The Stochastic Topic Block Model for the Clustering of Vertices in Networks with Textual edges

Presenter: Wei JIANG

Authors: Bouveyron Charles, Latouche Pierre, Zreik Rawya
Machine Learning Journal Club, CMAP

September 28th 2017
Outline

1. Introduction
2. STBM Model
3. Inference
4. Experiences
   - Simulation study
   - Real-world data
Motivation

Communication between individuals via
- Social media
  - Facebook
  - Twitter
  - Linkedin
- Electronic formats
  - Email
  - Web
  - E-publication

**Figure** – Social network diagram displaying friendship ties among a set of Facebook users

Network Analysis
Network analysis for clustering users in email system

- Directed graph
- Node: users
- Edge: sender -> recipient
- Task: Clustering users based on person-to-person link only

How to improve \(\Rightarrow\) Also take into account the email **content**.
Structure of paper

- **Problem**: Discovering clusters of vertices \(\leftrightarrow\) the network interactions and the text content.
- **Model**: Stochastic topic block model (STBM) - a probabilistic model for networks with textual edges.
- **Inference**: Classification variational expectation-maximization (C-VEM)
- **Experience**: Simulated data to assess the approach and highlight its features. Real-world data sets to demonstrate the effectiveness.
Outline

1. Introduction
2. STBM Model
3. Inference
4. Experiences
   - Simulation study
   - Real-world data
Context and notations

- Directed network
- \( M \) vertices
- \( A : M \times M \) adjacency matrix.
  \[
  A_{ij} = \begin{cases} 
  1, & \text{if there is an edge from } i \text{ to } j \\
  0, & \text{otherwise}
  \end{cases}
  \]
- If \( A_{ij} = 1 \), then this edge is characterized by a set of \( D_{ij} \) documents: \( W_{ij} = (W^d_{ij})_d \)
- Each document is made by a collection of \( N^d_{ij} \) words: \( W_{ij}^d = (W_{ij}^{dn})_n \)
- \( W = (W_{ij})_{ij} \) The set of all documents exchanged for all the edges
Cluster the vertices into Q latent groups sharing the same connection profiles

- Presence of edges
- Documents between pairs of vertices

\[ \Leftrightarrow \text{estimate } Y = (Y_1, \cdots, Y_M) \text{ of latent variable } Y_i \text{ s.t.} \]

\[ Y_{iq} = \begin{cases} 
1, & \text{vertex } i \text{ belongs to cluster } q \\
0, & \text{otherwise}
\end{cases} \]
Assumptions

- Any kind of relationships between two vertices can be explained by their latent clusters only.
- Words in documents are drawn from a mixture distribution over topics, each document $d$ having its own vector of topic proportions $\theta_d$. 
Overview of STBM Model

**FIGURE** — Graphical representation of the stochastic topic block model
Modeling the presence of edges

Stochastic block model (Wang & Wong 1987; Nowicki & Snijders 2001)

- \( Y_i \sim \mathcal{M}(1, \rho = (\rho_1, \cdots, \rho_Q)) \)
- \( A_{ij} | Y_{iq} Y_{jr} = 1 \sim \mathcal{B}(\pi_{qr}) \)
  \( \Rightarrow \pi \) the \( Q \times Q \) matrix of connection probabilities

\( \Rightarrow p(A, Y | \rho, \pi) = p(A | Y, \pi)p(Y | \rho) \)
Modeling the construction of documents

Latent Dirichlet Allocation (Blei et al. 2003)

- Pair of clusters \((q, r)\) of vertices \(\rightarrow\) vector of topic proportions
  \[\theta_{qr} = (\theta_{qrk})^\top \sim \text{Dir}(\alpha = (\alpha_1, \cdots, \alpha_K))\]
  Here all components of \(\alpha\) are fixed to 1.

- The \(n\)th word of \(d\)th document between vertex \(i\) and \(j\) \(\rightarrow\) Latent topic vector
  \[Z_{ij}^{dn} | \{Y_{iq} Y_{jr} A_{ij} = 1, \theta\} \sim \mathcal{M}(1, \theta_{qr} = (\theta_{qr1}, \cdots, \theta_{qrK}))\]

- \(W_{ij}^{dn} | Z_{ij}^{dnk} = 1 \sim \mathcal{M}(1, \beta_k = (\beta_{k1}, \cdots, \beta_{kV}))\)

\(\Rightarrow\) Mixture model for words over topics

\[
W_{ij}^{dn} | \{Y_{iq} Y_{jr} A_{ij} = 1, \theta\} \sim \sum_{k=1}^{K} \theta_{qrk} \mathcal{M}(1, \beta_k)
\]
Modeling the construction of documents

Assume

- All the latent variables $Z_{ij}^{dn}$ are sampled independently.
- Given the latent variables, the words $W_{ij}^{dn}$ are independent.

