Generative Models for Complex Network Structure

Aaron Clauset
@aaronclauset
Computer Science Dept. & BioFrontiers Institute
University of Colorado, Boulder
External Faculty, Santa Fe Institute
• what is structure?

• generative models for complex networks
  ➤ general form
  ➤ types models
  ➤ opportunities and challenges

• weighted stochastic block models
  ➤ a parable about thresholding
  ➤ checking our models
  ➤ learning from data (approximately)
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:
  i. from one part of network to another
  ii. from one network to others of same type
  iii. from small scale to large scale (coarse-grained structure)
  iv. from past to future (dynamics)
statistical inference

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

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

- if $\theta$ is partly known, constrain inference and determine the rest
- if $G$ is partly known, infer $\theta$ and use $P(G \mid \theta)$ 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
• what is structure?

• **generative models for complex networks**
  ▶ general form
  ▶ types models
  ▶ opportunities and challenges

• **weighted stochastic block models**
  ▶ a parable about thresholding
  ▶ checking our models
  ▶ learning from data (approximately)
generative models for complex networks

general form

\[ P(G \mid \theta) = \prod_{i<j} P(A_{ij} \mid \theta) \]

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

consistency

\[ \lim_{n \to \infty} \Pr \left( \hat{\theta} \neq \theta \right) = 0 \]

requires that edges be conditionally independent [Shalizi, Rinaldo 2011]
$\mathcal{D}$, $\{p_r\}$

probability $p_r$
Pr($i, j$ connected) = $p_r$

= $P(\text{lowest common ancestor of } i, j)$
\[ \mathcal{L}(D, \{p_r\}) = \prod_{r} p_r^{E_r} (1 - p_r)^{L_r R_r - E_r} \]

- \(L_r\) = number nodes in left subtree
- \(R_r\) = number nodes in right subtree
- \(E_r\) = number edges with \(r\) as lowest common ancestor
classes of generative models

- stochastic block models
  \[ P(A_{ij} | z_i, z_j) \] depends only on types of \( i, j \)
  originally invented by sociologists [Holland, Laskey, Leinhardt 1983]

  many, many flavors, including
  - mixed-membership SBM [Airoldi, Blei, Feinberg, Xing 2008]
  - hierarchical SBM [Clauset, Moore, Newman 2006, 2008]
  - restricted hierarchical SBM [Leskovec, Chakrabarti, Kleinberg, Faloutsos 2005]
  - infinite relational model [Kemp, Tenenbaum, Griffiths, Yamada, Ueda 2006]
  - restricted SBM [Hofman, Wiggins 2008]
  - degree-corrected SBM [Karrer, Newman 2011]
  - SBM + topic models [Ball, Karrer, Newman 2011]
  - SBM + vertex covariates [Mariadassou, Robin, Vacher 2010]
  - SBM + edge weights [Aicher, Jacobs, Clauset 2013]
  + many others
classes of generative models

• latent space models
  nodes live in a latent space, $P(A_{ij} | f(x_i, x_j))$ depends only on vertex-vertex proximity

  many, many flavors, including
  logistic function on vertex features [Hoff, Raftery, Handcock 2002]
  social status / ranking [Ball, Newman 2013]
  nonparametric metadata relations [Kim, Hughes, Sudderth 2012]
  multiple attribute graphs [Kim, Leskovec 2010]
  nonparametric latent feature model [Miller, Griffiths, Jordan 2009]
  infinite multiple memberships [Morup, Schmidt, Hansen 2011]
  ecological niche model [Williams, Anandanadesan, Purves 2010]
  hyperbolic latent spaces [Boguna, Papadopoulos, Krioukov 2010]
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
  - latent space models

- 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?
• what is structure?

• generative models for complex networks
  ▶ general form
  ▶ types models
  ▶ opportunities and challenges

• weighted stochastic block models
  ▶ a parable about thresholding
  ▶ learning from data (approximately)
  ▶ checking our models
functional groups, not just clumps

• social “communities” (large, small, dense or empty)
• social: leaders and followers
• word adjacencies: adjectives and nouns
• economics: suppliers and customers
nodes have discrete attributes
each vertex $i$ has type $t_i \in \{1, \ldots, k\}$
$k \times k$ matrix $p$ of connection probabilities
if $t_i = r$ and $t_j = s$, edge $(i \rightarrow j)$ exists with probability $p_{rs}$
$p$ not necessarily symmetric, and we do not assume $p_{rr} > p_{rs}$
given some $G$, we want to simultaneously
label nodes (infer type assignment $t : V \rightarrow \{1, \ldots, k\}$)
learn the latent matrix $p$

classic stochastic block model
classic stochastic block model

model

assortative modules

instance

Pr(node in i connected to node in j) = p_{i,j}

likelihood

\[ P(G | t, \theta) = \prod_{(i,j) \in E} p_{t_i,t_j} \prod_{(i,j) \notin E} (1 - p_{t_i,t_j}) \]
thresholding edge weights

- 4 groups
- edge weights $\sim N(\mu_i, \sigma^2)$ with $\mu_1 < \mu_2 < \mu_3 < \mu_4$
- what threshold $t$ should we choose? $t = 1, 2, 3, 4$
• 4 groups
• edge weights $\sim N(\mu_i, \sigma^2)$ with $\mu_1 < \mu_2 < \mu_3 < \mu_4$
  ➤ set threshold $t \leq 1$, fit SBM
• 4 groups
• edge weights $\sim N(\mu_i, \sigma^2)$ with $\mu_1 < \mu_2 < \mu_3 < \mu_4$
• set threshold $t = 2$, fit SBM
• 4 groups
• edge weights \( \sim N(\mu_i, \sigma^2) \) with \( \mu_1 < \mu_2 < \mu_3 < \mu_4 \)
  ➤ set threshold \( t = 3 \), fit SBM
• 4 groups
• edge weights $\sim N(\mu_i, \sigma^2)$ with $\mu_1 < \mu_2 < \mu_3 < \mu_4$
  ➤ set threshold $t \geq 4$, fit SBM
weighted stochastic block model

adding auxiliary information:

