Sparsity-sensitive Diagonal Co-clustering Algorithms for the Effective Handling of Text Data

Melissa Ailem

Prix de Thèse EGC'2018

Thesis Supervisors : Mohamed Nadif and François Role
Paris Descartes University

University of Southern California (USA) and INRIA Lille (France)

January 26, 2018

Outline

1. Introduction
    - Context
    - Co-clustering
    - Motivations
2. Graph-based Co-clustering
    - Graph Modularity
    - Modularity for Co-clustering
    - Experiments
3. Model-based Co-clustering
    - Sparse Poisson Latent Block Model (SPLBM)
    - Soft SPLBM-based Co-clustering Algorithm
    - Hard SPLBM-based Co-clustering Algorithm
    - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
    - The Biomedical Framework
    - Results and Discussions
5. Conclusion and Perspectives

Outline

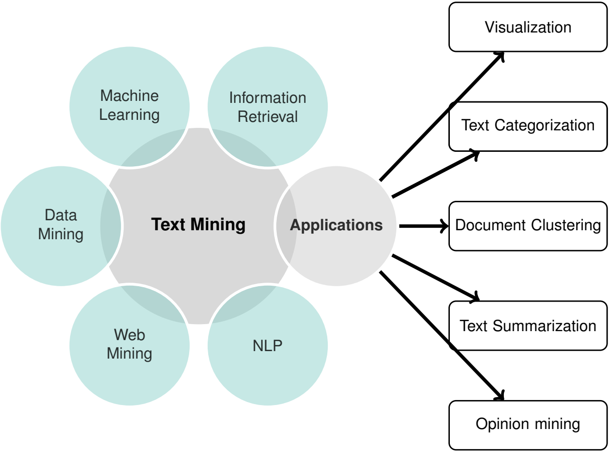
1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives

Outline

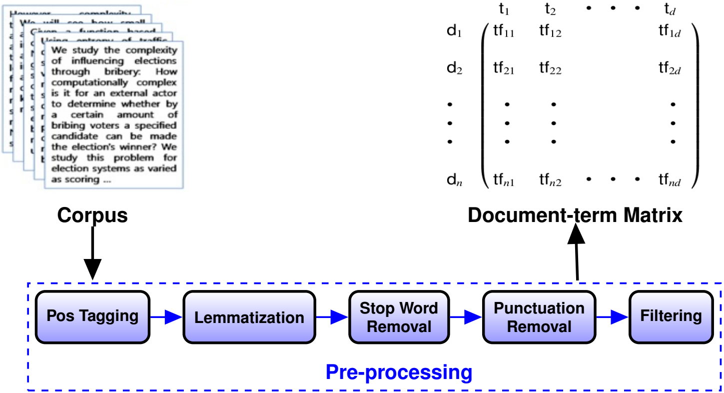
1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives

- Exponential growth of textual documents on the web, e.g. the PUBMED database contains more than 20 millions of biomedical articles
- It is become more laborious to access what we are looking for
- We need automated Text Mining tools to help us understand, interpret and organize this vast amount of information

- Exponential growth of textual documents on the web, e.g. the PUBMED database contains more than 20 millions of biomedical articles
- It is become more laborious to access what we are looking for
- We need automated Text Mining tools to help us understand, interpret and organize this vast amount of information



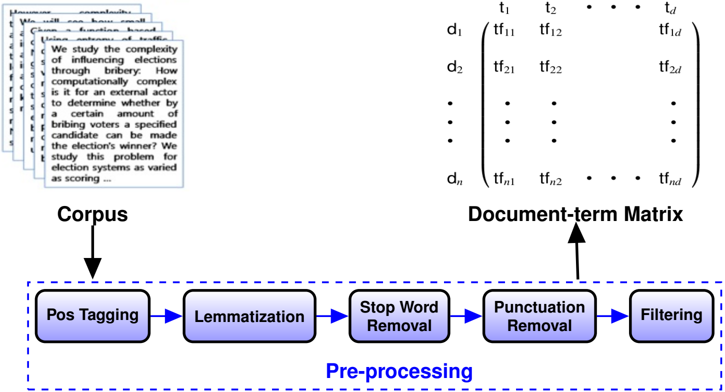
Data Representation



Document Representation

- Vector space model
- tf_{ij}=Frequency of term j in document i

Data Representation



$$
\begin{array}{c|cccc}
 & t_1 & t_2 & \cdots & t_d \\
\hline
d_1 & tf_{11} & tf_{12} & & tf_{1d} \\
d_2 & tf_{21} & tf_{22} & & tf_{2d} \\
\vdots & \vdots & \vdots & & \vdots \\
d_n & tf_{n1} & tf_{n2} & \cdots & tf_{nd}
\end{array}
$$

Document Representation

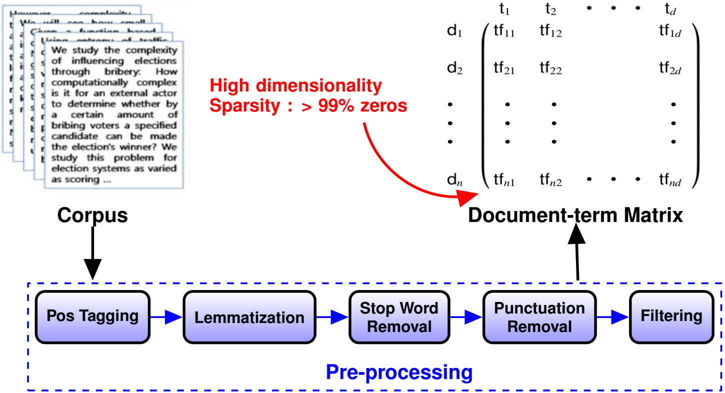
- Vector space model
- tf_{ij}=Frequency of term j in document i

Weighting scheme

- TF-IDF weighting scheme

$$w_{ij} = tf_{ij} \times \log \frac{N}{d_j}$$

Data Representation



$$
\begin{array}{c|cccc}
 & t_1 & t_2 & \cdots & t_d \\
\hline
d_1 & tf_{11} & tf_{12} & & tf_{1d} \\
d_2 & tf_{21} & tf_{22} & & tf_{2d} \\
\vdots & \vdots & \vdots & & \vdots \\
d_n & tf_{n1} & tf_{n2} & \cdots & tf_{nd}
\end{array}
$$

Document Representation

- Vector space model
- tf_{ij}=Frequency of term j in document i

Weighting scheme

- TF-IDF weighting scheme

$$w_{ij} = tf_{ij} \times \log \frac{N}{d_j}$$

Document Clustering :

- A widely used unsupervised learning technique, to group together similar documents based on their content
- Documents within a cluster are semantically coherent or deal with the same topics

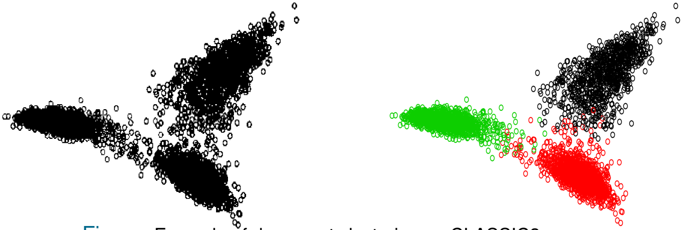


Figure: Example of document clustering on CLASSIC3 corpus

Document Clustering :

- A widely used unsupervised learning technique, to group together similar documents based on their content
- Documents within a cluster are semantically coherent or deal with the same topics

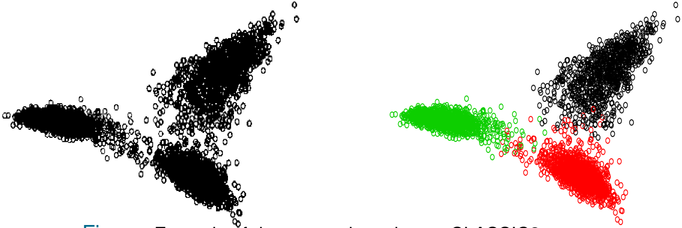


Figure: Example of document clustering on CLASSIC3 corpus

Advantages :

- Organization of documents, efficient browsing and navigation of huge text corpora, speed up search engines, etc.

Document Clustering :

- A widely used unsupervised learning technique, to group together similar documents based on their content
- Documents within a cluster are semantically coherent or deal with the same topics

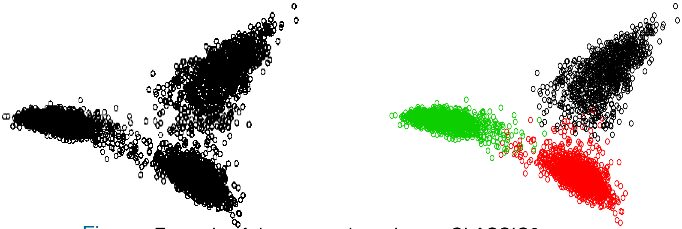


Figure: Example of document clustering on CLASSIC3 corpus

Advantages :

- Organization of documents, efficient browsing and navigation of huge text corpora, speed up search engines, etc.

Challenges :

- High dimensionality
- Sparsity

Outline

1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices Hartigan, 1972

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices Hartigan, 1972

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0

(a) Original Data

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices Hartigan, 1972

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0

(a) Original Data

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0

(b) Clustering

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices Hartigan, 1972

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0

(a) Original Data

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0

(b) Clustering

	Term 4	Term 10	Term 3	Term 6	Term 12	Term 2	Term 9	Term 5	Term 11	Term 7	Term 1	Term 8
Doc 1	1	1	1	1	0	0	0	0	1	1	1	1
Doc 2	1	1	1	1	0	0	1	0	1	1	1	1
Doc 3	1	1	0	1	0	0	0	0	1	1	1	1
Doc 4	0	0	0	0	1	1	1	1	0	0	0	0
Doc 5	0	0	0	0	1	1	1	0	0	0	1	0
Doc 6	0	0	0	0	1	1	1	1	0	0	0	0
Doc 7	0	0	0	0	0	0	0	0	1	1	1	1
Doc 8	0	0	0	0	0	0	0	0	1	1	1	1
Doc 9	0	0	0	0	0	1	0	0	1	1	1	1

(c) Co-clustering

Co-clustering

Co-clustering

- It is an important extension of traditional one-sided clustering, that addresses the problem of simultaneous clustering of both dimensions of data matrices Hartigan, 1972

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0

(a) Original Data

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10	Term 11	Term 12
Doc 7	1	0	1	1	0	1	1	1	0	1	1	0
Doc 3	1	0	1	1	0	1	1	1	1	1	1	0
Doc 2	1	0	0	1	0	1	1	1	0	1	1	0
Doc 8	0	1	0	0	1	0	0	0	1	0	0	1
Doc 6	1	1	0	0	0	0	0	0	1	0	0	1
Doc 1	0	1	0	0	1	0	0	0	1	0	0	1
Doc 4	1	0	0	0	0	0	1	1	0	0	1	0
Doc 9	1	0	0	0	0	0	1	1	0	0	1	0
Doc 5	1	1	0	0	0	0	1	1	0	0	1	0

(b) Clustering

	Term 4	Term 10	Term 3	Term 6	Term 12	Term 2	Term 9	Term 5	Term 11	Term 7	Term 1	Term 8
Doc 1	1	1	1	1	0	0	0	0	1	1	1	1
Doc 2	1	1	1	1	0	0	1	0	1	1	1	1
Doc 3	1	1	0	1	0	0	0	0	1	1	1	1
Doc 4	0	0	0	0	1	1	1	1	0	0	0	0
Doc 5	0	0	0	0	1	1	1	0	0	0	1	0
Doc 6	0	0	0	0	1	1	1	1	0	0	0	0
Doc 7	0	0	0	0	0	0	0	0	1	1	1	1
Doc 8	0	0	0	0	0	0	0	0	1	1	1	1
Doc 9	0	0	0	0	0	1	0	0	1	1	1	1

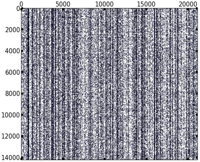
(c) Co-clustering

Why Co-clustering?

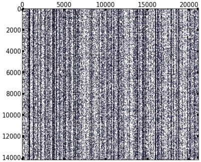
- Exploit the duality between object space and attribute space
- Cluster Characterization
- Technique for dimensionality reduction
- Reduce Computation time

Outline

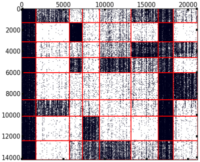
1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives



(a) Original Data



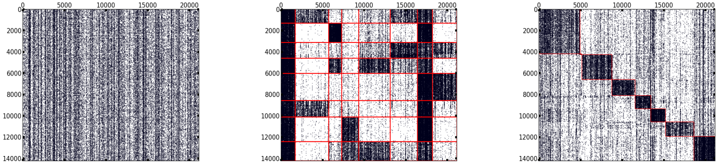
(a) Original Data



(b) General co-clustering

Motivations

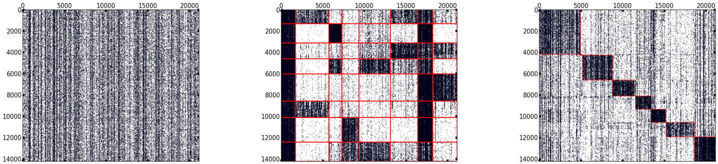
- When dealing with high dimensional sparse data, several co-clusters are primarily composed of zeros.
- Seeking homogeneous blocks is not sufficient to produce meaningful results.