Denote $Z = (Z_{ij}^{dn})_{ijdn} \Rightarrow$ Joint distribution

$$p(W, Z, \theta | A, Y, \beta) = p(W | A, Z, \beta)p(Z | A, Y, \theta)p(\theta)$$
The full joint distribution of STBM model

\[ p(A, W, Y, Z, \theta|\rho, \pi, \beta) = p(W, Z, \theta|A, Y, \beta)p(A, Y|\rho, \pi) \]
## Examples

<table>
<thead>
<tr>
<th>Scenario</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>M (nb of nodes)</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>K (topics)</td>
<td>100</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Q (groups)</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>$\rho$ (group prop.)</td>
<td>$\pi_{qq} = 0.25$, $\pi_{qr, r\neq q} = 0.01$</td>
<td>$\pi_{qr, \forall q, r} = 0.25$</td>
<td>$\pi_{qq} = 0.25$, $\pi_{qr, r\neq q} = 0.01$</td>
</tr>
<tr>
<td>$\pi$ (connection prob.)</td>
<td>$\theta_{111} = \theta_{222} = 1$, $\theta_{233} = 1$, $\theta_{qr4, r\neq q} = 1$, otherwise 0</td>
<td>$\theta_{111} = \theta_{222} = 1$, $\theta_{qr3, r\neq q} = 1$, otherwise 0</td>
<td>$\theta_{111} = \theta_{333} = 1$, $\theta_{222} = \theta_{142} = 1$, $\theta_{qr3, r\neq q} = 1$, otherwise 0</td>
</tr>
</tbody>
</table>

### Scenario A

### Scenario B

### Scenario C
The simulated messages (150 words) are from four texts from BBC news:

1. The birth of Princess Charlotte
2. Black holes in astrophysics
3. UK politics
4. Cancer diseases in medicine
Key Property of STBM Model

Assume that $Y$ is available.

- Recognize documents in $W$ s.t $W = (\tilde{W}_{qr})_{qr}$
  $\Rightarrow$ All words in $\tilde{W}_{qr}$ share the same mixture distribution over topic.
  $\Rightarrow$ Words in $W$ are drawn from LDA model with $D = Q^2$ independent documents $\tilde{W}_{qr}$.

- $p(A, Y | \rho, \pi)$ involves sampling of the clusters + construction of binary variables describing presence of edges
  $\Rightarrow$ correspond to likelihood of SBM model.

For given $Y$, the full joint distribution factorizes into LDA like term and SBM like term.
Outline

1. Introduction
2. STBM Model
3. Inference
4. Experiences
   - Simulation study
   - Real-world data
Aim: maximize log-likelihood

First fix the number of groups $Q$ and number of topics $K$

$$\log p(A, W, Y | \rho, \pi, \beta) = \log \sum_{Z} \int_{\theta} p(A, W, Y, Z, \theta | \rho, \pi, \beta) d\theta$$

- Model parameters $(\rho, \pi, \beta)$
- $Z$ and $\theta$ are latent variables.
- $Y = (Y_1, \cdots, Y_M)$ is seen as a set of binary vectors for which we aim at providing estimates. (Motivated by the key property of STBM)
Variational decomposition of log-likelihood

\[
\log p(A, W, Y | \rho, \pi, \beta) = \mathcal{L}(R(\cdot); Y, \rho, \pi, \beta) + \text{KL}(R(\cdot)||p(\cdot|A, W, Y, \rho, \pi, \beta))
\]

KL : the Kullback-Leibler divergence between the true and approximate posterior distribution \(R(\cdot)\) of \((Z, \theta)\), given the data and model parameters.

\[
\text{KL}(R(\cdot)||p(\cdot|A, W, Y, \rho, \pi, \beta)) = - \sum_Z \int_{\theta} R(Z, \theta) \log \frac{p(Z, \theta|A, W, Y, \rho, \pi, \beta)}{R(Z, \theta)} d\theta
\]

\(\Rightarrow\) Maximizing the lower bound \(\mathcal{L}\) w.r.t \(R(Z, \theta)\) induces a minimization of KL divergence.
Recall STBM property: The set of latent variables in \( Y \) allows the full joint distribution be decomposed to the sampling of \( Y \) and \( A \) + construction of documents given \( A \) and \( Y \).