- each edge has weight $w(i, j)$
- let $w(i, j) \sim f(x|\theta)$

\[
= h(x) \exp(T(x) \cdot \eta(\theta))
\]

- covers all exponential-family type distributions:
  - bernoulli, binomial (classic SBM), multinomial
  - poisson, beta
  - exponential, power law, gamma
  - normal, log-normal, multivariate normal
weighted stochastic block model

adding auxiliary information:

each edge has weight \( w(i, j) \)

let \( w(i, j) \sim f(x|\theta) \)

\[
= h(x) \exp(T(x) \cdot \eta(\theta))
\]

examples of weighted graphs:

- frequency of social interactions (calls, txt, proximity, etc.)
- cell-tower traffic volume
- other similarity measures
- time-varying attributes
- missing edges, active learning, etc.
weighted stochastic block model

- block structure: \( \mathcal{R} : k \times k \rightarrow \{1, \ldots, R\} \)
- weight distribution: \( f \)
- block assignment: \( z \)
- weighted graph: \( G \)
- likelihood function:
  \[
P(G \mid z, \theta, f) = \prod_{i<j} f(G_{i,j} \mid \theta_{\mathcal{R}(z_i, z_j)})
  \]

- given \( G \) and choice of \( f \), learn \( z \) and \( \theta \)

Technical difficulties:
- degeneracies in likelihood function
  (variance can go to zero. oops)
approximate learning

- edge generative model $P(G \mid z, \theta, f)$
- estimate model via variational Bayes
  - conjugate priors solve degeneracy problem
  - algorithms for dense and sparse graphs
approximate posterior distribution
\[ \pi^*(z, \theta \mid G) \approx q(z, \theta) = \prod_i q_i(z_i) \prod_r q(\theta_r) \]

estimate \( q \) by minimizing
\[ D_{KL}(q \| \pi^*) = \ln P(G \mid z, \theta, f) - \mathcal{G}(q) \]

where \( \mathcal{G}(q) = \mathbb{E}_q(\mathcal{L}) + \mathbb{E}_q \left( \log \frac{\pi(z, \theta)}{q(z, \theta)} \right) \)

for (conjugate) prior \( \pi \) for exponential family distribution \( f \)

taking derivative yields update equations for \( z, \theta \)

iterating equations yields local optima
checking the model

synthetic network with known structure

- given synthetic graph with known structure
- run VB algorithm to convergence
- compare against choose threshold + SBM (and others)

compute Variation of Information (partition distance)

\[ \text{VI}(P_1, P_2) \in [0, \ln N] \]

in this case \( \text{VI}(P_1, P_2) \in [0, \ln k^* + 1.5] = [0, 3.1] \)
checking the model

synthetic network with known structure

• variation of Newman’s four-groups test
• $k^* = 5$ latent groups
  $n_r = [48, 16, 32, 48, 16]$  
• Normal edge weights:
  \[ f = \mathcal{N}(\mu_r, \sigma_r^2) \]

in this case $\text{VI}(P_1, P_2) \in [0, \ln k^* + 1.5] = [0, 3.1]$
learn better with more data

increase network size $N$

- fix $k = k^*$, $f = N$
- bigger network, more data

we keep the $n_r/N$ constant
learn better with more data

increase network size $N$

- fix $k = k^*$, $f = N$
- bigger network, more data
- WSBM converges on correct solution more quickly
- thresholding + SBM particularly bad

we keep the $n_r / N$ constant
learning the number of groups

vary number of groups found $k$

- fix $f = N$
- too few / many blocks?
In fact, Bayesian marginalization will correctly choose $k=k^*$ in this case.
learning despite noise

increase variance in edge weights $\sigma_r^2$.

- fix $k = k^*$, $f = \mathcal{N}$
- bigger variance, less signal
increase variance in edge weights $\sigma_r^2$.

- fix $k = k^*$, $f = \mathcal{N}$
- bigger variance, less signal
- WSBM fails more gracefully than alternatives, even for very high variance
- thresholding + SBM particularly bad
• single-scale structural inference
  mixtures of assortative, disassortative groups

• inference is cheap (VB)
  approximate inference works well

• thresholding edge weights is bad, bad, bad
  one threshold (SBM) vs. many (WSBM)

• generalizations also for sparse graphs, degree-corrections, etc.
generative models

• 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?

• model checking
  have we learned correctly? check via generating synthetic networks

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

• anomaly detection
  low probability events under generative model
Generative models

- 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?

- Model checking
  - Have we learned correctly? Check via generating synthetic networks

- Partial or noisy data
  - Extrapolation, interpolation, hidden data, missing data

- Anomaly detection
  - Low probability events under generative model
Thank you to:

Cris Moore (Santa Fe)
Mark Newman (Michigan)
Cosma Shalizi (Carnegie Mellon)

Funding from:

Aicher, Jacobs, Clauset, “Adapting the Stochastic Block Model to Edge-Weighted Networks.” ICML (2013)
Moore, Yan, Zhu, Rouquier, Lane, “Active learning for node classification in assortative and disassortative networks.” KDD (2011)
fin