(a) Original Data (b) General co-clustering (c) Diagonal co-clustering

Motivations

- When dealing with high dimensional sparse data, several co-clusters are primarily composed of zeros.
- Seeking homogeneous blocks is not sufficient to produce meaningful results.
- Seeking diagonal structure turns out to be more beneficial.
  - In good agreement with sparsity
  - Produces directly the most relevant co-clusters and ignore noisy ones
  - Cluster hypothesis
  - Allows a direct interpretation of co-clusters
  - Parsimonious



(a) Original Data (b) General co-clustering (c) Diagonal co-clustering

Motivations

- When dealing with high dimensional sparse data, several co-clusters are primarily composed of zeros.
- Seeking homogeneous blocks is not sufficient to produce meaningful results.
- Seeking diagonal structure turns out to be more beneficial.
  - In good agreement with sparsity
  - Produces directly the most relevant co-clusters and ignore noisy ones
  - Cluster hypothesis
  - Allows a direct interpretation of co-clusters
  - Parsimonious

Contributions

- Graph-based block diagonal clustering
- Model-based block diagonal clustering

Outline

1 Introduction
- Context
- Co-clustering
- Motivations

2 Graph-based Co-clustering
- Graph Modularity
- Modularity for Co-clustering
- Experiments

3 Model-based Co-clustering
- Sparse Poisson Latent Block Model (SPLBM)
- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
- Experiments

4 Using Co-clustering in Biomedical Text Mining Framework
- The Biomedical Framework
- Results and Discussions

5 Conclusion and Perspectives

Contributions

Motivations

- Existing graph-based Co-clustering approaches use a spectral relaxation of the discrete optimization problem
  - Find minimum cut using spectral relaxation (Dhillon, 2001)
  - Find maximum Modularity using spectral relaxation (Labiod and Nadif, 2011)
- Eigen vector computation may be prohibitive when dealing with high dimensional matrices

Contributions

Motivations

- Existing graph-based Co-clustering approaches use a spectral relaxation of the discrete optimization problem
  - Find minimum cut using spectral relaxation (Dhillon, 2001)
  - Find maximum Modularity using spectral relaxation (Labiod and Nadif, 2011)
- Eigen vector computation may be prohibitive when dealing with high dimensional matrices

Contributions

- We propose a new block-diagonal clustering algorithm (Coclus) (Ailem, Role, and Nadif, 2015; Ailem, Role, and Nadif, 2016)
- Coclus is based on the direct maximization of graph modularity
- Use an iterative alternating optimization procedure

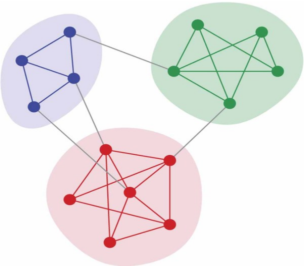
M. Ailem, F. Role, and M. Nadif (2015). "Co-clustering Document-term Matrices by Direct Maximization of Graph Modularity". In: *CIKM'2015*. ACM, pp. 1807–1810.

M. Ailem, F. Role, and M. Nadif (2016). "Graph modularity maximization as an effective method for co-clustering text data". In: *Knowledge-Based Systems Journal* 109, pp. 160–173.

Outline

1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives

Graph Modularity



- Introduced by Newman and Girvan (2004)
- Identify community structure in graphs
- Measure the strength of the community structure of a graph
- Maximize the difference between the original graph and its corresponding random version
- Q=(number of intra-cluster edges) - (expected number of edges)

Given the graph $G(V,E)$ and its corresponding adjacency matrix A :

$$Q(\mathbf{A},\mathbf{C}) = \frac{1}{2|E|}\sum_{i=1}^{n}\sum_{i'=1}^{n}(a_{ii'} - \frac{a_{i.}a_{i'.}}{2|E|})c_{ii'}, \quad (1)$$

- where $|E|$ represents the number of edges
- $a_{ii'} = 1$ if there is an edge between nodes i and i'
- $a_{i.}$ and $a_{i'.}$ the degree of nodes i and i' respectively, and $\frac{a_{i.}a_{i'.}}{2|E|}$ represents the expected number of edges between nodes i and i'
- $c_{ii'} = \sum_k z_{ik}z_{i'k}$ is equal to 1 if i and i' belong to the same community k

Outline

1 Introduction
- Context
- Co-clustering
- Motivations

2 Graph-based Co-clustering
- Graph Modularity
- Modularity for Co-clustering
- Experiments

3 Model-based Co-clustering
- Sparse Poisson Latent Block Model (SPLBM)
- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
- Experiments

4 Using Co-clustering in Biomedical Text Mining Framework
- The Biomedical Framework
- Results and Discussions

5 Conclusion and Perspectives

Modularity for Co-clustering

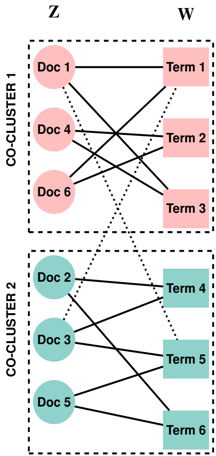
Given a rectangular positive matrix $\mathbf{A}$, modularity can be reformulated as follows in the co-clustering context:

$$Q(\mathbf{A},\mathbf{C}) = \frac{1}{a_{..}} \sum_{i=1}^{n} \sum_{j=1}^{d} (a_{ij} - \frac{a_{i.}a_{.j}}{a_{..}}) c_{ij}, \qquad (2)$$

$$Q(\mathbf{A},\mathbf{Z}\mathbf{W}^t) = \frac{1}{a_{..}} \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{g} (a_{ij} - \frac{a_{i.}a_{.j}}{a_{..}}) z_{ik} w_{jk}, \qquad (3)$$

where $a_{..} = \sum_{i,j} a_{ij} = |E|$ is the number of edges (or edge weights for weighted graphs) and $c_{ij} = \sum_k z_{ik} w_{jk} = 1$ if nodes i and j belong to the same co-cluster k and 0 otherwise

$$Q(\mathbf{A},\mathbf{C}) = \frac{1}{a_{..}} Trace[(\mathbf{A} - \boldsymbol{\delta})^t \mathbf{Z}\mathbf{W}^t] = Q(\mathbf{A},\mathbf{Z}\mathbf{W}^t). \qquad (4)$$



	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6
Doc 1	2	0	1	0	1	0
Doc 4	0	1	2	0	0	0
Doc 6	1	1	0	0	0	0
Doc 2	0	0	0	1	0	1
Doc 3	1	0	0	3	1	0
Doc 5	0	0	0	0	1	1

Alternated Maximization of Modularity

Proposition

Let $\mathbf{A}$ be a $(n \times d)$ positive data matrix and $\mathbf{C}$ be a $(n \times d)$ matrix defining a block seriation, the modularity measure $Q(\mathbf{A}, \mathbf{C})$ can be rewritten as

$$1) \quad Q(\mathbf{A},\mathbf{C}) = \frac{1}{a_{..}} \sum_{i=1}^{n} \sum_{k=1}^{g} (a_{ik}^{\mathbf{W}} - \frac{a_{i.} a_{.k}^{\mathbf{W}}}{a_{..}}) z_{ik} = \frac{1}{a_{..}} Trace[(\mathbf{A}^{\mathbf{W}} - \boldsymbol{\delta}^{\mathbf{W}})^t \mathbf{Z}] = Q(\mathbf{A}^{\mathbf{W}}, \mathbf{Z})$$

where $\boldsymbol{\delta}^{\mathbf{W}} := \{\delta_{ik}^{\mathbf{W}} = \frac{a_{i.} a_{.k}^{\mathbf{W}}}{a_{..}}; i = 1, \ldots, n; k = 1, \ldots, g\}$ with $a_{.k}^{\mathbf{W}} = \sum_{j=1}^{d} w_{jk} a_{.j}$

$$2) \quad Q(\mathbf{A},\mathbf{C}) = \frac{1}{a_{..}} \sum_{j=1}^{d} \sum_{k=1}^{g} (a_{kj}^{\mathbf{Z}} - \frac{a_{.j} a_{k.}^{\mathbf{Z}}}{a_{..}}) w_{jk} = \frac{1}{a_{..}} Trace[(\mathbf{A}^{\mathbf{Z}} - \boldsymbol{\delta}^{\mathbf{Z}}) W] = Q(\mathbf{A}^{\mathbf{Z}}, \mathbf{W})$$

where $\boldsymbol{\delta}^{\mathbf{Z}} := \{\delta_{kj}^{\mathbf{Z}} = \frac{a_{.j} a_{k.}^{\mathbf{Z}}}{a_{..}}; j = 1, \ldots, d; k = 1, \ldots, g\}$ with $a_{k.}^{\mathbf{Z}} = \sum_{i=1}^{n} z_{ik} a_{i.}$

`Coclus` Algorithm

Algorithm 1: `Coclus`

Input : positive data matrix $\mathbf{A}$, number of co-clusters g
Step 1. Initialization of $\mathbf{W}$
repeat

- **Step 2.** Compute $\mathbf{A}^{\mathbf{W}} = \mathbf{AW}$
- **Step 3.** Compute $\mathbf{Z}$ maximizing $Q(\mathbf{A}^{\mathbf{W}}, \mathbf{Z})$ by

$$z_{ik} = \arg\max_{1 \leq \ell \leq g} \left(a^{\mathbf{W}}_{i\ell} - \frac{a_{i.} a^{\mathbf{W}}_{.\ell}}{a_{..}} \right) \forall i = 1, \ldots, n; k = 1, \ldots, g$$

- **Step 4.** Compute $\mathbf{A}^{\mathbf{Z}} = \mathbf{Z}^t \mathbf{A}$
- **Step 5.** Compute $\mathbf{W}$ maximizing $Q(\mathbf{A}^{\mathbf{Z}}, \mathbf{W})$ by

$$w_{jk} = \arg\max_{1 \leq \ell \leq g} \left(a^{\mathbf{Z}}_{\ell j} - \frac{a^{\mathbf{Z}}_{\ell .} a_{.j}}{a_{..}} \right) \forall j = 1, \ldots, d; k = 1, \ldots, g$$

- **Step 6.** Compute $Q(\mathbf{A}, \mathbf{ZW}^t)$

until *Convergence*;
Output : partition matrices $\mathbf{Z}$ and $\mathbf{W}$, and modularity value $\mathbf{Q}$

Complexity : $O(nz \cdot it \cdot g)$

Outline

1 Introduction
- Context
- Co-clustering
- Motivations

2 Graph-based Co-clustering
- Graph Modularity
- Modularity for Co-clustering
- Experiments

3 Model-based Co-clustering
- Sparse Poisson Latent Block Model (SPLBM)
- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
- Experiments

4 Using Co-clustering in Biomedical Text Mining Framework
- The Biomedical Framework
- Results and Discussions

5 Conclusion and Perspectives

Datasets	Characteristics				
	#Documents	#Words	#Clusters	Sparsity (%)	Balance
CLASSIC4	7095	5896	4	99.41	0.323
NG20	19949	43586	20	99.99	0.991
SPORTS	8580	14870	7	99.99	0.036
REVIEWS	4069	18483	5	99.99	0.098

- Evaluation measure : Accuracy (Acc) and Normalized mutual information (NMI) (Strehl and Ghosh, 2003)
- Data Types : binary, contingency and TF-IDF

Method	**Data Type**	**References**	**Co-clustering**	**Type of implementation**
Spec	Positive data	(I. Dhillon, 2001)	Diagonal	Scickit Learn
Block	Binary	(Li, 2005)	Diagonal	Our python implementation
ITCC	Positive data	(I. S. Dhillon, Mallela, and D. S. Modha, 2003)	Non-diagonal	C++ implementation
SpecCo	Positive data	(Labiod and Nadif, 2011)	Diagonal	Our python implementation
χ-*Sim*	Positive data	(Bisson and Hussain, 2008)	Non-diagonal	MATLAB implementation of the authors
FNMTF	Positive data	(Wang et al., 2011)	Non-diagonal	MATLAB implementation of the authors

Datasets	Characteristics				
	#Documents	#Words	#Clusters	Sparsity (%)	Balance
CLASSIC4	7095	5896	4	99.41	0.323
NG20	19949	43586	20	99.99	0.991
SPORTS	8580	14870	7	99.99	0.036
REVIEWS	4069	18483	5	99.99	0.098

- Evaluation measure : Accuracy (Acc) and Normalized mutual information (NMI) (Strehl and Ghosh, 2003)
- Data Types : binary, contingency and TF-IDF

Method	**Data Type**	**References**	**Co-clustering**	**Type of implementation**
Spec	Positive data	(I. Dhillon, 2001)	Diagonal	Scickit Learn
Block	Binary	(Li, 2005)	Diagonal	Our python implementation
ITCC	Positive data	(I. S. Dhillon, Mallela, and D. S. Modha, 2003)	Non-diagonal	C++ implementation
SpecCo	Positive data	(Labiod and Nadif, 2011)	Diagonal	Our python implementation
χ-*Sim*	Positive data	(Bisson and Hussain, 2008)	Non-diagonal	MATLAB implementation of the authors
FNMTF	Positive data	(Wang et al., 2011)	Non-diagonal	MATLAB implementation of the authors

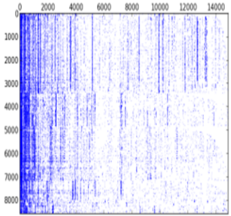
		Binary							**Contingency**						**TF-IDF**					
datasets	per.	Spec	ITCC	Block	SpecCo	χ-Sim	FNMTF	CoClus	Spec	ITCC	SpecCo	χ-Sim	FNMTF	CoClus	Spec	ITCC	SpecCo	χ-Sim	FNMTF	CoClus
CLASSIC4	Acc	0.34	0.65	0.52	0.45	0.31	0.50	**0.90**	0.53	0.87	0.58	0.31	0.56	**0.90**	0.44	0.60	0.45	0.35	0.76	**0.88**
	NMI	0.14	0.51	0.16	0.02	0.15	0.30	**0.72**	0.45	0.67	0.48	0.15	0.30	**0.73**	0.02	0.55	0.009	0.13	0.58	**0.70**
NG20	Acc	0.14	**0.43**	0.20	0.19	0.26	0.13	0.40	0.05	**0.45**	0.30	0.30	0.09	0.37	0.19	**0.41**	0.15	0.29	0.40	0.37
	NMI	0.29	**0.55**	0.22	0.42	0.33	0.03	**0.55**	0.02	**0.52**	0.49	0.37	0.07	**0.52**	0.32	0.44	0.38	0.41	0.40	**0.52**
SPORTS	Acc	0.56	0.45	0.47	0.59	0.57	0.28	**0.70**	0.44	0.56	0.68	0.53	0.36	**0.75**	0.45	0.54	0.61	0.67	0.57	**0.68**
	NMI	0.47	0.49	0.38	0.45	0.48	0.15	**0.54**	0.38	0.58	0.59	0.48	0.19	**0.62**	0.43	0.58	0.45	0.55	0.54	**0.59**
REVIEWS	Acc	0.56	0.58	0.53	0.59	0.46	0.34	**0.65**	0.50	0.71	0.45	0.41	0.38	**0.72**	0.35	0.63	0.46	0.44	0.43	**0.65**
	NMI	0.36	0.46	0.42	0.39	0.31	0.18	**0.54**	0.40	0.57	0.34	0.23	0.17	**0.58**	0.03	0.51	0.35	0.28	0.27	**0.52**

