A Random Graph Model embedding Vertices Information

Stevenn Volant, Hugo Zanghi and Christophe Ambroise

Exalead,
Laboratoire Statistique et Genome,
La genopole - Universite d’Evry

Journées MAS – 2008, August 28th
Introduction

The Web as a Graph
At Exalead

➤ Modelisation
  ➤ Nodes are pages or websites
  ➤ Edges are hyperlinks

➤ Use cases
  ➤ Ranking.
  ➤ Spam detection.
  ➤ website bounds.

➤ Data size
  ➤ billions of pages.
  ➤ $\sim 10$ hyperlinks by pages.
  ➤ $\sim 100$ millions of websites.
Introduction

Real networks
Embedding Vertices Information

Structure Analysis with Vertices Information

- Clustering,
- Function relationship.

Families of networks

- world wide web.
- social networks,
- biological networks,

⇝ Let us define a statistical model

Insurance sites into car insurance (blue), life insurance (orange), financial insurance (green), and health insurance (purple).

Volant, Zanghi, Ambroise, Matias
Introduction

Real networks
Embedding Vertices Information

Structure Analysis with Vertices Information

- Clustering,
- Function relationship.

Families of networks

- world wide web.
- social networks,
- biological networks,

⇝ Let us define a statistical model

Volant, Zanghi, Ambroise, Matias
Notations and basic model

▶ Notations:
- \( E \) a set of edges \( \in \{1, \ldots, n\}^2 \).
- \( X = (X_{ij}) \) the adjacency matrix such that \( \{X_{ij} = 1\} = \mathbb{I}\{i \leftrightarrow j\} \).

▶ Possible graphs:
- oriented: \( X_{ij} \neq X_{ji} \), valued: \( X_{ij} \in \mathbb{R} \).

▶ Random graph definition:
- the distribution of \( X \) describes the topology of the network.

▶ Erdös Rényi (ER) model (1959):
- \( (X_{ij}) \) independent, with Bernoulli distribution \( \mathcal{B}(p) \).
MixNet: an alternative probabilistic model

- **Origin**
  - model developed by J. Daudin et al. (2008),
  - ER model generalization,
  - application fields: biology, internet, social network . . .

- **Modelling connection heterogeneity**
  - hyp.: there exists a hidden structure into $Q$ classes of connectivity,
  - $\mathbf{Z} = (\mathbf{Z}_i)_i, Z_{iq} = \mathbb{I}\{i \in q\}$ are indep. hidden variables,
  - $\alpha = \{\alpha_q\}$, the prior proportions of groups,
  - $(\mathbf{Z}_i) \sim \mathcal{M}(1, \alpha)$.

- **Distribution of $X$**
  - Conditional distribution: $X_{ij}|\{Z_{iq}Z_{j\ell} = 1\} \sim \mathcal{B}(\pi_{q\ell})$
  - $X_{ij}|Z$ are independant.
  - $\pi = (\pi_{q\ell})$ is the connectivity matrix.
  - Erdős-Rényi Mixture for Network.
CohsMix: Covariables in Hidden Structures using Mixture models

Embedding informations

- Extending data relationships with exogenous variables: $n$ nodes with $p$ covariables. Example:
  - Web: text, geolocalisation, language, ranking, etc.
  - Social: age, geolocalisation, genre, etc.
  - Biological: gene expression data.

Handling Informations: two cases

- Adding a covariate vector $\mathbf{Y}_i$ ($p$ dimensions) for each node.
- Building a similarity matrix $\mathbf{Y}$ ($n$ by $n$).
Different dependencies

1. \( P(X, Y, Z) = P(Z)P(X, Y|Z) = P(Z)P(Y|Z)P(X|Z) \)
2. \( P(X, Y, Z) = P(Z)P(X|Z)P(Y|X, Z) \)
3. \( P(X, Y, Z) = P(Z)P(Y|Z)P(X|Y, Z) \Rightarrow \text{not considered in this talk.} \)

Common distributions

- \( Z_i \sim \mathcal{M}(1, \alpha) \) with \( \alpha = \{\alpha_q\} \), the prior proportions of groups.
- Conditional distribution: \( X_{ij}|\{Z_iqZ_{j\ell} = 1\} \sim \mathcal{B}(\pi_{q\ell}) \)
Model 1: Independence of $X$ and $Y / Z$

Considering independence between edges and covariates:

- Normal Distribution: $Y_{ij} | Z_i Z_j = 1 \sim \mathcal{N}(\mu_{ql}, \sigma^2)$

Affiliation model:

$$\begin{cases} 
\mu_{qq} = \mu_1, & \forall q \in [1, Q] \\
\mu_{ql} = \mu_2, & \forall q, l \in [1, Q]^2, q \neq l
\end{cases}$$

