# The dynamic stochastic topic block model

Marco Corneli (C. Bouveyron, P. Latouche, F. Rossi)

Université Côte d'Azur
Laboratoire LJAD
https://math.unice.fr/~mcorneli/

15 October, DS Meetup

# Outline

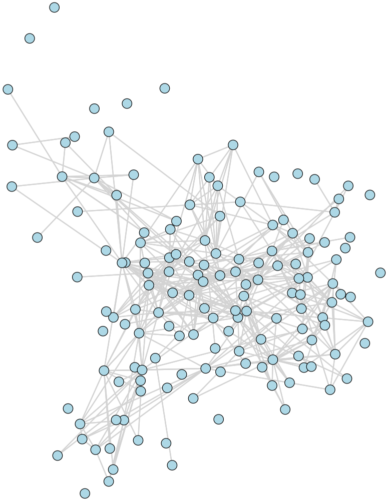
Introduction

STBM

DSTBM

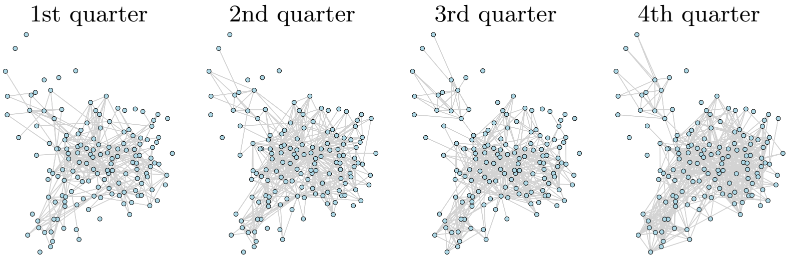
Analysis of the Enron scandal

# the Enron Email dataset (2001)



Nodes + edges

# the Enron Email dataset (2001)




Back to DSTBM

# Introduction

Types of networks: ($\rightarrow$ development of statistical approaches)

- Binary + static edges
- Discrete / continuous / categorical / ...
- Covariates on vertices / edges
- Dynamic edges:
  - Continous time $\rightarrow$ point processes
  - Discrete time $\rightarrow$ Markov,...

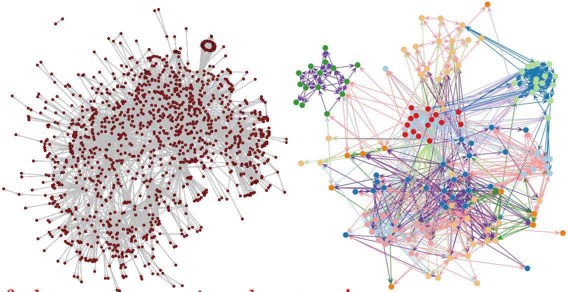
Types of clusters: ($\rightarrow$ development of statistical approaches)

- Communities (transitivity)
- Heterogeneous clusters
- Partitions, overlapping clusters, hierarchy

# Introduction

Networks can be observed directly or indirectly from a variety of sources:

- social websites (Facebook, Twitter, ...),
- personal emails (from your Gmail, Clinton's mails, ...),
- emails of a company (Enron Email data),
- digital/numeric documents (Panama papers, co-authorships, ...),
- and even archived documents in libraries (digital humanities).



⇒ most of these sources involve text!

# Introduction

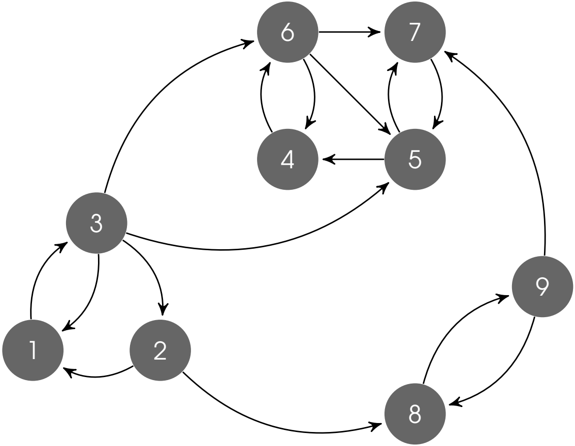



Figure: An (hypothetic) email network between a few individuals.