- Results obtained after running each algorithm 100 times with random initialization
- We retained the solution optimizing the associated criterion (maximizing the modularity for *CoClus*)
- Superiority of Coclus in almost all situations
- Robustness w.r.t the type of data (binary tables, contingency tables and TF-IDF weighted tables)

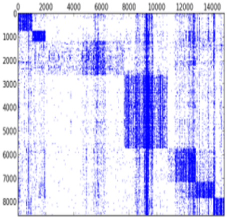
Assessing the Number of Co-clusters

- Most previous co-clustering algorithms require the number of co-clusters as an input parameter
- The modularity measure can be used to predict the right number of co-clusters
- Run *Coclus* algorithm with different values of g (number of co-clusters)
- For each number of co-cluster the modularity is computed
- Retain the number of co-clusters for which the modularity measure reaches it's maximum value

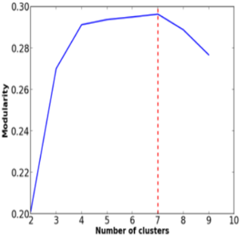
Assessing the Number of Co-clusters



Original matrix (document x term)



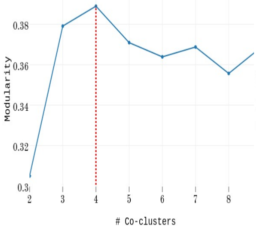
Reorganized matrix (after Co-clustering)



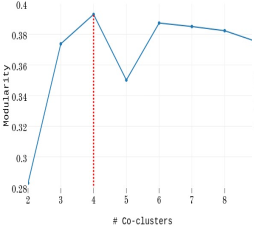
Number of Co-clusters

- Most previous co-clustering algorithms require the number of co-clusters as an input parameter
- The modularity measure can be used to predict the right number of co-clusters
- Run *Coclus* algorithm with different values of g (number of co-clusters)
- For each number of co-cluster the modularity is computed
- Retain the number of co-clusters for which the modularity measure reaches it's maximum value

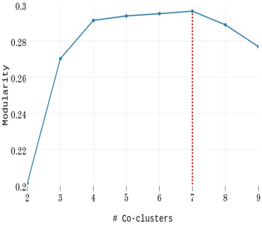
Assessing the right number of co-clusters



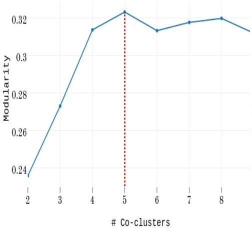
(a) CSTR (real : **4**, predicted : **4**)



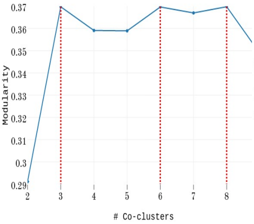
(b) CL4 (real : **4**, predicted : **4**)



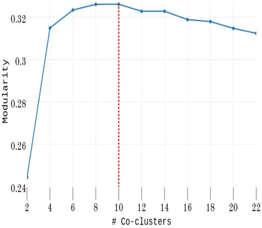
(c) SPORTS (real : **7**, predicted : **7**)



(d) Reviews (real : **5**, predicted : **5**)



(e) CL3 (real : **3**, predicted : **3**,6,8)



(f) NG20 (real : **20**, predicted : 10)

Outline

1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives

Motivation

- Investigate probabilistic mixture models allowing to make precise assumptions about the anatomy of diagonal co-clusters
- Flexibility
- Give rise to both soft and hard co-clustering

Motivation

- Investigate probabilistic mixture models allowing to make precise assumptions about the anatomy of diagonal co-clusters
- Flexibility
- Give rise to both soft and hard co-clustering

Contribution

- We present a sparse generative mixture model for co-clustering text data
- This model is based on the Poisson distribution, which arises naturally for contingency tables, such as document-term matrices
- The proposed model takes into account the sparsity in its formulation

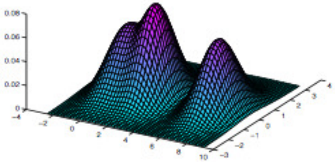
Model-based clustering - Finite mixture model

The matrix is assumed to be an i.i.d sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where $\mathbf{x}_i = (x_{i1}, \ldots x_{id}) \in \mathbb{R}^d$ is generated from a probability density function (pdf) with density :

$$f(\mathbf{x}_i, \theta) = \sum_{k=1}^{g} \pi_k f_k(\mathbf{x}_i, \alpha_k),$$

The likelihood of data $\mathbf{X}$ can be written as :

$$f(\mathbf{X}, \theta) = \prod_i \sum_{k=1}^{g} \pi_k f_k(\mathbf{x}_i, \alpha_k),$$

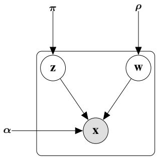


where

- $f_k(., \alpha_k)$ is the density of an observation $\mathbf{x}_i$ from the k-th component
- $\alpha_k' s$ are the corresponding class parameters
- π_k represents the proportions of each cluster.
- Each component k of the mixture represents a cluster.

Model-based co-clustering - Latent block model (LBM)

For each block $k\ell$, the values x_{ij} are generated according to a probability density function (pdf) $f(x_{ij};\alpha_{k\ell})$ (Govaert and Nadif, 2003)



Likelihood function

Denoting by $\mathcal{Z}$ and $\mathcal{W}$ the sets of all possible partitions, the likelihood function of a data matrix $\mathbf{X}$ of size $n \times d$ can be written

$$f(\mathbf{X};\boldsymbol{\theta}) = \sum_{(\mathbf{Z},\mathbf{W})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} f(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}},$$

Where

- $\boldsymbol{\theta} = (\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\alpha})$, is the parameters of the latent block model.
- $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ are the mixing proportions.
- $\boldsymbol{\alpha} = (\alpha_{k\ell}; k = 1,\ldots g, \ell = 1,\ldots,m)$ is the matrix of parameters of each block (k,ℓ).
- g (resp. m) represents the number of row (resp. column) clusters.

Latent block model (LBM)

Algorithm 2: Generative Process of LBM

Input : $n, d, g, m, \boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$

Output: data matrix $\mathbf{X}$, vector of row labellings $\mathbf{z} = (z_1, \ldots, z_n)$ and vector of column labellings $\mathbf{w} = (w_1, \ldots, w_d)$

for $i = 1$ **to** n **do**
- Generate the row label z_i according to the multinomial distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$

end

for $j = 1$ **to** d **do**
- Generate the column label w_j according to the multinomial distribution $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_g)$

end

for $i = 1$ **to** n **do**
  for $j = 1$ **to** d **do**
  - Generate a value x_{ij} according to the distribution $f(.; \alpha_{z_i, w_j})$

  end

end

Outline

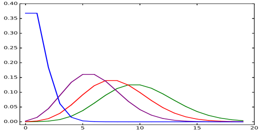
1 Introduction
- Context
- Co-clustering
- Motivations

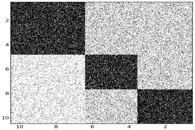
2 Graph-based Co-clustering
- Graph Modularity
- Modularity for Co-clustering
- Experiments

3 Model-based Co-clustering
- Sparse Poisson Latent Block Model (SPLBM)
- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
- Experiments

4 Using Co-clustering in Biomedical Text Mining Framework
- The Biomedical Framework
- Results and Discussions

5 Conclusion and Perspectives



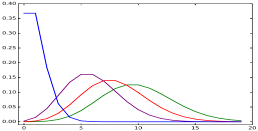


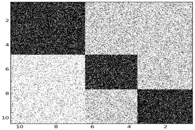
Intuition

- For each diagonal block kk the values x_{ij} are distributed according to the Poisson distribution $\mathcal{P}(\lambda_{ij})$ where the parameter λ_{ij} takes the following form :
$$\lambda_{ij} = x_{i.}x_{.j}\sum_k z_{ik}w_{jk}\gamma_{kk}.$$
- For each off-diagonal block $k\ell$ with $k \neq \ell$ the values x_{ij} are distributed according to the Poisson distribution $\mathcal{P}(\lambda_{ij})$ where the parameter λ_{ij} takes the following form :
$$\lambda_{ij} = x_{i.}x_{.j}\sum_{k,\ell\neq k} z_{ik}w_{j\ell}\gamma.$$
- Assuming $\forall \ell \neq k \quad \gamma_{k\ell} = \gamma$ leads to suppose that all blocks outside the diagonal share the same parameter.

Likelihood function

$$f(\mathbf{X};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,k} \rho_\ell^{w_{jk}} \times \prod_{i,j,k} (f(x_{ij};\alpha_{kk}))^{z_{ik}w_{jk}} \times \prod_{i,j,k,\ell\neq k} (f(x_{ij};\alpha_{k\ell}))^{z_{ik}w_{j\ell}}$$





Intuition

- For each diagonal block kk the values x_{ij} are distributed according to the Poisson distribution $\mathcal{P}(\lambda_{ij})$ where the parameter λ_{ij} takes the following form :
$$\lambda_{ij} = x_{i.}x_{.j}\sum_{k} z_{ik}w_{jk}\gamma_{kk}.$$
- For each off-diagonal block $k\ell$ with $k \neq \ell$ the values x_{ij} are distributed according to the Poisson distribution $\mathcal{P}(\lambda_{ij})$ where the parameter λ_{ij} takes the following form :
$$\lambda_{ij} = x_{i.}x_{.j}\sum_{k,\ell\neq k} z_{ik}w_{j\ell}\gamma.$$
- Assuming $\forall \ell \neq k \quad \gamma_{k\ell} = \gamma$ leads to suppose that all blocks outside the diagonal share the same parameter.

Likelihood function

$$\begin{aligned} f(\mathbf{X};\boldsymbol{\theta}) &= \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k}\pi_k^{z_{ik}}\prod_{j,k}\rho_\ell^{w_{jk}} \\ &\times \prod_{i,j,k}(f(x_{ij};\alpha_{kk}))^{z_{ik}w_{jk}} \times \prod_{i,j,k,\ell\neq k}(f(x_{ij};\alpha_{k\ell}))^{z_{ik}w_{j\ell}} \end{aligned}$$

Sparse Poisson Latent Block Model (SPLBM)

Complete Data Likelihood

$$f(\mathbf{X},\mathbf{Z},\mathbf{W};\boldsymbol{\theta}) = \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,k} \rho_k^{w_{jk}} \times \prod_{i,j,k} \left(\frac{e^{-x_{i.}x_{.j}\gamma_{kk}} (x_{i.}x_{.j}\gamma_{kk})^{x_{ij}}}{x_{ij}!} \right)^{z_{ik}w_{jk}}$$
$$\times \prod_{i,j,k,\ell \neq k} \left(\frac{e^{-x_{i.}x_{.j}\gamma} (x_{i.}x_{.j}\gamma)^{x_{ij}}}{x_{ij}!} \right)^{z_{ik}w_{j\ell}}$$

Complete Data Log-likelihood

$$L_C(\mathbf{Z},\mathbf{W},\boldsymbol{\theta}) = \log f(\mathbf{X},\mathbf{Z},\mathbf{W};\boldsymbol{\theta}) = \sum_{k=1}^{g} \mathcal{L}_C^k$$

$$\mathcal{L}_C^k = z_{.k}\log\pi_k + w_{.k}\log\rho_k + x_{kk}^{\mathbf{ZW}}\log\left(\frac{\gamma_{kk}}{\gamma}\right) - x_{k.}^{\mathbf{Z}}x_{.k}^{\mathbf{W}}(\gamma_{kk}-\gamma) + \frac{N}{g}(\log(\gamma) - \gamma N)$$

where $x_{kk}^{\mathbf{ZW}} = \sum_{ij} z_{ik}w_{jk}x_{ij}$, $z_{.k} = \sum_i z_{ik}$ and $w_{.k} = \sum_j w_{jk}$, $x_{k.}^{\mathbf{Z}} = \sum_i z_{ik}x_{i.}$ and $x_{.k}^{\mathbf{W}} = \sum_j w_{jk}x_{.j}$

Sparse Poisson Latent Block Model (SPLBM)

Complete Data Likelihood

$$f(\mathbf{X},\mathbf{Z},\mathbf{W};\boldsymbol{\theta}) = \prod_{i,k}\pi_k^{z_{ik}}\prod_{j,k}\rho_k^{w_{jk}} \quad\times\quad \prod_{i,j,k}\left(\frac{e^{-x_{i.}x_{.j}\gamma_{kk}}(x_{i.}x_{.j}\gamma_{kk})^{x_{ij}}}{x_{ij}!}\right)^{z_{ik}w_{jk}}$$

$$\times\quad \prod_{i,j,k,\ell\neq k}\left(\frac{e^{-x_{i.}x_{.j}\gamma}(x_{i.}x_{.j}\gamma)^{x_{ij}}}{x_{ij}!}\right)^{z_{ik}w_{j\ell}}$$

Complete Data Log-likelihood

$$\mathrm{L}_{\mathrm{C}}(\mathbf{Z},\mathbf{W},\boldsymbol{\theta}) = \log f(\mathbf{X},\mathbf{Z},\mathbf{W};\boldsymbol{\theta}) = \sum_{k=1}^{g}\mathcal{L}_C^k$$

$$\mathcal{L}_C^k = z_{.k}\log\pi_k + w_{.k}\log\rho_k + x_{kk}^{\mathbf{ZW}}\log\left(\frac{\gamma_{kk}}{\gamma}\right) - x_{k.}^{\mathbf{Z}}x_{.k}^{\mathbf{W}}(\gamma_{kk}-\gamma) + \frac{N}{g}(\log(\gamma) - \gamma N)$$

where $x_{kk}^{\mathbf{ZW}} = \sum_{ij} z_{ik}w_{jk}x_{ij}$, $z_{.k} = \sum_i z_{ik}$ and $w_{.k} = \sum_j w_{jk}$, $x_{k.}^{\mathbf{Z}} = \sum_i z_{ik}x_{i.}$ and $x_{.k}^{\mathbf{W}} = \sum_j w_{jk}x_{.j}$

Sparse Poisson Latent Block Model (SPLBM)

Complete Data Likelihood

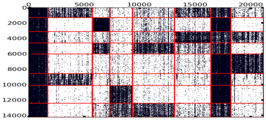
$$f(\mathbf{X},\mathbf{Z},\mathbf{W};\boldsymbol{\theta}) = \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,k} \rho_k^{w_{jk}} \times \prod_{i,j,k} \left(\frac{e^{-x_{i.}x_{.j}\gamma_{kk}} (x_{i.}x_{.j}\gamma_{kk})^{x_{ij}}}{x_{ij}!} \right)^{z_{ik}w_{jk}} \times \prod_{i,j,k,\ell \neq k} \left(\frac{e^{-x_{i.}x_{.j}\gamma} (x_{i.}x_{.j}\gamma)^{x_{ij}}}{x_{ij}!} \right)^{z_{ik}w_{j\ell}}$$