\[
\mathcal{L}(R(\cdot); Y, \rho, \pi, \beta) = \tilde{\mathcal{L}}(R(\cdot); Y, \beta) + \log p(A, Y|\rho, \pi)
\]

where

\[
\tilde{\mathcal{L}}(R(\cdot); Y, \beta) = \sum_Z \int_\theta R(Z, \theta) \log \frac{p(W, Z, \theta|A, Y, \beta)}{R(Z, \theta)} d\theta
\]

\( \Rightarrow \) For given \( Y \), the two terms can be maximized independently.
C-VEM algorithm

Aim: Maximize the lower bound $\mathcal{L}$.

C-VEM algorithm alternates between the optimization of $R(Z, \theta)$, $Y$ and $(\rho, \pi, \beta)$.

1. Estimate of $R(Z, \theta)$
   - Update $R(Z_{ij}^{dn})$ and $R(\theta)$ of the E-step of VEM

2. Estimate of model parameters $(\rho, \pi, \beta)$
   - Maximize the lower bound $\mathcal{L} \Rightarrow \beta$ only in $\tilde{\mathcal{L}}$; $\rho, \pi$ only in SBM log-likelihood. (M-step)

3. Estimate of $Y$
   - Fix $(\rho, \pi, \beta)$ and $R(Z, \theta) \Rightarrow$ Find $Y$ maximizing $\mathcal{L}$
   - Test $Q^M$ possible cluster assignments $\Rightarrow$ on line clustering methods

(Classification)
(Biernacki et al. 2003) For several initializations of a k-means like algorithm on a distance matrix between vertices

1. VEM for LDA is applied on all documents $i \rightarrow j \Rightarrow X_{ij} = k$ if $k$ is the majority topic.

2. Distance matrix

$$
\Delta(i, j) = \sum_{h=1}^{M} \delta(X_{ih} \neq X_{jh})A_{ih}A_{jh} + \sum_{h=1}^{M} \delta(X_{hi} \neq X_{hj})A_{hi}A_{hj}
$$

Look at all possible edges $i \rightarrow j$ towards a third vertex $h \Rightarrow$ compare the edge type

The distance matrix computes the number of discordances in the way both $i$ and $j$ connect to other vertices or vertices connect them.
Model selection

Model selection problem: Estimating number of groups $Q$ and number of topics $K$
Criterion: ICL (Biernacki et al. 2000)
Outline

1. Introduction
2. STBM Model
3. Inference
4. Experiences
   - Simulation study
   - Real-world data
Simulation setup

<table>
<thead>
<tr>
<th>Scenario</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>M (nb of nodes)</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K (topics)</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Q (groups)</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>ρ (group prop.)</td>
<td>(1/Q, ..., 1/Q)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>π (connection prob.)</td>
<td>( \pi_{qq} = 0.25 ) ( \pi_{qr, r \neq q} = 0.01 )</td>
<td>( \pi_{qr, \forall q, r} = 0.25 )</td>
<td>( \pi_{qq} = 0.25 ) ( \pi_{qr, r \neq q} = 0.01 )</td>
</tr>
<tr>
<td>θ (prop. of topics)</td>
<td>( \theta_{111} = \theta_{222} = 1 ) ( \theta_{333} = 1 ) ( \theta_{qr4, r \neq q} = 1 ) ( \text{otherwise} \ 0 )</td>
<td>( \theta_{111} = \theta_{222} = 1 ) ( \theta_{qr3, r \neq q} = 1 ) ( \text{otherwise} \ 0 )</td>
<td>( \theta_{111} = \theta_{331} = 1 ) ( \theta_{222} = \theta_{142} = 1 ) ( \theta_{qr3, r \neq q} = 1 ) ( \text{otherwise} \ 0 )</td>
</tr>
</tbody>
</table>

Scenario A

Scenario B

Scenario C
Simulation setup

The simulated messages (150 words) are from four texts from BBC news:

1. The birth of Princess Charlotte
2. Black holes in astrophysics
3. UK politics
4. Cancer diseases in medicine
Introductory example on scenario C

Run C-VEM for STBM on network of scenario C with the actual number of groups and topics ⇒ Both network structure and the topic information should be correctly recovered.