# Introduction

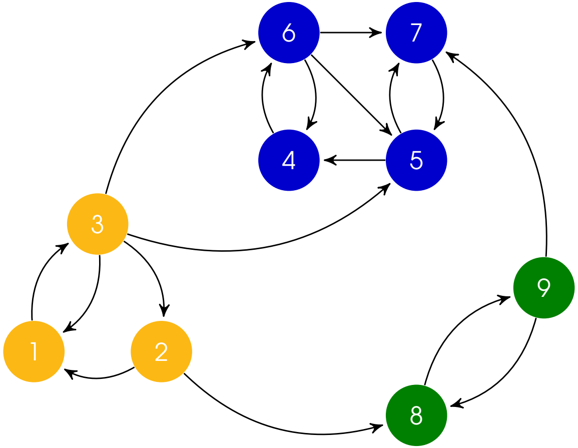



Figure: A typical clustering result for the (directed) binary network.

# Introduction

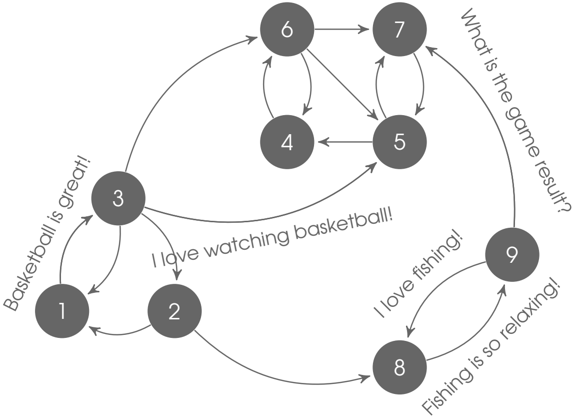



Figure: The (directed) network with textual edges.

# Introduction

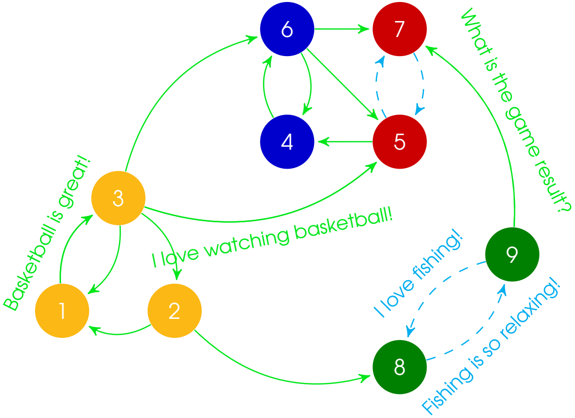



Figure: Expected clustering result for the (directed) network with textual edges.

# The stochastic topic block model

the stochastic topic block model (STBM) [BLZ16]:

- generalizes both SBM and LDA models
- allows to analyze (directed and undirected) networks with textual edges.

But: cannot deal with dynamic networks !!

# The stochastic topic block model

the stochastic topic block model (STBM) [BLZ16]:

- generalizes both SBM and LDA models
- allows to analyze (directed and undirected) networks with textual edges.

But: cannot deal with dynamic networks !!
Goal: develop a dynamic extension of STBM

# Outline

Introduction

STBM

DSTBM

Analysis of the Enron scandal

# Context and notations

We are interesting in clustering the nodes of a (directed) network of $M$ vertices into $Q$ groups:

# Context and notations

We are interesting in clustering the nodes of a (directed) network of $M$ vertices into $Q$ groups:

- the network is represented by its $M \times M$ adjacency matrix $A$:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

# Context and notations

We are interesting in clustering the nodes of a (directed) network of $M$ vertices into $Q$ groups:

- the network is represented by its $M \times M$ adjacency matrix $A$:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

- if $A_{ij} = 1$, the textual edge is characterized by a set of $D_{ij}$ documents:

$$W_{ij} = (W_{ij}^1, ..., W_{ij}^d, ..., W_{ij}^{D_{ij}})$$

# Context and notations

We are interesting in clustering the nodes of a (directed) network of $M$ vertices into $Q$ groups:

- the network is represented by its $M \times M$ adjacency matrix $A$:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