Complete Data Log-likelihood

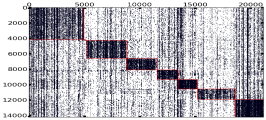
$$\mathrm{L_C}(\mathbf{Z},\mathbf{W},\boldsymbol{\theta}) = \log f(\mathbf{X},\mathbf{Z},\mathbf{W};\boldsymbol{\theta}) = \sum_{k=1}^{g} \mathcal{L}_C^k$$

$$\mathcal{L}_C^k = z_{.k}\log\pi_k + w_{.k}\log\rho_k + x_{kk}^{\mathbf{ZW}}\log\left(\frac{\gamma_{kk}}{\gamma}\right) - x_{k.}^{\mathbf{Z}}x_{.k}^{\mathbf{W}}(\gamma_{kk}-\gamma) + \frac{N}{g}(\log(\gamma) - \gamma N)$$

where $x_{kk}^{\mathbf{ZW}} = \sum_{ij} z_{ik}w_{jk}x_{ij}$, $z_{.k} = \sum_i z_{ik}$ and $w_{.k} = \sum_j w_{jk}$, $x_{k.}^{\mathbf{Z}} = \sum_i z_{ik}x_{i.}$ and $x_{.k}^{\mathbf{W}} = \sum_j w_{jk}x_{.j}$



(a) Traditional LBM - **64 parameters**



(b) Sparse PLBM - **9 parameters**

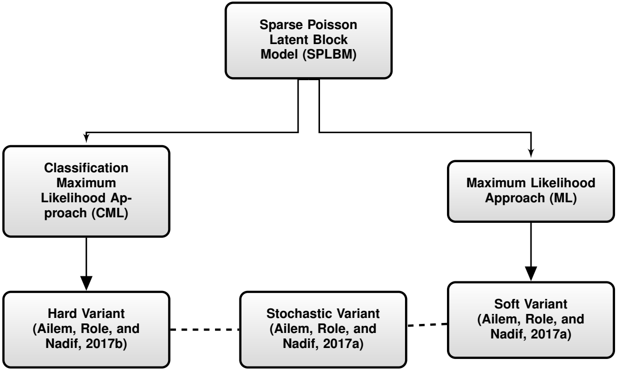


Figure: **SPLBM-based co-clustering algorithms**

M. Ailem, F. Role, and M. Nadif (2017b). "Sparse Poisson Latent Block Model for Document Clustering". In: *IEEE TKDE journal* 29.7, p. 1563.

M. Ailem, F. Role, and M. Nadif (2017a). "Model-based co-clustering for the effective handling of sparse data". In: *Pattern Recognition* 72, pp. 108–122.

Outline

1 Introduction
- Context
- Co-clustering
- Motivations

2 Graph-based Co-clustering
- Graph Modularity
- Modularity for Co-clustering
- Experiments

3 Model-based Co-clustering
- Sparse Poisson Latent Block Model (SPLBM)
- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
- Experiments

4 Using Co-clustering in Biomedical Text Mining Framework
- The Biomedical Framework
- Results and Discussions

5 Conclusion and Perspectives

Soft SPLBM-based Co-clustering Algorithm

- Estimate the model's parameters $\boldsymbol{\theta}$, $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{W}}$
- We rely on the Expectation-Maximization (EM) algorithm that consists in maximizing the expectation of the complete data likelihood $\mathrm{L_C}(\mathbf{Z},\mathbf{W},\boldsymbol{\theta})$ given by :

$$\begin{aligned}\mathbb{E}\left(\mathrm{L_C}(\mathbf{Z},\mathbf{W},\boldsymbol{\theta})|\boldsymbol{\theta}^{(t)},\mathbf{X}\right) &= \sum_{i,k}\tilde{z}_{ik}^{(t)}\log\pi_k+\sum_{j,k}\tilde{w}_{jk}^{(t)}\log\rho_k\\ &+ \sum_{i,j,k}\tilde{e}_{ijk}^{(t)}\left(x_{ij}\log(\gamma_{kk})-x_{i.}x_{.j}\gamma_{kk}\right)\\ &+ \sum_{i,j,k,\ell\neq k}\tilde{e}_{ikj\ell}^{(t)}\left(x_{ij}\log(\gamma)-x_{i.}x_{.j}\gamma\right),\end{aligned}$$

where $\tilde{z}_{ik}^{(t)}=\mathbb{E}(z_{ik}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})$, $\tilde{w}_{j\ell}=\mathbb{E}(w_{j\ell}^{(t)}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})$,
$\tilde{e}_{ikj\ell}^{(t)}=\mathbb{E}(e_{ikj\ell}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})=\mathbb{E}(z_{ik}w_{j\ell}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})$ and $\tilde{e}_{ijk}^{(t)}=\mathbb{E}(z_{ik}w_{jk}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})$.

Soft SPLBM-based Co-clustering Algorithm

- Estimate the model's parameters $\boldsymbol{\theta}$, $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{W}}$
- We rely on the Expectation-Maximization (EM) algorithm that consists in maximizing the expectation of the complete data likelihood $\mathrm{L_C}(\mathbf{Z},\mathbf{W},\boldsymbol{\theta})$ given by :

$$\begin{aligned}\mathbb{E}\left(\mathrm{L_C}(\mathbf{Z},\mathbf{W},\boldsymbol{\theta})|\boldsymbol{\theta}^{(t)},\mathbf{X}\right) &= \sum_{i,k}\tilde{z}_{ik}^{(t)}\log\pi_k+\sum_{j,k}\tilde{w}_{jk}^{(t)}\log\rho_k\\ &+ \sum_{i,j,k}\tilde{e}_{ijk}^{(t)}(x_{ij}\log(\gamma_{kk})-x_{i.}x_{.j}\gamma_{kk})\\ &+ \sum_{i,j,k,\ell\neq k}\tilde{e}_{ikj\ell}^{(t)}(x_{ij}\log(\gamma)-x_{i.}x_{.j}\gamma),\end{aligned}$$

where $\tilde{z}_{ik}^{(t)}=\mathbb{E}(z_{ik}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})$, $\tilde{w}_{j\ell}=\mathbb{E}(w_{j\ell}^{(t)}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})$,
$\tilde{e}_{ikj\ell}^{(t)}=\mathbb{E}(e_{ikj\ell}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})=\mathbb{E}(z_{ik}w_{j\ell}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})$ and $\tilde{e}_{ijk}^{(t)}=\mathbb{E}(z_{ik}w_{jk}=1|\mathbf{X},\boldsymbol{\theta}^{(t)})$.
The coupling of $\mathbf{Z}$ and $\mathbf{W}$ in e makes the direct application of the EM algorithm difficult, due to the determination of $\tilde{e}_{ijk}$ and $\tilde{e}_{ikj\ell}$

Model Fitting Using the Variational EM Algorithm

- Solution : Use a mean-field variational EM (VEM) algorithm for inferences
- The VEM algorithm is equivalent to maximizing the following soft co-clustering criteria:

$$F_C(\widetilde{\mathbf{Z}},\widetilde{\mathbf{W}},\boldsymbol{\theta}) = L_C(\widetilde{\mathbf{Z}},\widetilde{\mathbf{W}},\boldsymbol{\theta}) + H(\widetilde{\mathbf{Z}}) + H(\widetilde{\mathbf{W}}),$$

- where $H(\widetilde{\mathbf{Z}}) = -\sum_{i,k}\widetilde{z}_{ik}\log\widetilde{z}_{ik}$ and $H(\widetilde{\mathbf{W}}) = -\sum_{j,k}\widetilde{w}_{jk}\log\widetilde{w}_{jk}$ are respectively the entropy of the missing variables $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{W}}$
- $L_C(\widetilde{\mathbf{Z}},\widetilde{\mathbf{W}},\boldsymbol{\theta})$ is the soft complete data likelihood defined as follows :

$$\begin{aligned} L_C(\widetilde{\mathbf{Z}},\widetilde{\mathbf{W}},\boldsymbol{\theta}) &= \sum_{i,k}\widetilde{z}_{ik}\log\pi_k + \sum_{j,k}\widetilde{w}_{jk}\log\rho_k + \sum_{i,j,k}\widetilde{z}_{ik}\widetilde{w}_{jk}x_{ij}\log(\frac{\gamma_{kk}}{\gamma}) \\ &- \sum_k x_{k.}^{\widetilde{\mathbf{Z}}}x_{.k}^{\widetilde{\mathbf{W}}}\gamma_{kk} + \gamma\sum_k x_{k.}^{\widetilde{\mathbf{Z}}}x_{.k}^{\widetilde{\mathbf{W}}} + N(\log(\gamma) - \gamma N) \end{aligned}$$

- The SPLBvem algorithm consists of the expectation and maximization steps

Model Fitting Using the Variational EM Algorithm

M-step

- **Computation of $\hat{\gamma}_{kk}$ for all k.** It is easy to show that $\forall k$ the $\hat{\gamma}_{kk}$'s maximizing $\mathrm{F_C}$ can be computed separately for each k.

$$\hat{\gamma}_{kk} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}.$$

- **Computation of $\hat{\gamma}$ maximizing $\mathrm{F_C}$.** It is easy to show that $\hat{\gamma}$ is given by:

$$\hat{\gamma} = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}.$$

- **Computation of $\hat{\pi}_k$, $\hat{\rho}_k$ for all k.** Under the constraints $\sum_k \pi_k = \sum_k \rho_k = 1$, it is easy to show that each $\hat{\pi}_k$ and $\hat{\rho}_k$ maximizing $\mathrm{F_C}$ are respectively given by $\pi_k = \frac{\widetilde{\mathbf{Z}}_{.k}}{n}$ and $\rho_k = \frac{\widetilde{\mathbf{W}}_{.k}}{d}$.

Model Fitting Using the Variational EM Algorithm

E-step

- The E-step consists in computing the posterior probabilities $\widetilde{z}_{ik}$ and $\widetilde{w}_{jk}$ maximizing $\mathrm{F_C}$
- Plugging the estimation of γ_{kk}'s and γ (explicitly in some terms of $\mathrm{F_C}$) we obtain

$$\begin{aligned}\mathrm{F_C}(\widetilde{\mathbf{Z}},\widetilde{\mathbf{W}},\hat{\boldsymbol{\theta}}) &= \sum_{i,k}\widetilde{z}_{ik}\log\hat{\pi}_k + \sum_{j,k}\widetilde{\mathrm{w}}_{jk}\log\hat{\rho}_k + \sum_{i,j,k}\widetilde{z}_{ik}\widetilde{\mathrm{w}}_{jk}x_{ij}\log(\frac{\hat{\gamma}_{kk}}{\hat{\gamma}}) \\ &+ N(\log(\hat{\gamma})-1) - \sum_{i,k}\widetilde{z}_{ik}\log\widetilde{z}_{ik} - \sum_{j,k}\widetilde{w}_{jk}\log\widetilde{w}_{jk}.\end{aligned}$$

- Taking $x_{ik}^{\widetilde{\mathbf{W}}} = \sum_j \widetilde{\mathrm{w}}_{jk}x_{ij}$ and $x_{kj}^{\widetilde{\mathbf{Z}}} = \sum_i \widetilde{z}_{ik}x_{ij}$ it is easy to show that under the constraints:
  - $\sum_k \tilde{z}_{ik} = 1$
  - $\sum_k \tilde{w}_{jk} = 1$

$$\widetilde{z}_{ik} \propto \pi_k \exp(x_{ik}^{\widetilde{\mathbf{W}}}\log\frac{\gamma_{kk}}{\gamma}).$$

$$\widetilde{w}_{jk} \propto \rho_k \exp(x_{kj}^{\widetilde{\mathbf{Z}}}\log\frac{\gamma_{kk}}{\gamma}).$$

The `SPLBvem` Algorithm

Algorithm 3: `SPLBvem`

Input : $\mathbf{X}$, g

Initialization : $\widetilde{\mathbf{Z}}$, $\widetilde{\mathbf{W}}$, π_k, ρ_k, γ_{kk}, γ

repeat

$x^{\widetilde{\mathbf{W}}}_{ik} = \sum_j \widetilde{w}_{jk} x_{ij}$

step 1. $\widetilde{z}_{ik} \propto \pi_k \exp(x^{\widetilde{\mathbf{W}}}_{ik} \log \frac{\gamma_{kk}}{\gamma})$

step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{kk} = \frac{\sum_i \widetilde{z}_{ik} x^{\widetilde{\mathbf{W}}}_{ik}}{x^{\widetilde{\mathbf{Z}}}_{k.} x^{\widetilde{\mathbf{W}}}_{.k}} = \frac{x^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}_{kk}}{x^{\widetilde{\mathbf{Z}}}_{k.} x^{\widetilde{\mathbf{W}}}_{.k}}$ $\quad, \gamma = \frac{N - \sum_k x^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}_{kk}}{N^2 - \sum_k x^{\widetilde{\mathbf{Z}}}_{k.} x^{\widetilde{\mathbf{W}}}_{.k}}$

$x^{\widetilde{\mathbf{Z}}}_{kj} = \sum_i \widetilde{z}_{ik} x_{ij}$

step 3. $\widetilde{w}_{jk} \propto \rho_k \exp(x^{\widetilde{\mathbf{Z}}}_{kj} \log \frac{\gamma_{kk}}{\gamma})$

step 4. $\rho_k = \frac{\widetilde{w}_{.k}}{d}$, $\gamma_{kk} = \frac{\sum_j \widetilde{w}_{jk} x^{\widetilde{\mathbf{Z}}}_{kj}}{x^{\widetilde{\mathbf{Z}}}_{k.} x^{\widetilde{\mathbf{W}}}_{.k}} = \frac{x^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}_{kk}}{x^{\widetilde{\mathbf{Z}}}_{k.} x^{\widetilde{\mathbf{W}}}_{.k}}$ $\quad, \gamma = \frac{N - \sum_k x^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}_{kk}}{N^2 - \sum_k x^{\widetilde{\mathbf{Z}}}_{k.} x^{\widetilde{\mathbf{W}}}_{.k}}$

until *Convergence*;

Output : $\widetilde{\mathbf{Z}}$, $\widetilde{\mathbf{W}}$, π_k, ρ_k, γ_{kk}, γ

Outline

1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives

The hard SPLBM-based co-clustering algorithm (SPLBcem)

