Non-parametric Bayesian Models for Complex Networks

Morten Mørup
Associate Professor
Section for Cognitive Systems (CogSys)
DTU Informatics
Technical University of Denmark

Joint work with

Lars Kai Hansen  Mikkel N. Schmidt  Tue Herlau

DTU Informatik
Institut for Informatik og Matematisk Modellering
The father of Graph Theory is considered to be Leonhard Euler

The bridges of Königsberg

**Problem:** Find a walk through the city that cross each bridge once and only once.

Euler proved that the problem has **no solution** using a graph representation and proved further that a solution to the problem would require all vertices to have even degree

(degree of vertex: number of edges (i.e. bridges) touching the vertex (land mass)).
Graphs and their adjacency matrices

A graph $G(V,E)$ with vertices $V$ and edges $E$ can be represented by the corresponding adjacency matrix $A$ such that $A_{ij}=1$ if there is a link from vertex $i$ to vertex $j$ and $A_{ij}=0$ otherwise.

**Undirected Graph**

$$V = \{1,2,3,4,5,6\}$$
$$E_{Undirected} = \{(1,2),(1,5),(2,5),(2,3),(3,4),(4,5),(4,6)\}$$

**Directed Graph**

$$V = \{1,2,3,4,5,6\}$$
$$E_{Directed} = \{(1,5),(2,1),(2,5),(3,2),(3,4),(4,5),(5,2),(6,4)\}$$

$$A_{Undirected} = \begin{bmatrix}
0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
\end{bmatrix}, \quad A_{Directed} = \begin{bmatrix}
0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
\end{bmatrix}. $$
Bipartite Graphs

\[ V^{(1)} = \{1, 2, 3\} \]
\[ V^{(2)} = \{1, 2, 3, 4\} \]
\[ E_{Bipartite} = \{(1, 2), (2, 3), (2, 4), (3, 1), (3, 4)\} \]

\[ A_{Bipartite} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}. \]
Large-scale networks with complex patterns of connections between elements abound naturally in both nature and in man-made systems.

**Biology**
- Protein Interaction
- Food web
- Epidemic Spread
- Connectome

**Economics**
- Business Organization
- Trade network

**Social sciences:**
- Friendship network
- Collaboration network
- Sexual relations

**Technology**
- Tele communication network
- Airline connections
- The internet

Complex networks pervades both science and popular culture

Basic premise: Local connectivity leads to collective behavior and for such collective behavior to be detectable (i.e., by statistical machine learning methods) one needs to analyze large datasets.
Our research focus: Bayesian non-parametric modeling of relational data

Models we consider can:
1. Adapt to the complexity of the relational data (→ Bayesian non-parametrics)
2. Device generative model for graphs (→ A statistical process for creating networks)
3. Be used to predict missing entries (→ Evaluate significance of extracted patterns)
4. Form easy interpretable representations of structure (→ Facilitate human understanding)
5. Are scalable (→ models we research grow in the complexity of the edges of the graph)

Modeling challenges we consider

<table>
<thead>
<tr>
<th>Overlapping groups</th>
<th>Hierarchical structure</th>
<th>Community structure</th>
<th>Varying node degree</th>
<th>Varying connection strength</th>
<th>Connection Directionality</th>
<th>Multi-graphs</th>
</tr>
</thead>
</table>
Bayesian learning and the principle of parsimony

The explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory.

To get the posterior probability distribution, multiply the prior probability distribution by the likelihood function and then normalize.

Bayesian learning embodies Occam’s razor, i.e. Complex models are penalized.
Bayesian Statistical Learning

Infer parameters from the joint likelihood induced by the generative process

Device generative process for the data, i.e. a recipe for how the observed world could be generated.