Remarks

- Strong assumption: independence between covariates and edges.
- Not a realistic model.
Model 2: dependence between $X$ and $Y$

Considering dependence between edges and covariates:

\[ Y_{ij} | Z_{iq} Z_{jl} = 1 \sim \mathcal{N}(\mu_{ql}, \sigma^2), \quad \text{si } X_{ij} = 1, \]
\[ Y_{ij} | Z_{iq} Z_{jl} = 1 \sim \mathcal{N}(\tilde{\mu}_{ql}, \sigma^2), \quad \text{si } X_{ij} = 0. \]

Affiliation model:

\[ \begin{cases} 
\mu_{qq} = \mu_1, & \forall q \in [1, Q], \\
\mu_{ql} = \mu_2, & \forall q, l \in [1, Q]^2 q \neq l.
\end{cases} \]

and

\[ \begin{cases} 
\tilde{\mu}_{qq} = \tilde{\mu}_1, & \forall q \in [1, Q], \\
\tilde{\mu}_{ql} = \tilde{\mu}_2, & \forall q, l \in [1, Q]^2, q \neq l.
\end{cases} \]
Model 2: Log-Likelihood related to $Y/X, Z$

\[ P(Y|X, Z) = \prod_{i,j} \prod_{q,l} P(Y_{ij}|X_{ij} = x_{ij})^{x_{ij} z_{iq} z_{jl}} P(Y_{ij}|X_{ij} = x_{ij})^{(1-x_{ij}) z_{iq} z_{jl}}. \]

details . . .

\[
\log(P(Y|X, Z)) = \sum_{i,j} \sum_{q,l} z_{iq} z_{jl} x_{ij} \log(P(Y_{ij}|X_{ij} = x_{ij})) \\
+ \sum_{i,j} \sum_{q,l} z_{iq} z_{jl} (1-x_{ij}) \log(P(Y_{ij}|X_{ij} = x_{ij})) \\
= \sum_{i,j} \sum_{q,l} z_{iq} z_{jl} x_{ij} \left( -\frac{(y_{ij} - \mu_{ql}^{(1)})^2}{2\sigma^2} + \frac{(y_{ij} - \mu_{ql}^{(2)})^2}{2\sigma^2} \right) \\
- \sum_{i,j} \sum_{q,l} z_{iq} z_{jl} \frac{(y_{ij} - \mu_{ql}^{(2)})^2}{2\sigma^2} + \sum_{i,j} \sum_{q,l} z_{iq} z_{jl} \log\left( \frac{1}{\sqrt{2\pi\sigma^2}} \right).
\]
Covariates in Hidden Structures using Mixture models

Generative Models

\[ n = 150, \ q = 4, \ \alpha = (0.2, 0.3, 0.1, 0.4), \ \pi_{ql} = 0.2, \ \pi_{qq} = 0.05, \ \mu_{qq} = 2, \ \mu_{ql} = 4, \ \mu_{ql}^{-} = 5, \ \mu_{qq}^{-} = 7 \text{ and } \sigma^2 = 1 \]

model 1: \[ P(X, Y, Z) = P(Z)P(Y|Z)P(X|Z) \]

model 2: \[ P(X, Y, Z) = P(Z)P(X|Z)P(Y/X, Z) \]
How to estimate the model parameters?

- Log-likelihood(s) of the model:
  → Observed data: \( \mathcal{L}(X, Y) = \log \left( \sum_{Z} \exp \mathcal{L}(X, Y, Z) \right) \).
  → Complete data: \( \mathcal{L}(X, Y, Z) \).
  ⇒ \( Q^n \) partitions: not tractable.

- EM-based Strategies:
  → Expectation of Complete data:
  \( Q(\theta) = \mathbb{E} [\mathcal{L}(X, Y, Z) | X, Y] \).
  → EM-like strategies require the knowledge of \( \Pr(Z|X, Y) \).
  ⇒ In our case, this distribution is not tractable (no conditional independence).

Objective

Restrict the set of distributions of \( Z \): Variationnal Approach.

Approximation: \( R(Z) = \prod_{i=1}^{n} P(Z_i | X, Y, \theta^{(m)}) \).
**Variational method**
Daudin et. al, 2008

**Principle**

→ Optimizing $J(\theta)$ defined by:

$$J(\theta) = \mathcal{L}(X, Y) - KL(R(Z), \Pr(Z|X, Y, \theta^{(m)})).$$

→ $R(Z)$ chosen such that $KL(R(Z), \Pr(Z|X, Y))$ is minimal.

→ If $R(Z) = \Pr(Z|X, Y)$ then $J(R(Z)) = \mathcal{L}(X, Y)$.