- if $A_{ij} = 1$, the textual edge is characterized by a set of $D_{ij}$ documents:

$$W_{ij} = (W_{ij}^{1}, ..., W_{ij}^{d}, ..., W_{ij}^{D_{ij}})$$

- each document $W_{ij}^{d}$ is made of $N_{ij}^{d}$ words:

$$W_{ij}^{d} = (W_{ij}^{d1}, ..., W_{ij}^{dn}, ..., W_{ij}^{dN_{ij}^{d}}).$$

# Modeling of the edges

Let us assume that edges are generated according to a SBM model:

- each node $i$ is associated with an (unobserved) group among $Q$ according to:

$$Y_i \sim \mathcal{M}(1, \rho),$$

where $\rho \in [0, 1]^Q$ is the vector of group proportions,

# Modeling of the edges

Let us assume that edges are generated according to a SBM model:

- each node $i$ is associated with an (unobserved) group among $Q$ according to:

$$Y_i \sim \mathcal{M}(1, \rho),$$

  where $\rho \in [0, 1]^Q$ is the vector of group proportions,

- the presence of an edge $A_{ij}$ between $i$ and $j$ is drawn according to:

$$A_{ij} | Y_{iq} Y_{jr} = 1 \sim \mathcal{B}(\pi_{qr}),$$

  where $\pi_{qr} \in [0, 1]$ is the connection probability between clusters $q$ and $r$.

# Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters $(q, r)$ is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:

$$\theta_{qr} \sim \mathrm{Dir}\left(\alpha\right),$$

  such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.

# Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters $(q,r)$ is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:

  $$\theta_{qr} \sim \mathrm{Dir}\left(\alpha\right),$$

  such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q,r)$.

- the $n$th word $W_{ij}^{dn}$ of documents $d$ in $W_{ij}$ is then associated to a latent topic vector $Z_{ij}^{dn}$ according to:

  $$Z_{ij}^{dn} | \left\{A_{ij} Y_{iq} Y_{jr} = 1, \theta\right\} \sim \mathcal{M}\left(1, \theta_{qr}\right).$$

# Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters $(q, r)$ is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:

$$\theta_{qr} \sim \mathrm{Dir}\left(\alpha\right),$$

  such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.

- the $n$th word $W_{ij}^{dn}$ of documents $d$ in $W_{ij}$ is then associated to a latent topic vector $Z_{ij}^{dn}$ according to:

$$Z_{ij}^{dn} | \left\{A_{ij} Y_{iq} Y_{jr} = 1, \theta\right\} \sim \mathcal{M}\left(1, \theta_{qr}\right).$$

- then, given $Z_{ij}^{dn}$, the word $W_{ij}^{dn}$ is assumed to be drawn from a multinomial distribution:

$$W_{ij}^{dn} | Z_{ij}^{dnk} = 1 \sim \mathcal{M}\left(1, \beta_k = (\beta_{k1}, \ldots, \beta_{kV})\right),$$

  where $V$ is the vocabulary size.

# Modeling of the documents

- notice that the two previous equations lead to the following mixture model for words over topics:

$$W_{ij}^{dn} | \{Y_{iq} Y_{jr} A_{ij} = 1, \theta\} \sim \sum_{k=1}^{K} \theta_{qrk} \mathcal{M}(1, \beta_k).$$

# STBM at a glance...

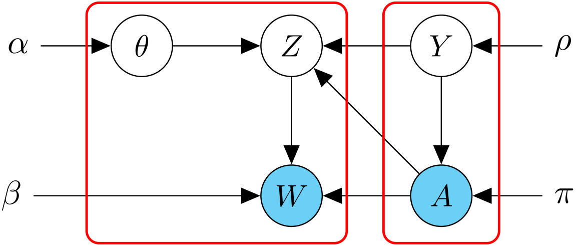



Figure: The stochastic topic block model.

# STBM at a glance...

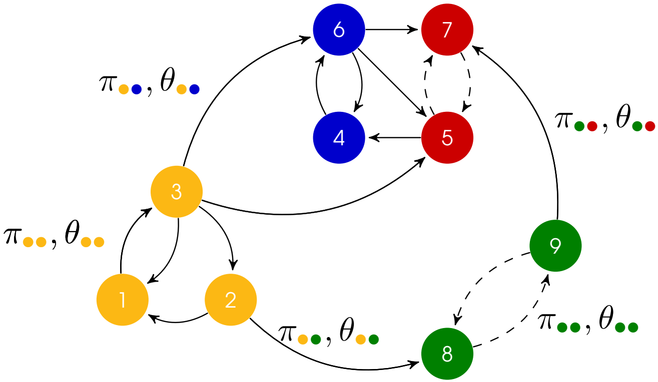



Figure: The stochastic topic block model.

