \Pr(G \mid M, z, \theta, f) = \alpha \ln \Pr(G \mid M, z)
+ (1 - \alpha) \ln \Pr(G \mid \theta, z, f)
\]

Aicher et al., *J. Complex Networks* 3, 221-228 (2015).
weighted networks

weighted stochastic block model (WSBM)

- model edge existence and edge weight separately
- edge existence: SBM
- edge weights: exponential family distribution

log-likelihood:

$$\ln \Pr(G \mid M, z, \theta, f) = \alpha \ln \Pr(G \mid M, z) + (1 - \alpha) \ln \Pr(G \mid \theta, z, f)$$

mixing parameter

$\alpha = 1$ only model edge existence (ignore weights)
$\alpha = 0$ only model edge weights (ignore non-edges)

Aicher et al., J. Complex Networks 3, 221-228 (2015).
American football, NFL 2009 season

- 32 teams, 2 “divisions”, 4 “subdivisions”
- edge existence: who plays whom
- edge weight: mean score difference
American football, NFL 2009 season

- 32 teams, 2 “divisions”, 4 “subdivisions”
- SBM ($\alpha = 1$) recovers subdivisions perfectly
American football, NFL 2009 season

- 32 teams, 2 “divisions”, 4 “subdivisions”
- WSBM ($\alpha = 0$) recovers team skill hierarchy

weighted networks

Aicher et al., J. Complex Networks 3, 221-228 (2015).
adding weights to the SBM

• what does $A_{ij} = 0$ mean?
  no edge or weight=0 edge or non-observed edge?
• how will we model the distribution of edge weights?
• edge existences and edge weights may contain different large-scale structure (conference structure vs. skill hierarchy)
weighted networks

other approaches
weighted networks

other approaches

bin the edge weights into \( C \) ranges

\( C \) bins = \( C \) layers of a multi-layer SBM

common node labels

each layer has its own block matrix

infer \( C, M_c, z \)
weighted networks

other approaches (discretized weights = SBM layers)

OpenFlights airport network
edge weights = distance (km)
weighted networks

other approaches (discretized weights = SBM layers)

OpenFlights airport network

fin