**After Simplification**:

$$E_{R(Z)}(J(\theta)) = E_{R(Z)}(\log(P_\theta(X, Y, Z))|X, Y, \theta) - \sum_{Z} R(Z) \log(R(Z)).$$

⇒ tractable
Covariates in Hidden Structures using Mixture models

Estimation : Variational approach of the EM algorithm

Quantity to maximize

\[
E_{R(Z)}(J(\theta)) = \sum_{i=1}^{n} \sum_{q=1}^{Q} R(z_{iq}) \log(\alpha_q) + \sum_{i,j} R(z_{iq}) R(z_{jl}) (x_{ij} \log(\pi_{ql}) + (1 - x_{ij}) \log(1 - \pi_{ql}))
\]

\[
+ \sum_{i,j} \sum_{q,l} R(z_{iq}) R(z_{jl}) \left[ x_{ij} \left( -\frac{(y_{ij} - \mu_{ql}^{(1)})^2}{2\sigma^2} + \frac{(y_{ij} - \mu_{ql}^{(2)})^2}{2\sigma^2} \right) - \frac{(y_{ij} - \mu_{ql}^{(2)})^2}{2\sigma^2} \right]
\]

\[
+ \sum_{i,j} \sum_{q,l} R(z_{iq}) R(z_{jl}) \log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \sum_{Z} R(Z) \log(R(Z)).
\]
Covariates in Hidden Structures using Mixture models

Estimation: Variational approach of the EM algorithm

Optimal parameters

\[ \hat{\pi}_{ql} = \frac{\sum_{i,j} R(Z_{iq}) R(Z_{jl}) x_{ij}}{\sum_{i,j} R(Z_{iq}) R(Z_{jl})} , \]

\[ \hat{\alpha}_q = \frac{\sum_{i=1}^{n} R(Z_{iq})}{n} \text{, } \hat{\mu}_{ql} = \frac{\sum_{i,j} R(Z_{iq}) R(Z_{jl}) x_{ij} y_{ij}}{\sum_{i,j} R(Z_{iq}) R(Z_{jl}) x_{ij}} \quad \hat{\bar{\mu}}_{ql} = \frac{\sum_{i,j} R(Z_{iq}) R(Z_{jl}) (1 - x_{ij}) y_{ij}}{\sum_{i,j} R(Z_{iq}) R(Z_{jl}) (1 - x_{ij})} , \]

\[ \hat{\sigma}^2 = \frac{\sum_{i,j} R(Z_{iq}) R(Z_{jl}) \left[ x_{ij} \left( (y_{ij} - \hat{\mu}_{ql})^2 - (y_{ij} - \hat{\bar{\mu}}_{ql})^2 \right) + (y_{ij} - \hat{\bar{\mu}}_{ql})^2 \right]}{\sum_{i,j} R(Z_{iq}) R(Z_{jl})} . \]

\[ \hat{\tau}_{iq}^{(m+1)} \propto \alpha_q \prod_j \prod_l \left[ \frac{\hat{\pi}_{ql} x_{ij} (1 - \hat{\pi}_{ql})^{1 - x_{ij}}}{\sqrt{2\pi\sigma^2}} \exp\left( \frac{1}{2\sigma^2} \left[ x_{ij} \left( -(y_{ij} - \mu_{ql})^2 + (y_{ij} - \bar{\mu}_{ql})^2 \right) - (y_{ij} - \bar{\mu}_{ql})^2 \right] \right) \right]^{\tau_{jl}^{(m)}} . \]
Model selection: ICL algorithm

The Integrated Classification Likelihood

\[
ICL(Q) = \max_{\theta} \log(X, Y, Z | \theta, Q)
\]

\[
- \frac{1}{2} \times Q^2 \log\left(\frac{n(n - 1)}{2}\right)
\]

Penalization term related to \(\pi_{q_l}\)

\[
- Q^2 \log\left(\frac{n(n - 1)}{2}\right) - \log\left(\frac{n(n - 1)}{2}\right)
\]

Penalization term related to \(\mu_{q_l}, \tilde{\mu}_{q_l}\) and \(\sigma^2\)

\[
- \frac{Q - 1}{2} \log(n)
\]

Penalization term related to \(\alpha_q\)
Application : Simulated data

Synthetic network

\[
n = 150, \ Q = 4 \\
\alpha = (0.2, 0.3, 0.1, 0.4) \\
\pi_{qq} = 0.05, \ \forall q \in [1, Q] \\
\pi_{ql} = 0.2, \ \forall q, l \in [1, Q], q \neq l \\
\mu_{qq} = 2, \ \forall q \in [1, Q] \\
\mu_{ql} = 4, \ \forall q, l \in [1, Q], q \neq l \\
\tilde{\mu}_{qq} = 5, \ \forall q \in [1, Q] \\
\tilde{\mu}_{ql} = 7, \ \forall q, l \in [1, Q], q \neq l \\
\sigma^2 = 1
\]

Results matching the simulated number of groups.