# Inference

A likelihood based approach is adopted. The aim is to maximize the complete data log-likelihood

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

with respect to $(\rho, \pi, \beta)$ and $Y = (Y_1, \ldots, Y_M)$.

# Inference

A likelihood based approach is adopted. The aim is to maximize the complete data log-likelihood

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

with respect to $(\rho, \pi, \beta)$ and $Y = (Y_1, \dots, Y_M)$.

Strategy:

- an approximated Expectation-Maximization (**VEM**) algorithm is used to estimate the optimal $(\rho, \pi, \theta)$,
- a *greedy* classification (**C**) over the node labels $Y$ is performed.

The above two steps (C-VEM) are repeated until convergence.

# Outline

Introduction

STBM

DSTBM

Analysis of the Enron scandal

# Dynamic data

Several network **snapshots** are observed during the time interval $[0, T]$: ▸ Enron Snapshots

# Dynamic data

Several network **snapshots** are observed during the time interval $[0, T]$: ▸ Enron Snapshots

We need a partition!

# Dynamic data

Several network **snapshots** are observed during the time interval $[0, T]$: ▸ Enron Snapshots

We need a partition!

A partition of the time interval $[0, T]$ is introduced

$$0 = t_0 < t_1 < \cdots < t_U = T$$

and $I_u := [t_{u-1}, t_u[$, $\Delta_u$ is the size of $I_u$.

A sequence of static networks is built by "summing" the interactions between all pairs $(i, j)$ over all $I_u$.

# The dynamic stochastic topic block model

- From $A_{ij}$ to $D_{iju}$: the number of interactions that occurred between $i$ and $j$ during $I_u$
- From $W_{ij}^{dn}$ to $W_{ij}^{un}$: $n$th word during $I_u$
- Each time interval is assumed to belong to an unknown time cluster:

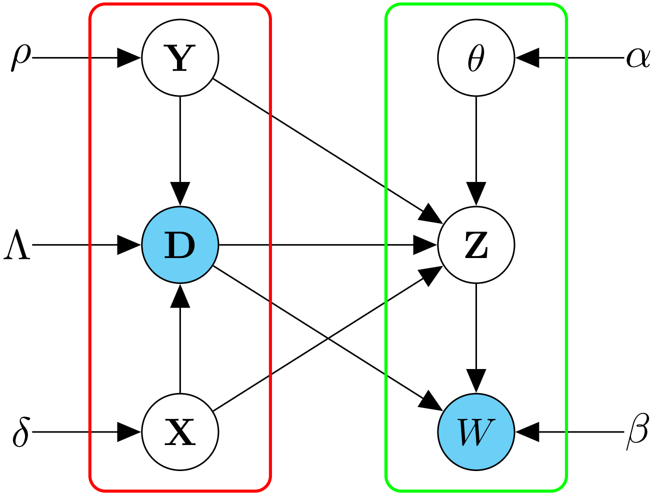
$$X_u \sim \mathcal{M}(1, \delta),$$

  where $\delta \in [0, 1]^L$ is the vector of time cluster proportions.

# The dynamic stochastic topic block model

- $Y_i \sim \mathcal{M}(1, \rho)$ iid
- $X_u \sim \mathcal{M}(1, \delta)$ iid
- $D_{iju} | Y_{iq} Y_{jr} X_{ul} = 1 \sim \mathcal{P}(\lambda_{qrl} \Delta_u)$
- $\theta_{qrl} \sim \mathrm{Dir}(\alpha)$
- $Z_{ij}^{un} | \{D, Y_{iq} Y_{jr} X_{ul} = 1\} \sim \mathcal{M}(1, \theta_{qrl})$
- $W_{ij}^{un} | Z_{ij}^{unk} = 1 \sim \mathcal{M}(1, \beta_k)$.

# The dynamic stochastic topic block model



# Inference

- Same trick as for STBM: $Y$ and $X$ are pivotal
- Consider $\log(D, W, Y, X | \rho, \delta, \Lambda, \beta)$
- C-VEM: maximize the log-likelihood with respect to $R(\cdot), Y, X, \rho, \delta, \Lambda, \beta$, in turn.