Evaluate model performance by how well the model is able to predict unobserved quantities or account for observed structures using frequentist hypothesis testing with the generative model as null-distribution.
The Infinite Relational Model/Stochastic Blockmodel
(A prominent Bayesian generative model for graphs)

Learning Systems of Concepts with an Infinite Relational Model (AAAI2006)

Charles Kemp
Josh Tenenbaum
Thomas Griffith
Takeshi Yamada
Naonori Ueda

Infinite Hidden Relational Model (UAI 2006)

Zhao Xu
Kai Yu
Volker Tresp
Hans-Peter Kriegel
The Stochastic Blockmodel and the Infinite Relational Model

Graph with elements $A_{ij}$ given by the relation $j$ "defers to" $i$


Bayesian Statistical Learning

Infer parameters from the joint likelihood induced by the generative process.

Device generative process for the data, i.e. a recipe for how the observed world could be generated.

Evaluate model performance by how well the model is able to predict unobserved quantities or account for observed structures using frequentist hypothesis testing with the generative model as null-distribution.
Non-parametric modeling by the Chinese Restaurant Process (CRP)

Imagine a (Chinese) restaurant with countable infinitely many tables, labeled 1, 2,...... Customers (i.e. vertices) walk in and sit down at some table (cluster). The tables are chosen according to the following random process.

1. **The first customer always chooses the first table.**
2. **The i\textsuperscript{th} customer chooses:**
   a) first unoccupied table with probability: \( \frac{\alpha}{i-1+\alpha} \)
   b) an occupied table with probability: \( \frac{m_k}{i-1+\alpha} \)

where \( m_k \) is the number of people already sitting at that table.

\[
p(z_1, \ldots, z_{10}) = p(z_1)p(z_2|z_1)p(z_3|z_1, z_2) \cdots p(z_{10}|z_1, \ldots, z_9) \\
= \frac{\alpha}{\alpha+1} \frac{\alpha}{\alpha+2} \frac{\alpha}{\alpha+3} \frac{\alpha}{\alpha+4} \frac{\alpha}{\alpha+5} \frac{\alpha}{\alpha+6} \frac{\alpha}{\alpha+7} \frac{\alpha}{\alpha+8} \frac{\alpha}{\alpha+9} \frac{\alpha}{\alpha+2}
\]

\[
p(Z|\alpha) = \frac{\alpha^K (N_k - 1)! (\alpha - 1)!}{(J + \alpha - 1)!} = \frac{\alpha^K \Gamma(N_k) \Gamma(\alpha)}{\Gamma(J + \alpha)}
\]

See also tutorial: S.J. Gershman, D.M. Blei / Journal of Mathematical Psychology 56 (2012) 1–12
Generative model of the IRM
(A statistical recipie for simulating networks)

- Draw clustering assignement from CRP
  \[ z : z \sim CRP(\gamma) \]

- Draw the relations between the generated clusters
  \[ \eta : \eta_{lm} \sim Beta(\beta^+, \beta^-) \]

- Draw the graph
  \[ A : A_{ij} \sim Bernoulli(\eta_{zi,zj}) \]

Main Assumption:
Relationships are conditionally independent given cluster assignments.
Different kinds of relational systems

IRM extends well to many types of graphs

UnDirected

\[ \pi_{ij} = \eta_{zi} z_j \]

\[ \eta = \eta^\top \]

Directed

\[ \pi_{ij} = \eta_{zi} z_j \]

Bipartite

\[ \pi_{ij} = \eta_{z_i^{(1)} z_j^{(2)}} \]

Multi-graph

\[ \pi_{ij}^{(r)} = \eta_{z_i^{(r)} z_j^{(r)}} \]

\[ \pi_{ijk} = \eta_{z_i^{(1)} z_j^{(1)} z_k^{(2)}} \]
Bayesian Statistical Learning

Infer parameters from the joint likelihood induced by the generative process

Evaluate model performance by how well the model is able to predict unobserved quantities or account for observed structures using frequentist hypothesis testing with the generative model as null-distribution.