Volant, Zanghi, Ambroise, Matias
Application: Simulated data

Parameters estimation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Simulated</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Model 1</td>
</tr>
<tr>
<td>$\pi_{qq}$</td>
<td>0.2</td>
<td>0.18</td>
</tr>
<tr>
<td>$\pi_{ql}$</td>
<td>0.05</td>
<td>0.044</td>
</tr>
<tr>
<td>$\mu_{qq}$</td>
<td>2</td>
<td>2.04</td>
</tr>
<tr>
<td>$\mu_{ql}$</td>
<td>4</td>
<td>4.02</td>
</tr>
<tr>
<td>$\tilde{\mu}_{qq}$</td>
<td>5</td>
<td>/</td>
</tr>
<tr>
<td>$\tilde{\mu}_{ql}$</td>
<td>7</td>
<td>/</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table: Means of the parameter estimates of the two models computed over 10 runs.
Algorithm comparison

- reference: HMRF with covariates (Besag (1986)).
- X is known:
- favorable Markov fields settings:
  - Simulate X: classical Erdős-Rényi (parameter $\lambda$)
  - Simulate $Z|X$: Gibbs sampling and Potts model
  - Simulate $Y_i^{(p)}|Z_i: Y_i^{(p)}|Z_i \sim \mathcal{N}(\zeta_i^{(p)}, \sigma^2)$
  - Build $Y_{ij}: Y_{ij} = \langle Y_i, Y_j \rangle$
Algorithm comparison

Markov favorable settings.

CohsMix favorable settings.
Application: Web Search Results Clustering using hypertextuality

Aims

- Reduce ambiguous queries ("avocat", "orange", "jaguar", etc.)
- Organize search results into groups, one for each meaning of the query.

Innovation

- Use the hypertextuality.
- Competitive communities ("abortion", "political", "jo", etc.)
Document similarities matrix

- **Vector Space Model**: Weight vector for document $i$:
  $$v_i = [w_{1,i}, w_{2,i}, ..., w_{p,i}]^T,$$
  where
  $$w_{t,i} = TF_t \cdot \log \frac{|D|}{|t \in D|}.$$  

- $TF_t$: term frequency of term $t$ in document $i$.
- $\log \frac{|D|}{|t \in D|}$: inverse document frequency with $|D|$ the total number of documents and $|t \in D|$ the number of documents containing the term $t$.

- **Document similarities**: comparing deviations of angles between each pair of document vectors:
  $$\cos \theta_{i,j} = Y_{i,j} = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}.$$
A query time process

1. Retrieve web pages matching the query;
2. Compute the text similarities;
3. Fetch websites graph (because webpages are underconnected);
4. Multithreading of CohsMix algo. to find the optimal models;
5. Shows clusters;
Application : Example

Example: query "orange", \( Q = 6, n = 269 \)

<table>
<thead>
<tr>
<th>( Q )</th>
<th>( \alpha )</th>
<th>Analysis</th>
<th>url examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.25</td>
<td>Phone</td>
<td>pressphone.com, comparatel.fr</td>
</tr>
<tr>
<td>3</td>
<td>0.11</td>
<td>Orange</td>
<td>orange.fr, <a href="http://www.orange-wifi.com">www.orange-wifi.com</a></td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>French Town</td>
<td>ville-orange.fr, immobilier.orange.fr</td>
</tr>
<tr>
<td>6</td>
<td>0.16</td>
<td>ADSL</td>
<td>dedibox-news.com, infosadsl.com</td>
</tr>
</tbody>
</table>

Junk: \( Q = 1 \cup Q = 4 \)

Proof of concept

Need tuning:
  - Document similarities.
  - WebSites graph.
Conclusion and perspectives

- **CohsMix**:
  - Uses MixNet: a probabilistic model which captures features of real-networks,
  - Embedding vertex informations to detect hidden structure.

- **References**:
  - Daudin J-J., Picard F., Robin S. (2008), A mixture model for random graphs, Statistic and Computing
  - Zanghi, H, Ambroise, C. and Miele, V. (2008), Fast online Graph Clustering via Erdös-Rényi Mixture, Pattern Recognition

- **Softwares**:
  - CohsMix, a R code of this talk (available on demand ?)

- **Perspectives**:
  - Compare with other methods (bi-clustering).
  - demo online: [http://labs.exalead.com](http://labs.exalead.com)