**Figure** – Clustering result for the introductory example (scenario C)
Simulation study

Introductory example on scenario C

- Evolution of the lower bound $\mathcal{L}$ along iterations (top-left)
- The most frequent words in the 3 found topics (left-bottom)
- The estimated model parameters ($\rho, \pi$) (right)
Introductory example on scenario C

**Figure** – Summary of connexion probabilities between groups ($\pi$, edge widths), group proportions ($\rho$, node sizes) and most probable topics for group interactions (edge colors).
## Experiment on model selection

<table>
<thead>
<tr>
<th>Scenario A ($Q = 3, K = 4$)</th>
<th>Scenario B ($Q = 2, K = 3$)</th>
<th>Scenario C ($Q = 4, K = 3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K \backslash Q$</td>
<td>$K \backslash Q$</td>
<td>$K \backslash Q$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0 0 0 0 0 0 0 0</td>
<td>0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>2</td>
<td>12 0 0 0 0 0 0 0</td>
<td>12 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>0 0 0 0 0 0 0 0</td>
<td>0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>4</td>
<td>0 0 82 2 0 0 2 0</td>
<td>0 0 88 0 0 0 0 0</td>
</tr>
<tr>
<td>5</td>
<td>0 0 2 0 0 0 0 0</td>
<td>0 0 2 0 0 0 0 0</td>
</tr>
<tr>
<td>6</td>
<td>0 0 0 0 0 0 0 0</td>
<td>0 0 0 0 0 0 0 0</td>
</tr>
</tbody>
</table>

- Percentage of selections by ICL for each STBM model ($Q, K$) on 50 simulated networks of each of three scenarios.
- Highlighted rows and columns correspond to the actual values for $Q$ and $K$. 
Run SBM, LDA and STBM on 20 networks simulated according to the three scenarios. Average ARI values (Rand, 1971) are reported with standard deviations for both node and edge clustering.

- **Easy**: same as the previous simulations of three scenarios.

- **Hard 1**: the communities are very few differentiated ($p_{iqq} = 0.25$ and $\pi_{q \neq r} = 0.2$).

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario A</th>
<th>Scenario B</th>
<th>Scenario C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>node ARI</td>
<td>edge ARI</td>
<td>node ARI</td>
</tr>
<tr>
<td>SBM</td>
<td>1.00±0.00</td>
<td>-</td>
<td>0.01±0.01</td>
</tr>
<tr>
<td>LDA</td>
<td>-</td>
<td>0.97±0.06</td>
<td>-</td>
</tr>
<tr>
<td>STBM</td>
<td>0.98±0.04</td>
<td>0.98±0.04</td>
<td>1.00±0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Scenario A</th>
<th>Scenario B</th>
<th>Scenario C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>node ARI</td>
<td>edge ARI</td>
<td>node ARI</td>
</tr>
<tr>
<td>SBM</td>
<td>0.01±0.01</td>
<td>-</td>
<td>0.01±0.01</td>
</tr>
<tr>
<td>LDA</td>
<td>-</td>
<td>0.90±0.17</td>
<td>-</td>
</tr>
<tr>
<td>STBM</td>
<td>1.00±0.00</td>
<td>0.90±0.13</td>
<td>1.00±0.00</td>
</tr>
</tbody>
</table>
Benchmark study

- Hard 2: 40% of message words are sampled in different topics than the actual topic.

<table>
<thead>
<tr>
<th>Hard 2</th>
<th>Method</th>
<th>Scenario A node ARI</th>
<th>Scenario A edge ARI</th>
<th>Scenario B node ARI</th>
<th>Scenario B edge ARI</th>
<th>Scenario C node ARI</th>
<th>Scenario C edge ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBM</td>
<td>1.00±0.00</td>
<td>-</td>
<td>-0.01±0.01</td>
<td>-</td>
<td>0.65±0.05</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>LDA</td>
<td>-</td>
<td>0.21±0.13</td>
<td>-</td>
<td>0.08±0.06</td>
<td>-</td>
<td>0.09±0.05</td>
<td></td>
</tr>
<tr>
<td>STBM</td>
<td>0.99±0.02</td>
<td>0.99±0.01</td>
<td>0.59±0.35</td>
<td>0.54±0.40</td>
<td>0.68±0.07</td>
<td>0.62±0.14</td>
<td></td>
</tr>
</tbody>
</table>

The joint model of network structure and topics allows to recover the complex hidden structure in a network with textual edges.
Email communications between 149 employees from 1999-2002. All messages sent between 2 individuals were coerced in a single meta-message $\Rightarrow 1234$ directed edges

Run V-CEM for STBM, for number of groups $Q = 1 : 14$ and number of topics $K = 2 : 20$ $\Rightarrow$ Model selection $(Q, K) = (10, 5)$
### Real-world data