Device generative process for the the data, i.e. a recipe for how the observed world could be generated.
From the generative model

\[ z \sim CRP(\alpha) \]

\[ \eta_{lm} \sim Beta(\beta^+, \beta^-) \]

\[ A_{ij} \sim Bernoulli(\eta_{zi, zj}) \]

We can write down the joint likelihood

\[
P(A, z, \eta|\alpha, \beta^+, \beta^-) = \prod_{i \neq j} \text{Bernoulli}(A_{ij}|\eta_{zi, zj}) \prod_{lm} \text{Beta}(\eta_{lm}|\beta^+_{lm}, \beta^-_{lm}) \text{CRP}(z|\alpha)
\]

\[
= \prod_{lm} \eta_{lm}^{N^+_{lm}} (1 - \eta_{lm})^{N^-_{lm}} \prod_{lm} \frac{\Gamma(\beta^+_{lm} + \beta^-_{lm})}{\Gamma(\beta^+_{lm}) \Gamma(\beta^-_{lm})} \eta_{lm}^{\beta^+_{lm} - 1} (1 - \eta_{lm})^{\beta^-_{lm} - 1} \frac{\alpha^K \Gamma(\alpha) \prod_k \Gamma(n_k)}{\Gamma(J + \alpha)}
\]

and we observe that \( \eta \) can be analytically integrated out (i.e., collapsed):

\[
P(A, z|\alpha, \beta^+, \beta^-) = \int P(A, z, \eta|\alpha, \beta^+, \beta^-) d\eta
\]

\[
= \prod_{lm} \frac{\Gamma(\beta^+_{lm} + \beta^-_{lm})}{\Gamma(\beta^+_{lm}) \Gamma(\beta^-_{lm})} \int \eta_{lm}^{N^+_{lm} + \beta^+_{lm} - 1} \eta_{lm}^{N^-_{lm} + \beta^-_{lm} - 1} d\eta_{lm} \frac{\alpha^K \Gamma(\alpha) \prod_k \Gamma(n_k)}{\Gamma(J + \alpha)}
\]

\[
\text{Beta}(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a + b)}
\]
From the joint distribution

\[
P(A, z | \alpha, \beta^+, \beta^-) = \prod_{lm} \frac{\text{Beta}(N_{lm}^+ + \beta_{lm}^+, N_{lm}^- + \beta_{lm}^-)}{\text{Beta}(\beta_{lm}^+, \beta_{lm}^-)} \frac{\alpha^K \Gamma(\alpha) \prod_k \Gamma(n_k)}{\Gamma(J + \alpha)}
\]

We can use Bayes theorem to obtain the posterior distribution of \(z_i\)

\[
P(z_i = l | A, z \setminus i, \alpha, \beta^+, \beta^-) = \frac{P(A, z \setminus i, z_i = l | \alpha, \beta^+, \beta^-)}{\sum_l P(A, z \setminus i, z_i = l | \alpha, \beta^+, \beta^-)}
\]

And infer \(Z\) by Gibbs sampling each vertex at a time from the above posterior distribution.
By sampling we can infer $z$ that maximizes $P(z|A, \theta)$
Bayesian Statistical Learning

Inference

Modeling

Evaluation

Infer parameters from the joint likelihood induced by the generative process.

Evaluate model performance by how well the model is able to predict unobserved quantities or account for observed structures using frequentist hypothesis testing with the generative model as null-distribution.

Device generative process for the data, i.e. a recipe for how the observed world could be generated.
How do we evaluate the quality of the inferred parameters?

- Compare with ground truth if available

\[ Z^{(true)} \text{ Vs. } Z^{(estimated)} \]

- Evaluate how well the model predicts edges/links or account for structure in the data using generative process as null-model.
Ground truth evaluation: Use metric that is permutation invariant such as Mutual Information

\[ I(X, Y) = \sum_{\mathbf{x} \in X, \mathbf{y} \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right) = KL(p(x, y) || p(x)p(y)) \]
When ground truth is not available

- **Evaluate models predictive ability**
  Two commonly used measures based on prediction:
  1) Generalization error, i.e. Evaluate predictive distribution estimated on hold out data (cross-validation)
  2) Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) on hold out data (cross-validation)

- **Frequentist testing of structure in network**
  Use generative model to draw graphs and evaluate the properties of these graphs to the actual graph using the generative model as null-distribution.
Bayesian Statistical Learning

Infer parameters from the joint likelihood induced by the generative process.