Intuition

- It consists in maximizing the classification likelihood instead of its expectation
- This is done by incorporating a classification step (C-step) between the E and M steps of the SPLBvem

Algorithm 4: `SPLBcem`

Input : $\mathbf{X}$, g
Initialization : $\mathbf{Z}$, $\mathbf{W}$, π_k, ρ_k,γ_{kk}, γ
repeat

$x_{ik}^{\widetilde{\mathbf{W}}} = \sum_j \widetilde{w}_{jk} x_{ij}$

step 1. $\widetilde{z}_{ik} \propto \pi_k \exp(x_{ik}^{\widetilde{\mathbf{W}}} \log \frac{\gamma_{kk}}{\gamma})$

step 1'. $z_{ik} = \arg\max_k \widetilde{z}_{ik}$

step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{kk} = \frac{\sum_i \widetilde{z}_{ik} x_{ik}^{\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

$x_{kj}^{\widetilde{\mathbf{Z}}} = \sum_i \widetilde{z}_{ik} x_{ij}$

step 3. $\widetilde{w}_{jk} \propto \rho_k \exp(x_{kj}^{\widetilde{\mathbf{Z}}} \log \frac{\gamma_{kk}}{\gamma})$

step 3'. $w_{jk} = \arg\max_k \widetilde{w}_{jk}$

step 4. $\rho_k = \frac{\widetilde{w}_{.k}}{d}$, $\gamma_{kk} = \frac{\sum_j \widetilde{w}_{jk} x_{kj}^{\widetilde{\mathbf{Z}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

until *Convergence*;
Output : $\mathbf{Z}$, $\mathbf{W}$, π_k, ρ_k, γ_{kk}, γ

Advantages

- SPLBcem is considerably faster and scalable than SPLBvem
- It allows us to avoid numerical difficulties, related to the computation of the posterior probabilities $\widetilde{z}_{ik}$ and $\widetilde{w}_{jk}$

The hard SPLBM-based co-clustering algorithm (SPLBcem)

Intuition

- It consists in maximizing the classification likelihood instead of its expectation
- This is done by incorporating a classification step (C-step) between the E and M steps of the SPLBvem

Algorithm 4: `SPLBcem`

Input : $\mathbf{X}$, g
Initialization : $\mathbf{Z}$, $\mathbf{W}$, π_k, ρ_k,γ_{kk}, γ
repeat

$x_{ik}^{\widetilde{\mathbf{W}}} = \sum_j \widetilde{w}_{jk} x_{ij}$

step 1. $\widetilde{z}_{ik} \propto \pi_k \exp(x_{ik}^{\widetilde{\mathbf{W}}} \log \frac{\gamma_{kk}}{\gamma})$

step 1'. $z_{ik} = \arg\max_k \tilde{z}_{ik}$

step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{kk} = \frac{\sum_i \widetilde{z}_{ik} x_{ik}^{\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

$x_{kj}^{\widetilde{\mathbf{Z}}} = \sum_i \widetilde{z}_{ik} x_{ij}$

step 3. $\widetilde{w}_{jk} \propto \rho_k \exp(x_{kj}^{\widetilde{\mathbf{Z}}} \log \frac{\gamma_{kk}}{\gamma})$

step 3'. $w_{jk} = \arg\max_k \tilde{w}_{jk}$

step 4. $\rho_k = \frac{\widetilde{w}_{.k}}{d}$, $\gamma_{kk} = \frac{\sum_j \widetilde{w}_{jk} x_{kj}^{\widetilde{\mathbf{Z}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

until *Convergence*;
Output : $\mathbf{Z}$, $\mathbf{W}$, π_k, ρ_k, γ_{kk}, γ

Advantages

- SPLBcem is considerably faster and scalable than SPLBvem
- It allows us to avoid numerical difficulties, related to the computation of the posterior probabilities $\tilde{z}_{ik}$ and $\tilde{w}_{jk}$

The stochastic SPLBM-based co-clustering algorithm (SPLBsem)

`SPLBvem` and `SPLBcem` are very dependant on their starting points!

Algorithm 5: `SPLBsem`

Input : $\mathbf{X}, g$
Initialization : $\widetilde{\mathbf{Z}}, \widetilde{\mathbf{W}}, \pi_k, \rho_k, \gamma_{kk}, \gamma$
repeat

$x_{ik}^{\widetilde{\mathbf{W}}} = \sum_j \widetilde{w}_{jk} x_{ij}$

step 1. $\widetilde{z}_{ik} \propto \pi_k \exp(x_{ik}^{\widetilde{\mathbf{W}}} \log \frac{\gamma_{kk}}{\gamma})$

step 1'. simulation of z_i according to $\mathcal{M}(\widetilde{z}_{i1}, \ldots, \widetilde{z}_{ig})$

step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}, \gamma_{kk} = \frac{\sum_i \widetilde{z}_{ik} x_{ik}^{\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}, \gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

$x_{kj}^{\widetilde{\mathbf{Z}}} = \sum_i \widetilde{z}_{ik} x_{ij}$

step 3. $\widetilde{w}_{jk} \propto \rho_k \exp(x_{kj}^{\widetilde{\mathbf{Z}}} \log \frac{\gamma_{kk}}{\gamma})$

step 3'. simulation of w_j according to $\mathcal{M}(\widetilde{w}_{j1}, \ldots, \widetilde{w}_{jg})$

step 4. $\rho_k = \frac{\widetilde{w}_{.k}}{d}, \gamma_{kk} = \frac{\sum_j \widetilde{w}_{jk} x_{kj}^{\widetilde{\mathbf{Z}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}, \gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

until *Convergence*;
Output : $\mathbf{Z}, \mathbf{W}, \pi_k, \rho_k, \gamma_{kk}, \gamma$

The stochastic SPLBM-based co-clustering algorithm (SPLBsem)

Algorithm 5: `SPLBsem`

Input : $\mathbf{X}$, g
Initialization : $\widetilde{\mathbf{Z}}$, $\widetilde{\mathbf{W}}$, π_k, ρ_k,γ_{kk}, γ
repeat

$x_{ik}^{\widetilde{\mathbf{W}}} = \sum_j \widetilde{w}_{jk} x_{ij}$

step 1. $\widetilde{z}_{ik} \propto \pi_k \exp(x_{ik}^{\widetilde{\mathbf{W}}} \log \frac{\gamma_{kk}}{\gamma})$

step 1'. simulation of z_i according to $\mathcal{M}(\widetilde{z}_{i1}, \ldots, \widetilde{z}_{ig})$

step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{kk} = \frac{\sum_i \widetilde{z}_{ik} x_{ik}^{\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

$x_{kj}^{\widetilde{\mathbf{Z}}} = \sum_i \widetilde{z}_{ik} x_{ij}$

step 3. $\widetilde{w}_{jk} \propto \rho_k \exp(x_{kj}^{\widetilde{\mathbf{Z}}} \log \frac{\gamma_{kk}}{\gamma})$

step 3'. simulation of w_j according to $\mathcal{M}(\widetilde{w}_{j1}, \ldots, \widetilde{w}_{jg})$

step 4. $\rho_k = \frac{\widetilde{w}_{.k}}{d}$, $\gamma_{kk} = \frac{\sum_j \widetilde{w}_{jk} x_{kj}^{\widetilde{\mathbf{Z}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

until *Convergence*;
Output : $\mathbf{Z}$, $\mathbf{W}$, π_k, ρ_k, γ_{kk}, γ

The stochastic SPLBM-based co-clustering algorithm (SPLBsem)

Algorithm 5: `SPLBsem`

Input : $\mathbf{X}$, g
Initialization : $\widetilde{\mathbf{Z}}$, $\widetilde{\mathbf{W}}$, π_k, ρ_k, γ_{kk}, γ
repeat

$x_{ik}^{\widetilde{\mathbf{W}}} = \sum_j \widetilde{w}_{jk} x_{ij}$

step 1. $\widetilde{z}_{ik} \propto \pi_k \exp(x_{ik}^{\widetilde{\mathbf{W}}} \log \frac{\gamma_{kk}}{\gamma})$

step 1'. simulation of z_i according to $\mathcal{M}(\widetilde{z}_{i1}, \ldots, \widetilde{z}_{ig})$

step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{kk} = \frac{\sum_i \widetilde{z}_{ik} x_{ik}^{\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

$x_{kj}^{\widetilde{\mathbf{Z}}} = \sum_i \widetilde{z}_{ik} x_{ij}$

step 3. $\widetilde{w}_{jk} \propto \rho_k \exp(x_{kj}^{\widetilde{\mathbf{Z}}} \log \frac{\gamma_{kk}}{\gamma})$

step 3'. simulation of w_j according to $\mathcal{M}(\widetilde{w}_{j1}, \ldots, \widetilde{w}_{jg})$

step 4. $\rho_k = \frac{\widetilde{w}_{.k}}{d}$, $\gamma_{kk} = \frac{\sum_j \widetilde{w}_{jk} x_{kj}^{\widetilde{\mathbf{Z}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

until *Convergence*;
Output : $\mathbf{Z}$, $\mathbf{W}$, π_k, ρ_k, γ_{kk}, γ

Advantages : It does not stop at the first stationary point of the likelihood function, which makes it possible to avoid bad local maxima due to the initial position

The stochastic SPLBM-based co-clustering algorithm (SPLBsem)

Algorithm 5: `SPLBsem`

Input : $\mathbf{X}$, g
Initialization : $\widetilde{\mathbf{Z}}$, $\widetilde{\mathbf{W}}$, π_k, ρ_k,γ_{kk}, γ
repeat

$x_{ik}^{\widetilde{\mathbf{W}}} = \sum_j \widetilde{w}_{jk} x_{ij}$

step 1. $\widetilde{z}_{ik} \propto \pi_k \exp(x_{ik}^{\widetilde{\mathbf{W}}} \log \frac{\gamma_{kk}}{\gamma})$

step 1'. simulation of z_i according to $\mathcal{M}(\widetilde{z}_{i1}, \ldots, \widetilde{z}_{ig})$

step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{kk} = \frac{\sum_i \widetilde{z}_{ik} x_{ik}^{\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

$x_{kj}^{\widetilde{\mathbf{Z}}} = \sum_i \widetilde{z}_{ik} x_{ij}$

step 3. $\widetilde{w}_{jk} \propto \rho_k \exp(x_{kj}^{\widetilde{\mathbf{Z}}} \log \frac{\gamma_{kk}}{\gamma})$

step 3'. simulation of w_j according to $\mathcal{M}(\widetilde{w}_{j1}, \ldots, \widetilde{w}_{jg})$

step 4. $\rho_k = \frac{\widetilde{w}_{.k}}{d}$, $\gamma_{kk} = \frac{\sum_j \widetilde{w}_{jk} x_{kj}^{\widetilde{\mathbf{Z}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

until *Convergence*;
Output : $\mathbf{Z}$, $\mathbf{W}$, π_k, ρ_k, γ_{kk}, γ

Advantages : It does not stop at the first stationary point of the likelihood function, which makes it possible to avoid bad local maxima due to the initial position
Weakness : `SPLBsem` does not share the convergence properties of `SPLBvem` and `SPLBcem` and may require a large number of iterations to reach a steady state

The stochastic SPLBM-based co-clustering algorithm (SPLBsem)

Algorithm 5: `SPLBsem`

Input : $\mathbf{X}$, g

Initialization : $\widetilde{\mathbf{Z}}$, $\widetilde{\mathbf{W}}$, π_k, ρ_k,γ_{kk}, γ

repeat

$x_{ik}^{\widetilde{\mathbf{W}}} = \sum_j \widetilde{w}_{jk} x_{ij}$

step 1. $\widetilde{z}_{ik} \propto \pi_k \exp(x_{ik}^{\widetilde{\mathbf{W}}} \log \frac{\gamma_{kk}}{\gamma})$

step 1'. simulation of z_i according to $\mathcal{M}(\widetilde{z}_{i1}, \ldots, \widetilde{z}_{ig})$

step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{kk} = \frac{\sum_i \widetilde{z}_{ik} x_{ik}^{\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

$x_{kj}^{\widetilde{\mathbf{Z}}} = \sum_i \widetilde{z}_{ik} x_{ij}$

step 3. $\widetilde{w}_{jk} \propto \rho_k \exp(x_{kj}^{\widetilde{\mathbf{Z}}} \log \frac{\gamma_{kk}}{\gamma})$

step 3'. simulation of w_j according to $\mathcal{M}(\widetilde{w}_{j1}, \ldots, \widetilde{w}_{jg})$

step 4. $\rho_k = \frac{\widetilde{w}_{.k}}{d}$, $\gamma_{kk} = \frac{\sum_j \widetilde{w}_{jk} x_{kj}^{\widetilde{\mathbf{Z}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}} = \frac{x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$, $\gamma = \frac{N - \sum_k x_{kk}^{\widetilde{\mathbf{Z}}\widetilde{\mathbf{W}}}}{N^2 - \sum_k x_{k.}^{\widetilde{\mathbf{Z}}} x_{.k}^{\widetilde{\mathbf{W}}}}$

until *Convergence*;

Output : $\mathbf{Z}$, $\mathbf{W}$, π_k, ρ_k, γ_{kk}, γ

Advantages : It does not stop at the first stationary point of the likelihood function, which makes it possible to avoid bad local maxima due to the initial position

Weakness : `SPLBsem` does not share the convergence properties of `SPLBvem` and `SPLBcem` and may require a large number of iterations to reach a steady state

- Solution ⇒ initialize `SPLBvem` with the parameters resulting from `SPLBsem` ⇒ `SPLBsvem`

Outline

1 Introduction
- Context
- Co-clustering
- Motivations

2 Graph-based Co-clustering
- Graph Modularity
- Modularity for Co-clustering
- Experiments

3 Model-based Co-clustering
- Sparse Poisson Latent Block Model (SPLBM)
- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
- Experiments

4 Using Co-clustering in Biomedical Text Mining Framework
- The Biomedical Framework
- Results and Discussions

5 Conclusion and Perspectives

Global Performance Comparison - Document Clustering

Datasets	Characteristics				
	#Documents	#Words	#Clusters	Sparsity (%)	Balance
SPORTS	8580	14870	7	99.14	0.036
TDT2	9394	36771	30	99.64	0.028
Yahoo_K1B	2340	21839	6	99.41	0.043
Reuters40	8203	18914	40	99.75	0.003