# Model selection

$$
\begin{aligned}
ICL = \tilde{\mathcal{L}}(R(\cdot); D, Y, X, D, \beta) - \frac{K(V-1)}{2} \log(LQ^2) \\
+ \max_{\Lambda,\rho,\delta} \log p(D, Y, X|\Lambda, \rho, \delta) - \frac{LQ^2}{2} \log(MU(M-1)) \\
- \frac{Q-1}{2} \log(M) - \frac{L-1}{2} \log(U).
\end{aligned}
$$

# Outline

Introduction

STBM

DSTBM

Analysis of the Enron scandal

# Analysis of the Enron scandal

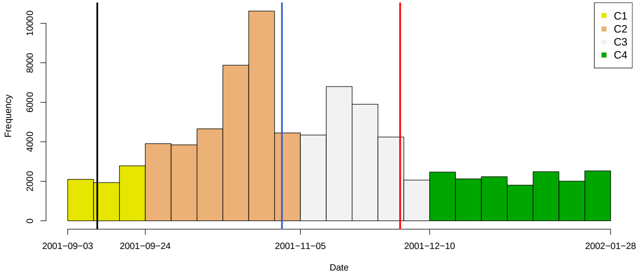
- All email exchanges between 149 Enron employees
- Time window considered: September, 3rd, 2001 to January, 28th, 2002
- Three key dates:
  - September, 11th, 2001: the terrorist attacks to the Twin Towers and the Pentagon
  - October, 31st, 2001: the Securities and Exchange Commission (SEC) opened on investigation for fraud concerning Enron.
  - December, 2nd, 2001: Enron filed for bankruptcy, resulting in more than 4,000 lost jobs.

# Analysis of the Enron scandal

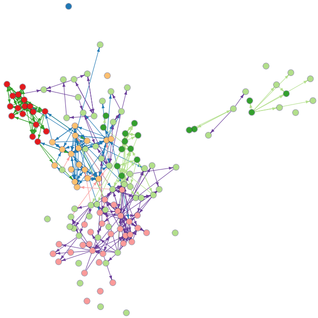
- The selected time window is partitioned into subintervals, each interval corresponding to a week.
- $U = 21$ weeks
- 4321 directed edges
- Dictionary: 49955 words
- Test models $(Q, K, L) \in \{1, \ldots, 10\}^3$

# Results

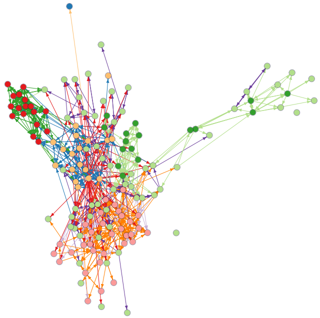
- Model selection: $Q = 6$, $K = 9$, $L = 4$
- Time clusters:



# Results : clusters of nodes

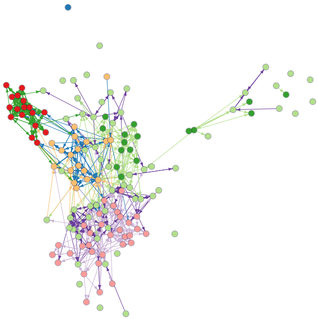


(a) Time cluster $\mathcal{C}_1$.

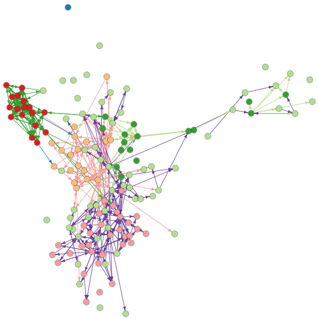


(b) Time cluster $\mathcal{C}_2$.