#### Enron email network

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>cycle</td>
<td>griegby</td>
<td>edison</td>
<td>backup</td>
<td>mbhlud</td>
</tr>
<tr>
<td>afghanistan</td>
<td></td>
<td>pac</td>
<td>seal</td>
<td>harrie</td>
</tr>
<tr>
<td>viewing</td>
<td>interview</td>
<td>test</td>
<td></td>
<td>watson</td>
</tr>
<tr>
<td>desk</td>
<td>table</td>
<td>location</td>
<td></td>
<td>capacity</td>
</tr>
<tr>
<td>philip</td>
<td>enery crisis</td>
<td>building</td>
<td>transaction</td>
<td></td>
</tr>
<tr>
<td>ground</td>
<td>day</td>
<td>supplies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>park</td>
<td>july</td>
<td>computer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>date</td>
<td></td>
<td>announcement</td>
<td>master</td>
<td></td>
</tr>
<tr>
<td>israel</td>
<td></td>
<td>notified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>named</td>
<td>cat</td>
<td>phones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sheppard</td>
<td>october</td>
<td>seats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fundamental</td>
<td>october</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tori</td>
<td>california</td>
<td>locations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>routine</td>
<td>california</td>
<td></td>
<td>regular</td>
<td></td>
</tr>
<tr>
<td>allen</td>
<td>masa</td>
<td></td>
<td></td>
<td>indy</td>
</tr>
<tr>
<td>erms</td>
<td>descouche</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>isa</td>
<td>phone</td>
<td></td>
<td></td>
<td>sara</td>
</tr>
<tr>
<td>keyyork</td>
<td>power</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gaskill</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heizerneader</td>
<td>contracts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>topics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure** – Most specific words for the 5 found topics with STBM on the Enron data set.

1. **Financial and trading activity**
2. **Enron activities in Afghanistan**
3. **California electricity crisis**
4. **Usual logistic issues (building equipment, computers, ...)**
5. **technical discussions on gas deliveries**
Group 10 contains a single individual who
- has a central place in the network
- frequently discusses about logistic issues (topic 4) with groups 4, 5, 6 and 7.
Group 8 contains 6 individuals who mainly communicate about Enron activities in Afghanistan (topic 2) between them and with other groups.

Group 4 and 6 are more focused on trading activities (topic 1).

Group 1, 3 and 9 deal with technical issues on gas deliveries (topic 5).
Enron email network

**Figure** – Clustering results with SBM (left, $Q = 8$) and STBM (right) on the Enron data set.

- Some clusters found by SBM (ex. red) have been split by STBM since some nodes use different topics than the rest.
- SBM isolates two "hubs" (light green) $\leftrightarrow$ STBM identify a unique "hub" and the second is gathered with others using similar topics.
Enron email network

**FIGURE** – Clustering results with SBM (left, $Q = 8$) and STBM (right) on the Enron data set.

STBM allows a better and deeper understanding of the Enron network.
NIPS co-authorship network

1988-2003 editions (Nips 1-17 http://robotics.stanford.edu/~gal/data.html) contains the abstracts of 2,484 accepted papers from 2,740 contributing authors.

⇒ undirected network between 2,740 authors with 22,640 textual edges.

Model selection by ICL : \((Q, K) = (13, 7)\)
FIGURE — Clustering result with STBM on the Nips co-authorship network
**Figure** – Most specific words for the 5 found topics with STBM on the Nips co-authorship network.
STBM has proved its ability to bring out concise and relevant analyses on the structure of a large and dense network.
Conclusion

- **STBM**: modeling and clustering vertices in networks with textual edges
  - directed or undirected network
  - application to various types of network
- **C-VEM**: model inference
- **ICL**: model selection
- Numerical experiments on simulated data
- Two real worlds networks
  - large co-authorship network → scalability
Authors

- Bouveyron Charles
  http://w3.mi.parisdescartes.fr/~cbouveyr/

- Pierre Latouche
  http://samm.univ-paris1.fr/Pierre-Latouche

- Zreik Rawya
  http://samm.univ-paris1.fr/Rawya-ZREIK
https://linkage.fr/

Innovative and efficient cluster analysis of networks with textual edges

Linkage allows you to cluster the nodes of networks with textual edges while identifying topics which are used in communications. You can analyze with Linkage networks such as email networks or co-authorship networks. Linkage allows you to upload your own network data or to make requests on scientific databases (Arxiv, Pubmed, HAL).
Thank you!