- Data : contingency tables
- Evaluation measures : Acc, NMI (Strehl and Ghosh, 2003) and ARI (Rand, 1971)

Comparative study

- Proposed diagonal co-clustering : Coclus, SPLBcem, SPLBvem, SPLBsem, SPLBsvem
- Non-diagonal co-clustering : ITCC (I. S. Dhillon, Mallela, and D. S. Modha, 2003), PLBvem (Govaert and Nadif, 2010) and LDA (Blei, Ng, and Jordan, 2003)
- Clustering : Spherical kmeans (I. Dhillon and D. Modha, 2001)

Global Performance Comparison - Document Clustering

Datasets	Characteristics				
	#Documents	#Words	#Clusters	Sparsity (%)	Balance
SPORTS	8580	14870	7	99.14	0.036
TDT2	9394	36771	30	99.64	0.028
Yahoo_K1B	2340	21839	6	99.41	0.043
Reuters40	8203	18914	40	99.75	0.003

- Data : contingency tables
- Evaluation measures : Acc, NMI (Strehl and Ghosh, 2003) and ARI (Rand, 1971)

Comparative study

- Proposed diagonal co-clustering : Coclus, SPLBcem, SPLBvem, SPLBsem, SPLBsvem
- Non-diagonal co-clustering : ITCC (I. S. Dhillon, Mallela, and D. S. Modha, 2003), PLBvem (Govaert and Nadif, 2010) and LDA (Blei, Ng, and Jordan, 2003)
- Clustering : Spherical kmeans (I. Dhillon and D. Modha, 2001)

datasets	per.	Skmeans	ITCC	LDA	PLBvem	CoClus	SPLBcem	SPLBvem	SPLBsem	SPLBsvem
SPORTS	Acc	0.49	0.53	0.53	0.47	0.75	0.85	0.85	**0.86**	0.81
	NMI	0.50	0.60	0.54	0.64	0.62	0.69	0.70	**0.71**	0.67
	ARI	0.30	0.44	0.33	0.49	0.55	0.76	0.75	**0.77**	0.69
TDT2	Acc	0.57	0.59	0.60	0.59	**0.87**	0.83	0.84	0.84	0.85
	NMI	0.76	0.78	0.73	0.76	**0.84**	0.81	0.82	**0.84**	**0.84**
	ARI	0.46	0.52	0.49	0.51	**0.85**	0.81	0.80	**0.85**	**0.85**
Yahoo_K1B	Acc	0.57	0.61	0.62	0.58	0.60	0.79	0.84	0.86	**0.88**
	NMI	0.64	0.58	0.58	0.62	0.54	0.66	0.69	0.72	**0.75**
	ARI	0.39	0.40	0.37	0.38	0.31	0.60	0.72	0.76	**0.79**
REUTERS40	Acc	0.26	0.27	0.47	0.25	0.61	0.73	0.74	0.73	**0.77**
	NMI	0.50	0.52	0.51	0.52	0.54	0.57	0.58	0.57	**0.62**
	ARI	0.11	0.18	0.42	0.15	0.51	0.71	0.75	0.73	**0.76**

- Diagonal co-clustering are better in almost all situations
- In particular the SPLBsvem which leverages the benefits of both soft and stochastic variants

Global Performance Comparison - Document Clustering

datasets	per.	Skmeans	ITCC	LDA	PLBvem	CoClus	SPLBcem	SPLBvem	SPLBsem	SPLBsvem
SPORTS	Acc	0.49	0.53	0.53	0.47	0.75	0.85	0.85	**0.86**	0.81
	NMI	0.50	0.60	0.54	0.64	0.62	0.69	0.70	**0.71**	0.67
	ARI	0.30	0.44	0.33	0.49	0.55	0.76	0.75	**0.77**	0.69
TDT2	Acc	0.57	0.59	0.60	0.59	**0.87**	0.83	0.84	0.84	0.85
	NMI	0.76	0.78	0.73	0.76	**0.84**	0.81	0.82	**0.84**	**0.84**
	ARI	0.46	0.52	0.49	0.51	**0.85**	0.81	0.80	**0.85**	**0.85**
Yahoo_K1B	Acc	0.57	0.61	0.62	0.58	0.60	0.79	0.84	0.86	**0.88**
	NMI	0.64	0.58	0.58	0.62	0.54	0.66	0.69	0.72	**0.75**
	ARI	0.39	0.40	0.37	0.38	0.31	0.60	0.72	0.76	**0.79**
REUTERS40	Acc	0.26	0.27	0.47	0.25	0.61	0.73	0.74	0.73	**0.77**
	NMI	0.50	0.52	0.51	0.52	0.54	0.57	0.58	0.57	**0.62**
	ARI	0.11	0.18	0.42	0.15	0.51	0.71	0.75	0.73	**0.76**

- Diagonal co-clustering are better in almost all situations
- In particular the SPLBsvem which leverages the benefits of both soft and stochastic variants

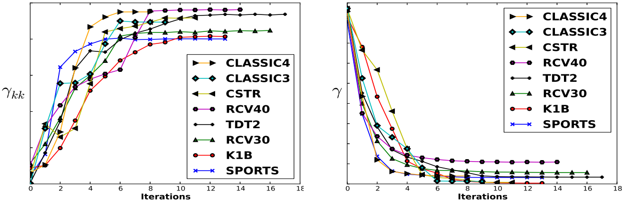
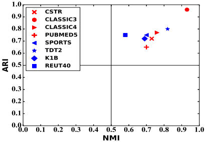


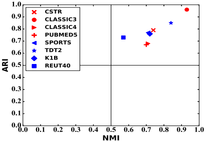
Figure: Behaviour of the γ_{kk}'s (left) and γ (right) parameters at each iteration.

- The proposed diagonal approaches deal well with unbalanced datasets

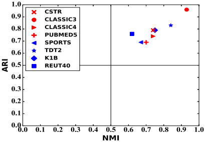
- The proposed diagonal approaches deal well with unbalanced datasets
- The diagonal approaches reach good performance in both NMI and ARI on unbalanced datasets
- ARI, unlike NMI, is more sensitive to cluster merging/splitting



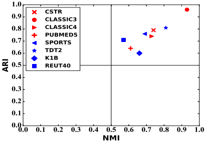
(a) **SPLBvem**



(b) **SPLBsem**

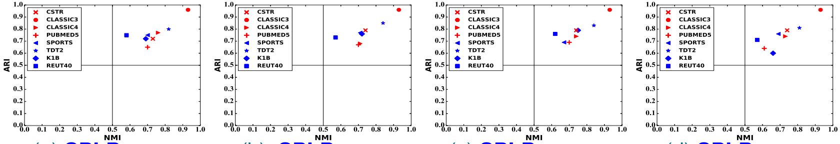


(c) **SPLBsvem**

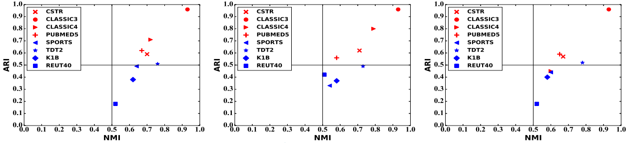


(d) **SPLBcem**

- The proposed diagonal approaches deal well with unbalanced datasets
- The diagonal approaches reach good performance in both NMI and ARI on unbalanced datasets
- ARI, unlike NMI, is more sensitive to cluster merging/splitting

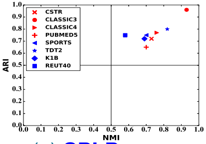


(a) **SPLBvem** (b) **SPLBsem** (c) **SPLBsvem** (d) **SPLBcem**

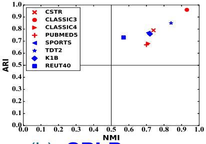


(e) `PLBvem` (f) `LDA` (g) `ITCC`

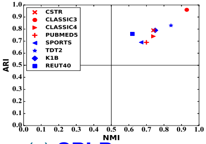
- The proposed diagonal approaches deal well with unbalanced datasets
- The diagonal approaches reach good performance in both NMI and ARI on unbalanced datasets
- ARI, unlike NMI, is more sensitive to cluster merging/splitting



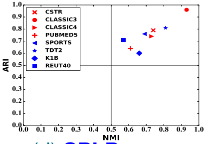
(a) **SPLBvem**



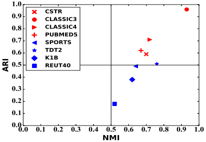
(b) **SPLBsem**



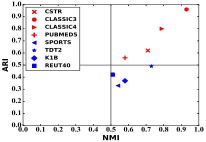
(c) **SPLBsvem**



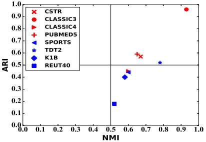
(d) **SPLBcem**



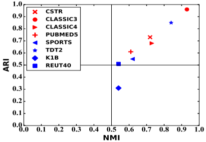
(e) `PLBvem`



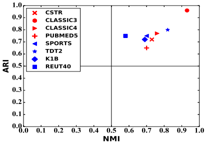
(f) `LDA`



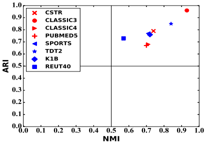
(g) `ITCC`



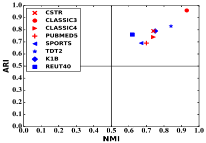
(h) **CoClus**



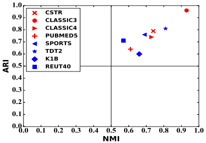
(a) **SPLBvem**



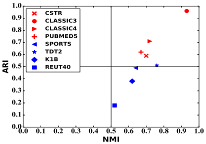
(b) **SPLBsem**



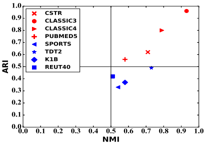
(c) **SPLBsvem**



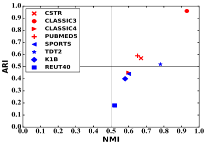
(d) **SPLBcem**



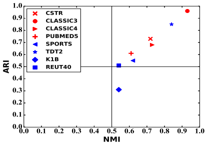
(e) `PLBvem`



(f) `LDA`



(g) `ITCC`



(h) **CoClus**

- Comparison of the standard deviation in cluster size (SDCS) of clusters obtained by each method
  SDCS = $\left(\frac{1}{g-1}\sum_{k=1}^{g}(z_{.k} - \frac{n}{g})^2\right)^{0.5}$
- The SDCS values of the clusters obtained with `SPLBcem` are the closest to the real SDCS of the datasets

Data	Clustering	Co-clustering				Real SDCS
		Non-diagonal			Diagonal	
	`Skmeans`	`ITCC`	`LDA`	`PLBcem`	`SPLBcem`	
REUTERS40	112.638	144.195	362.102	201.162	**642.839**	**654.556**
REUTERS30	161.797	238.353	414.568	261.291	**752.129**	**747.879**
K1B	154.3684	198.828	261.765	189.849	**336.555**	**513.303**
TDT2	154.143	216.152	189.609	235.698	**516.685**	**481.830**
SPORTS	760.099	346.066	482.714	393.510	**1359.321**	**1253.011**

Assessing the Quality of Term Clusters

- Lack of benchmark datasets providing the true cluster labels of both the objects and attributes.
- Most studies evaluate the co-clustering algorithms based on the object (document) clustering only.
- We propose two different approaches to evaluate term clusters :
  - Visual assessment of term cluster coherence
  - Quantitative evaluation of term cluster quality
- We use a biomedical document-term matrix, namely the PUBMED5 dataset.
- PUBMED5 dataset is a document-term matrix of size 12648×19518 that contains documents about 5 different diseases.

Disease	Number of documents
Migraine	3703
Age-related Macular Degeneration	3283
Otitis	2596
Kidney Calculi	1549
Hay Fever	1517

Assessing the Quality of Term Clusters

- Lack of benchmark datasets providing the true cluster labels of both the objects and attributes.
- Most studies evaluate the co-clustering algorithms based on the object (document) clustering only.
- We propose two different approaches to evaluate term clusters :
  - Visual assessment of term cluster coherence
  - Quantitative evaluation of term cluster quality
- We use a biomedical document-term matrix, namely the PUBMED5 dataset.
- PUBMED5 dataset is a document-term matrix of size 12648×19518 that contains documents about 5 different diseases.

Disease	Number of documents
Migraine	3703
Age-related Macular Degeneration	3283
Otitis	2596
Kidney Calculi	1549
Hay Fever	1517

Assessing the Quality of Term Clusters

- Lack of benchmark datasets providing the true cluster labels of both the objects and attributes.
- Most studies evaluate the co-clustering algorithms based on the object (document) clustering only.
- We propose two different approaches to evaluate term clusters :
  - Visual assessment of term cluster coherence
  - Quantitative evaluation of term cluster quality
- We use a biomedical document-term matrix, namely the PUBMED5 dataset.
- PUBMED5 dataset is a document-term matrix of size 12648×19518 that contains documents about 5 different diseases.

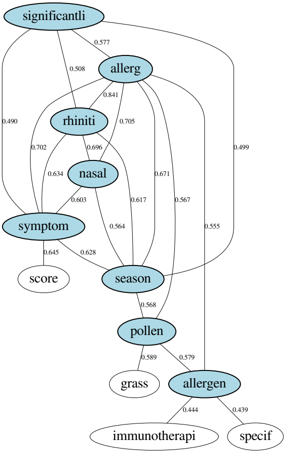
Disease	Number of documents
Migraine	3703
Age-related Macular Degeneration	3283
Otitis	2596
Kidney Calculi	1549
Hay Fever	1517

Visual assessment of term cluster coherence

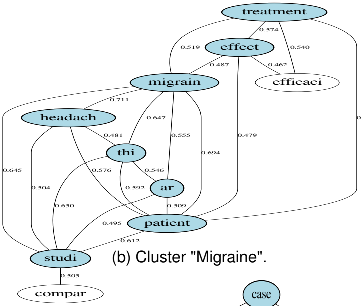
Assess if the top terms present in a co-cluster are densely interconnected and form a semantically coherent set.