Evaluate model performance by how well the model is able to predict unobserved quantities or account for observed structures using frequentist hypothesis testing with the generative model as null-distribution.

Device generative process for the data, i.e. a recipe for how the observed world could be generated.
Relational modeling of animal features

- 50 Animals
- 85 Features

\[ \pi_{ij} = \eta_{z_i^{(1)}} z_j^{(2)} \]

Figure 3: Animal clusters, feature clusters, and a sorted matrix showing the relationships between them. The matrix includes seven of the twelve animal clusters and all of the feature clusters. Some features refer to habitat (jungle, tree, coastal), and others are anatomical (bulbous body shape, has teeth for straining food from the water) or behavioral (swims, slow).

Relational modeling of Countries their characteristics and relations

- 14 Countries
- 54 Predicates
- 90 Features

Figure 6: (a) Country clusters and feature clusters for the political data. Every second feature is labelled, and gray entries indicate missing data. (b) – (i) Representatives of eight predicate clusters found by the IRM. The countries in each matrix are Kemp, C. Tenenbaum, J.B., Griffiths, T.L., Yamada, T and Ueda, N “Learning systems of concepts with an infinite relational model, AAAI 2006.
Example of our use of the IRM on fMRI data

Pairwise Mutual Information (MI) between 2x2x2 voxel groups

\[ I(i,j) = \sum_{uv} P_{ij}(u,v) \log \frac{P_{ij}(u,v)}{P_i(u)P_j(v)} \]

Top 100,000 MI links

5039 Voxel groups

A(72)

IRM

5039 Voxel groups

Components

72 Subjects

Functional units defined by coherent Groups of Volumes (Z)

Communication between the functional units (\(\rho^{(n)}\))

\[ \pi^{(n)}_{ij} = Z_i^{(1)} \eta^{(n)} z_j^{(1)} \]

Infinite Relational Modeling of Functional Connectivity in Resting State fMRI
Our Generative Model of Hierarchical Structure:

\[ z \sim CRP(\alpha) \]
\[ T \sim UT(z) \]
\[ \eta_{kl}^{(n)} \sim Beta(\beta, \beta) \]
\[ A_{ij} \sim Bernoulli(\eta_{c(i,j),c(j,i)}) \]

generate a partitioning of the nodes by the CRP
draw a tree with K leaves from a uniform distribution over all multifurcating trees
draw probability of link between group k and l at each internal node of the tree
draw links by flipping a biased coin
Synthetic data

Zachary’s Karate Club

Our model

Existing binary tree model

(Herlau, Mørup, Schmidt, Hansen, Cognitive Information Processing, 2012)
“The organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters.” (Fortunato, 2010)

Our Generative Model of Community Structure

\[
z \sim \text{CRP}(\alpha), \\
\eta_{\ell \ell} \sim \text{Beta}(\beta, \beta), \\
\gamma_{\ell} \sim \text{Beta}(\vartheta, \vartheta), \\
\eta_{m} \sim \text{Beta}(\beta, x_{lm}), \\
\eta_{m} \sim \text{Beta}(\beta, x_{lm})
\]

Cluster assignment, Within-cluster link probability, Cluster gap, Between-cluster link probability, Link,

\[
A_{ij} \sim \text{Bernoulli}(\eta_{z_{i}z_{j}}),
\]

IRM

BCD

\[
\gamma = 1 \quad \gamma = 0.25 \quad \gamma = 0
\]

(Mørup&Schmidt, in press Neural Computation 2012)
Our applications so far...

- User recommendation on social/internet services such as Facebook and Netflix
- Topic modeling of large databases, i.e. PubMed and NYTimes.
- Music preferential behaviour in large groups of people attending the Roskilde Festival
- Modeling functional and structural connectivity in the brain
- Link prediction in social networks