# Results : clusters of nodes

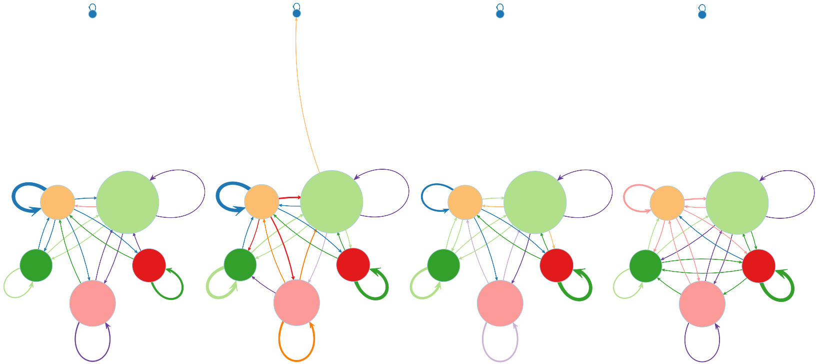


(c) Time cluster $\mathcal{C}_3$.



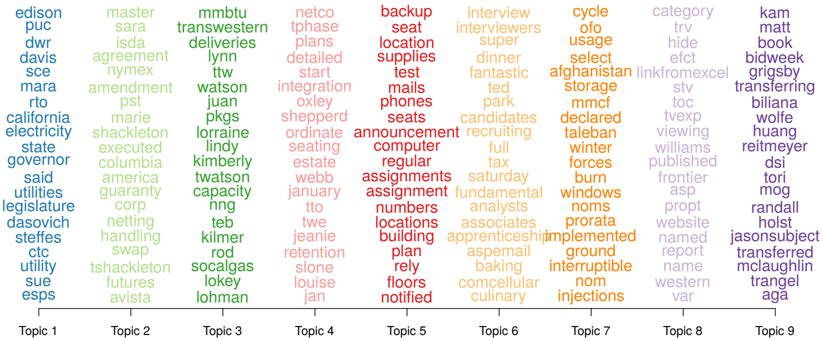
(d) Time cluster $\mathcal{C}_4$.

# Results : clusters of nodes



(e) $\mathcal{C}_1$. (f) $\mathcal{C}_2$. (g) $\mathcal{C}_3$. (h) $\mathcal{C}_4$.

# Results : topics



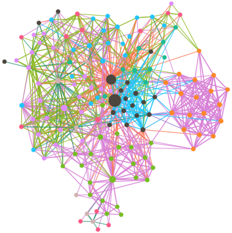
# Conclusion

- DSTBM : allows to model temporal networks with textual edges
- C-VEM algorithm for inference
- Model selection criterion
- Find clusters of nodes and topics of discussions

Thanks for your attention

# Linkage.fr

Linkage    New Job    Jobs    pierre



## 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).

Try Linkage

# Simulations

| Model | Setup A node ARI | time ARI | edge ARI |
|---|---|---|---|
| dSTBM | 0.99 (0.06) | 1 (0) | 0.99 (0.06) |
| dSBM | 1 (0) | 1 (0) | - |
| STBM | 1 (0) | - | 0.66 (0.21) |
| SBM | 0.01 (0.06) | - | - |
| LDA | - | - | 0.73 (0.20) |

| Model | Setup B node ARI | time ARI | edge ARI |
|---|---|---|---|
| dSTBM | 1 (0) | 1 (0) | 1 (0) |
| dSBM | 0.98 (0.03) | 0.00 (0.01) | - |
| STBM | 0.5 (0.5) | - | 0.02 (0.03) |
| SBM | 0.99 (0.04) | - | - |
| LDA | - | - | 1 (0) |

| Model | Setup C node ARI | time ARI | edge ARI |
|---|---|---|---|
| dSTBM | 1 (0) | 1(0) | 1 (0) |
| dSBM | 0.67 (0.05) | 0.00 (0.01) | - |
| STBM | 1 (0) | - | 0.70 (0.10) |
| SBM | 0.65 (0.04) | - | - |
| LDA | - | - | 0.69 (0.15) |

# Biblio I


- Charles Bouveyron, Pierre Latouche, and Rawya Zreik, *The stochastic topic block model for the clustering of vertices in networks with textual edges*, Statistics and Computing (2016), 1–21.
- Peter D Hoff, Adrian E Raftery, and Mark S Handcock, *Latent space approaches to social network analysis*, Journal of the american Statistical association **97** (2002), no. 460, 1090–1098.
- K. Nowicki and T.A.B. Snijders, *Estimation and prediction for stochastic blockstructures*, Journal of the American Statistical Association **96** (2001), 1077–1087.
- Y.J. Wang and G.Y. Wong, *Stochastic blockmodels for directed graphs*, Journal of the American Statistical Association **82** (1987), 8–19.