Principle

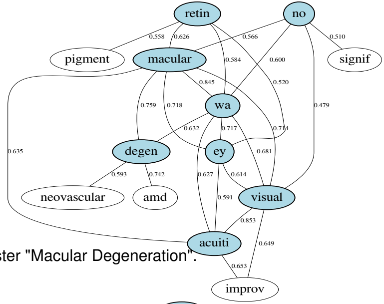
1. Co-clustering with `SPLBcem` on the PUBMED5 dataset into $g = 5$ blocks
2. For each diagonal block c, we extract the corresponding matrix X_c
3. Build a term-term cosine similarity matrix $S_c = X_c^{norm\prime} X_c^{norm}$ for each diagonal block
4. Place the $n = 8$ top terms of c in a graph
5. Connect each top word their $k = 5$ most similar neighbors according to the cosine similarity recorded in S_c



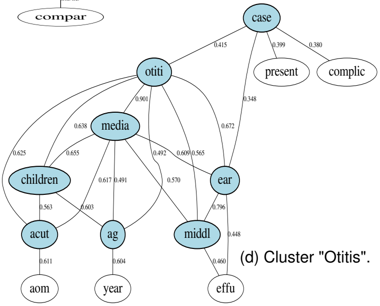
(a) Cluster "Hay fever".



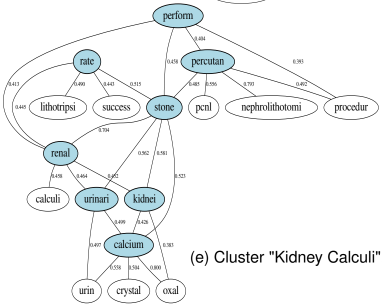
(b) Cluster "Migraine".



(c) Cluster "Macular Degeneration".



(d) Cluster "Otitis".



(e) Cluster "Kidney Calculi".

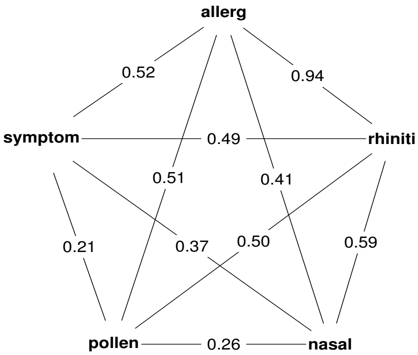
Quantitative evaluation of term cluster quality

Principle

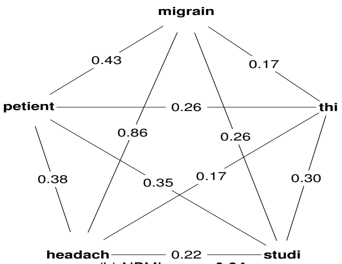
- Use the Point-wise Mutual Information (PMI) to measure the degree of association between word pairs

$$\text{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

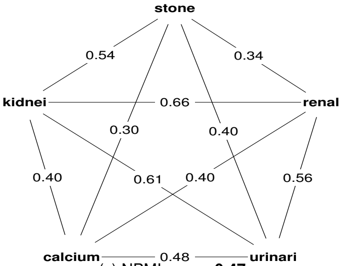
- PMI can be estimated using an external corpus
- Use the whole English WIKIPEDIA corpus that consists of approximately 4 millions of documents and 2 billions of words
- The $\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log(p(w_i, w_j))}$ ranges between -1 and +1, the higher the NPMI, the greater the correlation between words w_i and w_j



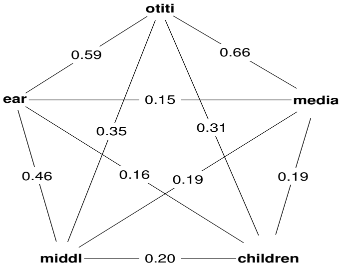
(a) NPMI score: **0.48**.



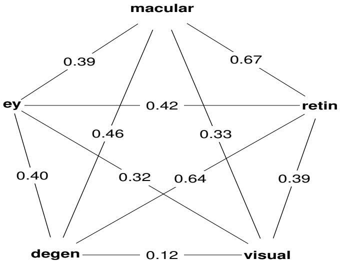
(b) NPMI score: **0.34**.



(c) NPMI score: **0.47**.



(d) NPMI score: **0.33**.



(e) NPMI score: **0.41**.

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
    - A hard variant SPLBcem
    - A soft variant SPLBvem
    - A stochastic variant SPLBsem
- Extensive numerical experiments show that
    - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
    - Reduce the computational time
    - Robust against highly unbalanced datasets
    - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
    - A hard variant SPLBcem
    - A soft variant SPLBvem
    - A stochastic variant SPLBsem
- Extensive numerical experiments show that
    - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
    - Reduce the computational time
    - Robust against highly unbalanced datasets
    - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Concluding remarks

- Diagonal co-clustering algorithm (Coclus) by direct maximization of graph modularity
- Coclus is able to effectively co-cluster different kind of positive document-term matrices
- Sparse Poisson Latent Block Model (SPLBM)
- SPLBM is also very parsimonious
- SPLBM has been designed from the ground up to deal with data sparsity problems
- From this model, three co-clustering algorithms have been inferred
  - A hard variant SPLBcem
  - A soft variant SPLBvem
  - A stochastic variant SPLBsem
- Extensive numerical experiments show that
  - Seeking diagonal structure is more effective when dealing with high dimensional sparse data
  - Reduce the computational time
  - Robust against highly unbalanced datasets
  - Discover pure and well separated document/word clusters

Outline

1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives

Context

- Exponential growth of biomedical text data (PUBMED, GO, . . .)
- There is a genuine need for text mining techniques to analyse and interpret these large amounts of information
- Help researchers to characterize relationships between biomedical entities (genes, diseases, . . .) quickly and efficiently

Motivations

- Genome-wide association studies (GWAS) : examination of many genetic variants (SNPs) in different individuals to study their correlations with phenotypic traits
- GWAS allow to identify groups of genes associated with a common phenotype
- GWAS do not provide information about associations in these gene groups

Contributions

- A biomedical text mining framework (Ailem et al., 2016) to augment the results of GWAS
- Benefits of co-clustering in biomedical text mining application
- Illustration on GWAS of asthma disease (Moffatt et al., 2010), which reported 10 genes associated with asthma
- Assess the strength of association between these genes and infer new candidate genes likely associated with asthma

---

M. Ailem et al. (2016). “Unsupervised text mining for assessing and augmenting GWAS results”. In: *Journal of biomedical informatics* 60, pp. 252–259.

Context

- Exponential growth of biomedical text data (PUBMED, GO, . . .)
- There is a genuine need for text mining techniques to analyse and interpret these large amounts of information
- Help researchers to characterize relationships between biomedical entities (genes, diseases, . . .) quickly and efficiently

Motivations

- Genome-wide association studies (GWAS) : examination of many genetic variants (SNPs) in different individuals to study their correlations with phenotypic traits
- GWAS allow to identify groups of genes associated with a common phenotype
- GWAS do not provide information about associations in these gene groups

Contributions

- A biomedical text mining framework (Ailem et al., 2016) to augment the results of GWAS
- Benefits of co-clustering in biomedical text mining application
- Illustration on GWAS of asthma disease (Moffatt et al., 2010), which reported 10 genes associated with asthma
- Assess the strength of association between these genes and infer new candidate genes likely associated with asthma

M. Ailem et al. (2016). “Unsupervised text mining for assessing and augmenting GWAS results”. In: *Journal of biomedical informatics* 60, pp. 252–259.

Context

- Exponential growth of biomedical text data (PUBMED, GO, . . .)
- There is a genuine need for text mining techniques to analyse and interpret these large amounts of information
- Help researchers to characterize relationships between biomedical entities (genes, diseases, . . .) quickly and efficiently

Motivations

- Genome-wide association studies (GWAS) : examination of many genetic variants (SNPs) in different individuals to study their correlations with phenotypic traits
- GWAS allow to identify groups of genes associated with a common phenotype
- GWAS do not provide information about associations in these gene groups

Contributions

- A biomedical text mining framework (Ailem et al., 2016) to augment the results of GWAS
- Benefits of co-clustering in biomedical text mining application
- Illustration on GWAS of asthma disease (Moffatt et al., 2010), which reported 10 genes associated with asthma
- Assess the strength of association between these genes and infer new candidate genes likely associated with asthma

M. Ailem et al. (2016). "Unsupervised text mining for assessing and augmenting GWAS results". In: *Journal of biomedical informatics* 60, pp. 252–259.

Context

- Exponential growth of biomedical text data (PUBMED, GO, . . .)
- There is a genuine need for text mining techniques to analyse and interpret these large amounts of information
- Help researchers to characterize relationships between biomedical entities (genes, diseases, . . .) quickly and efficiently

Motivations

- Genome-wide association studies (GWAS) : examination of many genetic variants (SNPs) in different individuals to study their correlations with phenotypic traits
- GWAS allow to identify groups of genes associated with a common phenotype
- GWAS do not provide information about associations in these gene groups

Contributions

- A biomedical text mining framework (Ailem et al., 2016) to augment the results of GWAS
- Benefits of co-clustering in biomedical text mining application
- Illustration on GWAS of asthma disease (Moffatt et al., 2010), which reported 10 genes associated with asthma
- Assess the strength of association between these genes and infer new candidate genes likely associated with asthma

---

M. Ailem et al. (2016). "Unsupervised text mining for assessing and augmenting GWAS results". In: *Journal of biomedical informatics* 60, pp. 252–259.

Outline

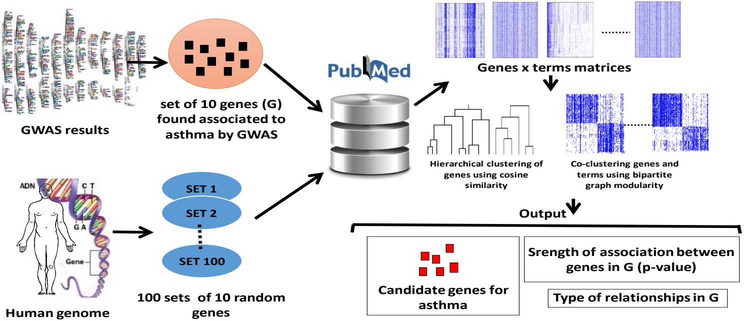
1 Introduction
- Context
- Co-clustering
- Motivations

2 Graph-based Co-clustering
- Graph Modularity
- Modularity for Co-clustering
- Experiments

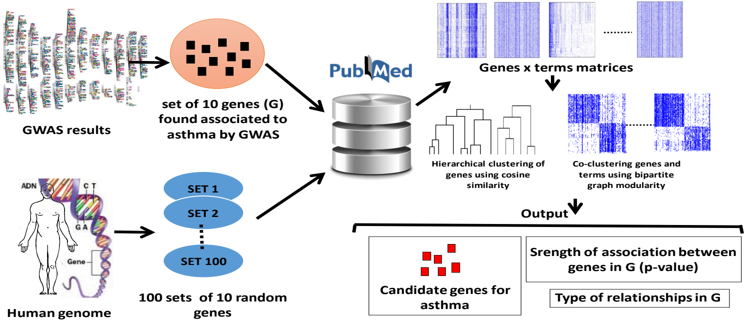
3 Model-based Co-clustering
- Sparse Poisson Latent Block Model (SPLBM)
- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
- Experiments

4 Using Co-clustering in Biomedical Text Mining Framework
- The Biomedical Framework
- Results and Discussions

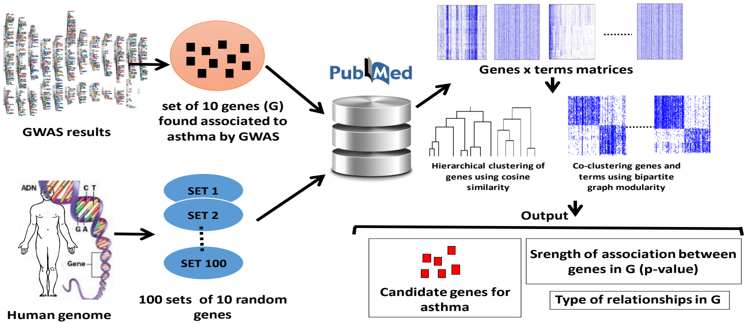
5 Conclusion and Perspectives



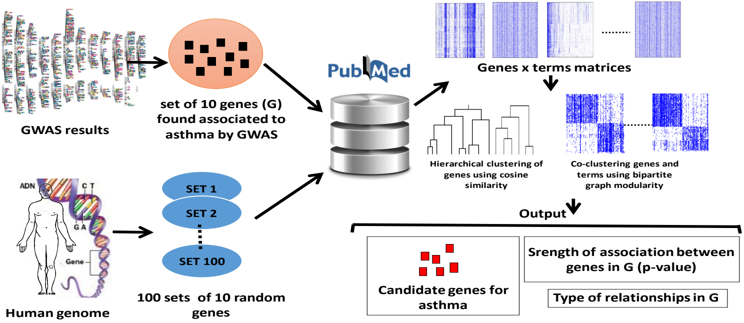
1. Input : set of 10 asthma genes ($\mathbf{G}$) and 100 sets of random genes $\{\mathbf{R}_1, \ldots, \mathbf{R}_{100}\}$ selected randomly from the human genome
2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set
   - Compare the cosine similarity between asthma gene vectors and random gene vectors
3. Assess the purity of asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set $(\mathbf{G} + \mathbf{R}_i)$ (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem
4. New candidate genes for asthma



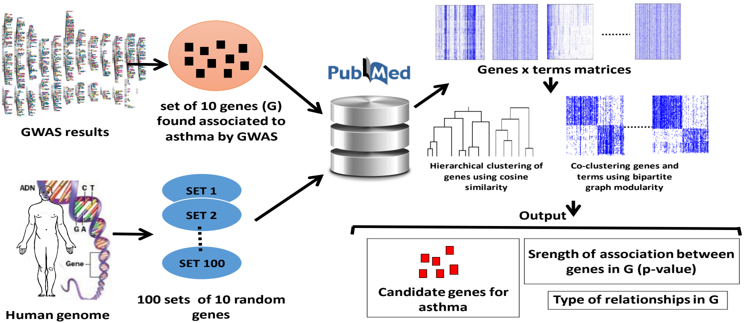
1. Input : set of 10 asthma genes ($\mathbf{G}$) and 100 sets of random genes $\{\mathbf{R}_1, \ldots, \mathbf{R}_{100}\}$ selected randomly from the human genome
2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set
   - Compare the cosine similarity between asthma gene vectors and random gene vectors
3. Assess the purity of asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set $(\mathbf{G} + \mathbf{R}_i)$ (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem
4. New candidate genes for asthma



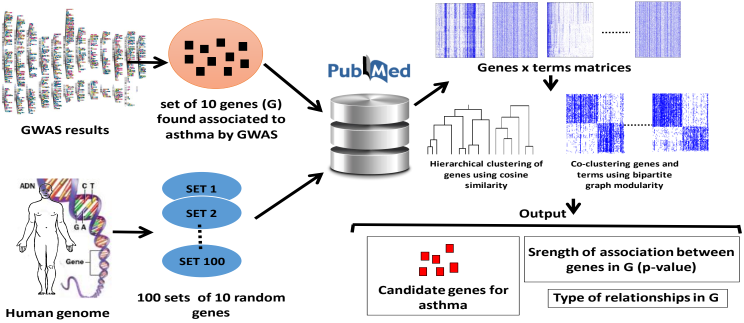
1. Input : set of 10 asthma genes ($\mathbf{G}$) and 100 sets of random genes $\{\mathbf{R}_1, \ldots, \mathbf{R}_{100}\}$ selected randomly from the human genome
2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set
   - Compare the cosine similarity between asthma gene vectors and random gene vectors
3. Assess the purity of asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set $(\mathbf{G} + \mathbf{R}_i)$ (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem
4. New candidate genes for asthma



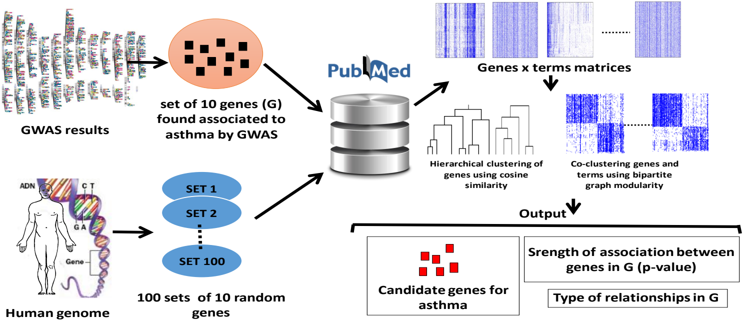
1. Input : set of 10 asthma genes ($\mathbf{G}$) and 100 sets of random genes $\{\mathbf{R}_1, \ldots, \mathbf{R}_{100}\}$ selected randomly from the human genome
2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set
   - Compare the cosine similarity between asthma gene vectors and random gene vectors
3. Assess the purity of asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set $(\mathbf{G} + \mathbf{R}_i)$ (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem
4. New candidate genes for asthma



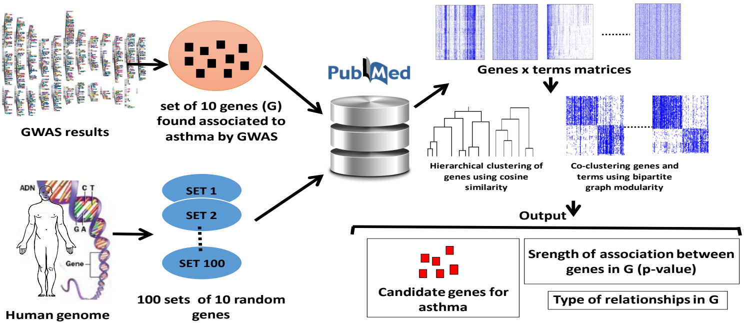
1. Input : set of 10 asthma genes ($\mathbf{G}$) and 100 sets of random genes $\{\mathbf{R}_1, \ldots, \mathbf{R}_{100}\}$ selected randomly from the human genome
2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set
   - Compare the cosine similarity between asthma gene vectors and random gene vectors
3. Assess the purity of asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set $(\mathbf{G} + \mathbf{R}_i)$ (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem
4. New candidate genes for asthma



1. Input : set of 10 asthma genes ($\mathbf{G}$) and 100 sets of random genes $\{\mathbf{R}_1, \ldots, \mathbf{R}_{100}\}$ selected randomly from the human genome
2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set
   - Compare the cosine similarity between asthma gene vectors and random gene vectors
3. Assess the purity of asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set $(\mathbf{G} + \mathbf{R}_i)$ (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem
4. New candidate genes for asthma



1. Input : set of 10 asthma genes ($\mathbf{G}$) and 100 sets of random genes $\{\mathbf{R}_1, \ldots, \mathbf{R}_{100}\}$ selected randomly from the human genome
2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set
   - Compare the cosine similarity between asthma gene vectors and random gene vectors
3. Assess the purity of asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set $(\mathbf{G} + \mathbf{R}_i)$ (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem
4. New candidate genes for asthma



1. Input : set of 10 asthma genes ($\mathbf{G}$) and 100 sets of random genes $\{\mathbf{R}_1, \ldots, \mathbf{R}_{100}\}$ selected randomly from the human genome
2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set
   - Compare the cosine similarity between asthma gene vectors and random gene vectors
3. Assess the purity of asthma-associated genes
   - Use the PUBMED database to create a $gene \times term$ matrix for each set $(\mathbf{G} + \mathbf{R}_i)$ (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem
4. New candidate genes for asthma

Outline

1 Introduction
- Context
- Co-clustering
- Motivations

2 Graph-based Co-clustering
- Graph Modularity
- Modularity for Co-clustering
- Experiments

3 Model-based Co-clustering
- Sparse Poisson Latent Block Model (SPLBM)
- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
- Experiments

4 Using Co-clustering in Biomedical Text Mining Framework
- The Biomedical Framework
- Results and Discussions

5 Conclusion and Perspectives

Results and Discussions

- The mean cosine similarities of asthma gene vectors is greater than would be expected by chance (empirical p-value $< 1\%$)
- Application of clustering and co-clustering to 100 sets of 20 genes that each included the 10 asthma genes plus 10 random genes, returned an average purity of 89%
- 20 Top terms of asthma genes co-cluster

Smoking	diabetes	th2	environmental
immune-mediated	chronic	enterotoxin	proinflammatory
child	microenvironment	cytokine	autoimmune
immunohistochemistry	childhood	influenza	asthma
drug	inflammation	crohn	necrosis

Results and Discussions

- The mean cosine similarities of asthma gene vectors is greater than would be expected by chance (empirical p-value $< 1\%$)
- Application of clustering and co-clustering to 100 sets of 20 genes that each included the 10 asthma genes plus 10 random genes, returned an average purity of 89%
- 20 Top terms of asthma genes co-cluster

Smoking	diabetes	th2	environmental
immune-mediated	chronic	enterotoxin	proinflammatory
child	microenvironment	cytokine	autoimmune
immunohistochemistry	childhood	influenza	asthma
drug	inflammation	crohn	necrosis

Results and Discussions

- The mean cosine similarities of asthma gene vectors is greater than would be expected by chance (empirical p-value $< 1\%$)
- Application of clustering and co-clustering to 100 sets of 20 genes that each included the 10 asthma genes plus 10 random genes, returned an average purity of 89%
- 20 Top terms of asthma genes co-cluster

Smoking	diabetes	th2	environmental
immune-mediated	chronic	enterotoxin	proinflammatory
child	microenvironment	cytokine	autoimmune
immunohistochemistry	childhood	influenza	asthma
drug	inflammation	crohn	necrosis

Candidate genes for asthma

- Moreover, 104 random genes were grouped with the 10 asthma associated-genes and, therefore, might be new candidates for asthma
- We ranked these candidate genes according to their cosine similarity with the group of asthma genes ($\mathbf{G}$)
- Study the Top 20 genes
- Use the biomedical literature and experts to validate the results

IL1RL1	*RAG1*	*CLEC1B*	***IL23R***
STAT6	*EFNA3*	*S1PR5*	*TGFBR1*
FCMR	***CXCL8/IL8***	*CHRNB4*	***NFKB1***
TNFRSF1A	*TMED1*	***NOD2***	***TSLP***
NLRP10	*POMP*	***SPINK1***	***PTGES***

- **Reported associated with asthma or allergy**
- **Reported associated with auto-immune diseases**
- **Encode proteins that are involved in immune-related mechanisms**

Candidate genes for asthma

- Moreover, 104 random genes were grouped with the 10 asthma associated-genes and, therefore, might be new candidates for asthma
- We ranked these candidate genes according to their cosine similarity with the group of asthma genes ($\mathbf{G}$)
- Study the Top 20 genes
- Use the biomedical literature and experts to validate the results

IL1RL1	*RAG1*	*CLEC1B*	***IL23R***
STAT6	*EFNA3*	*S1PR5*	*TGFBR1*
FCMR	***CXCL8/IL8***	*CHRNB4*	***NFKB1***
TNFRSF1A	*TMED1*	***NOD2***	***TSLP***
NLRP10	*POMP*	***SPINK1***	***PTGES***

- **Reported associated with asthma or allergy**
- **Reported associated with auto-immune diseases**
- **Encode proteins that are involved in immune-related mechanisms**

Outline

1. Introduction
   - Context
   - Co-clustering
   - Motivations
2. Graph-based Co-clustering
   - Graph Modularity
   - Modularity for Co-clustering
   - Experiments
3. Model-based Co-clustering
   - Sparse Poisson Latent Block Model (SPLBM)
   - Soft SPLBM-based Co-clustering Algorithm
   - Hard SPLBM-based Co-clustering Algorithm
   - Experiments
4. Using Co-clustering in Biomedical Text Mining Framework
   - The Biomedical Framework
   - Results and Discussions
5. Conclusion and Perspectives

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Main contributions

- Three main contributions
  1. Graph-based Diagonal co-clustering approach
  2. Model-based Diagonal co-clustering approach
  3. Using Co-clustering for Biomedical Text Mining
- Assessing the right number of co-clusters
- Methods for assessing term clusters
- Soft, hard and stochastic assignments
- Extensive experiments on real world text datasets
- Availability : Coclust python module (https://pypi.python.org/pypi/coclust)

Toward Semantic (co)-clustering

Motivation

- Existing (co)-clustering methods ignore the semantic relationships between words, which may result in a significant loss of semantics since documents that are about the same topic may not necessarily use exactly the same vocabulary.

Contribution

- We propose a new (co)-clustering models which goes beyond the bag of word representation so as to preserve more semantics.
- We achieve our objective by successfully integrating `word2vec` into a (co)-clustering framework.
- The proposed models substantially outperforms existing (co)-clustering models in terms of document clustering, cluster interpretability as well as document/word embedding.

M. Ailem, A. Salah, and M. Nadif (2017). "Non-negative Matrix Factorization Meets Word Embedding". In: *SIGIR*. ACM, pp. 1081–1084.

A. Salah, M. Ailem, and M. Nadif (2017). "A Way to Boost Semi-NMF for Document Clustering". In: *CIKM*. ACM, pp. 2275–2278.

A. Salah, M. Ailem, and M. Nadif (2018). "Word Co-occurrence Regularized Non-Negative MatrixTri-Factorization for Text Data Co-clustering". In: *AAAI'2018*.

Perspectives

- Investigate an overlapping version of the Coclus algorithm
- Study the theoretical link between graph-based and model-based approaches
- Assessing the number of (co-)clusters for model-based approaches using information criteria such as BIC, AIC, ICL ...
- Investigate Bayesian non-parametric formulations of SPLBM, which would allows us to overcome the problem of the number of clusters as well as handle evolving data

Thank you for your attention!

References I

Ailem, M., F. Role, and M. Nadif (2015). “Co-clustering Document-term Matrices by Direct Maximization of Graph Modularity”. In: *CIKM'2015*. ACM, pp. 1807–1810.

– (2016). “Graph modularity maximization as an effective method for co-clustering text data”. In: *Knowledge-Based Systems Journal* 109, pp. 160–173.

– (2017a). “Model-based co-clustering for the effective handling of sparse data”. In: *Pattern Recognition* 72, pp. 108–122.

Ailem, M., F. Role, and M. Nadif (2017b). “Sparse Poisson Latent Block Model for Document Clustering”. In: *IEEE TKDE journal* 29.7, p. 1563.

Ailem, M., A. Salah, and M. Nadif (2017). “Non-negative Matrix Factorization Meets Word Embedding”. In: *SIGIR*. ACM, pp. 1081–1084.

Ailem, M. et al. (2016). “Unsupervised text mining for assessing and augmenting GWAS results”. In: *Journal of biomedical informatics* 60, pp. 252–259.

References II

- Bisson, G. and F. Hussain (2008). "Chi-sim: A new similarity measure for the co-clustering task". In: *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*. IEEE, pp. 211–217.
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.
- Dhillon, I. S., S. Mallela, and D. S. Modha (2003). "Information-Theoretic Co-Clustering". In: *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pp. 89–98.
- Dhillon, I. (2001). "Co-clustering documents and words using bipartite spectral graph partitioning". In: KDD '01. San Francisco, California, pp. 269–274.
- Dhillon, I. and D. Modha (2001). "Concept Decompositions for Large Sparse Text Data Using Clustering". In: *Mach. Learn.* 42.1-2, pp. 143–175.

References III

Govaert, G. and M. Nadif (2003). “Clustering with block mixture models”. In: *Pattern Recognition* 36.2, pp. 463–473.

– (2010). “Latent block model for contingency table”. In: *Communications in Statistics—Theory and Methods* 39.3, pp. 416–425.

Hartigan, J. A. (1972). “Direct clustering of a data matrix”. In: *Journal of the American Statistical Association*, pp. 123–129.

Labiod, L. and M. Nadif (2011). “Co-clustering for Binary and Categorical Data with Maximum Modularity.”. In: *ICDM*, pp. 1140–1145.

Li, T. (2005). “A general model for clustering binary data.”. In: *KDD ’05*, pp. 188–197.

Moffatt, M. F. et al. (2010). “A large-scale, consortium-based genomewide association study of asthma”. In: *New England Journal of Medicine* 363, pp. 1211–1221.

Rand, W. (1971). “Objective criteria for the evaluation of clustering methods”. In: *Journal of the American Statistical association* 66.336, pp. 846–850.

References IV

- Salah, A., M. Ailem, and M. Nadif (2017). “A Way to Boost Semi-NMF for Document Clustering”. In: *CIKM*. ACM, pp. 2275–2278.
- – (2018). “Word Co-occurrence Regularized Non-Negative MatrixTri-Factorization for Text Data Co-clustering”. In: *AAAI’2018*.
- Strehl, A. and J. Ghosh (2003). “Cluster ensembles—a knowledge reuse framework for combining multiple partitions”. In: *The Journal of Machine Learning Research* 3, pp. 583–617.
- Wang, H. et al. (2011). “Fast nonnegative matrix tri-factorization for large-scale data co-clustering”. In: *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*. Vol. 22. 1, p. 1553.
- Zhao, Y. and G. K (2002). *Comparison of agglomerative and partitional document clustering algorithms*. Tech. rep. DTIC Document.