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
• 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

- **Corpus**
- **Document-term Matrix**

**Pre-processing**
- Pos Tagging
- Lemmatization
- Stop Word Removal
- Punctuation Removal
- Filtering

**Document Representation**
- Vector space model
- \( tf_{ij} = \text{Frequency of term } j \text{ in document } i \)
Data Representation

**Context**

**Co-clustering**

**Motivations**

**Data Representation**

- **Corpus**
- **Document-term Matrix**

- **Pre-processing**
  - **Pos Tagging**
  - **Lemmatization**
  - **Stop Word Removal**
  - **Punctuation Removal**
  - **Filtering**

**Document Representation**

- **Vector space model**
- \( tf_{ij} = \text{Frequency of term } j \text{ in document } i \)

**Weighting scheme**

- **TF-IDF weighting scheme**

\[
 w_{ij} = tf_{ij} \times \log \frac{N}{d_j}
\]
Introduction

Graph-based Co-clustering
Model-based Co-clustering
Using Co-clustering in Biomedical Text Mining Framework
Conclusion and Perspectives
References

Context

Co-clustering

Motivations

Data Representation

![Diagram showing data representation process]

High dimensionality
Sparsity : > 99% zeros

**Corpus**

```
[Documents]
```

**Document-term Matrix**

```
<table>
<thead>
<tr>
<th>Document</th>
<th>t_1</th>
<th>t_2</th>
<th>...</th>
<th>t_d</th>
</tr>
</thead>
<tbody>
<tr>
<td>d_1</td>
<td>tf_{11}</td>
<td>tf_{12}</td>
<td>...</td>
<td>tf_{1d}</td>
</tr>
<tr>
<td>d_2</td>
<td>tf_{21}</td>
<td>tf_{22}</td>
<td>...</td>
<td>tf_{2d}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>d_n</td>
<td>tf_{n1}</td>
<td>tf_{n2}</td>
<td>...</td>
<td>tf_{nd}</td>
</tr>
</tbody>
</table>
```

Pre-processing

- Pos Tagging
- Lemmatization
- Stop Word Removal
- Punctuation Removal
- Filtering

**Document Representation**

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

**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

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>Term 4</th>
<th>Term 5</th>
<th>Term 6</th>
<th>Term 7</th>
<th>Term 8</th>
<th>Term 9</th>
<th>Term 10</th>
<th>Term 11</th>
<th>Term 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doc 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doc 5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doc 6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doc 9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) Original Data
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

<table>
<thead>
<tr>
<th></th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>Term 4</th>
<th>Term 5</th>
<th>Term 6</th>
<th>Term 7</th>
<th>Term 8</th>
<th>Term 9</th>
<th>Term 10</th>
<th>Term 11</th>
<th>Term 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>0 1 0 0 1 0 0 0 1 0 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 2</td>
<td>1 0 0 1 0 1 1 0 1 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 3</td>
<td>1 0 1 1 0 1 1 1 1 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 4</td>
<td>1 0 0 0 0 1 1 0 0 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 5</td>
<td>1 1 0 0 0 1 1 0 0 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 6</td>
<td>1 1 0 0 0 0 0 1 0 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 7</td>
<td>1 0 1 1 0 1 1 1 0 1 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 8</td>
<td>0 1 0 0 1 0 0 0 1 0 0 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 9</td>
<td>1 0 0 1 0 0 1 0 0 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Original Data

(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

<table>
<thead>
<tr>
<th>Doc 1</th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>Term 4</th>
<th>Term 5</th>
<th>Term 6</th>
<th>Term 7</th>
<th>Term 8</th>
<th>Term 9</th>
<th>Term 10</th>
<th>Term 11</th>
<th>Term 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc 3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc 5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doc 8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doc 9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(a) Original Data

<table>
<thead>
<tr>
<th>Doc 1</th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>Term 4</th>
<th>Term 5</th>
<th>Term 6</th>
<th>Term 7</th>
<th>Term 8</th>
<th>Term 9</th>
<th>Term 10</th>
<th>Term 11</th>
<th>Term 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc 6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc 7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(b) Clustering

<table>
<thead>
<tr>
<th>Doc 1</th>
<th>Term 4</th>
<th>Term 10</th>
<th>Term 6</th>
<th>Term 12</th>
<th>Term 3</th>
<th>Term 9</th>
<th>Term 5</th>
<th>Term 7</th>
<th>Term 8</th>
<th>Term 11</th>
<th>Term 1</th>
<th>Term 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc 6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc 7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc 9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(c) 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

(Original Data)

(Why Co-clustering?)

(a) Original Data

(b) Clustering

(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
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 ignores noisy ones
• Cluster hypothesis
• Allows a direct interpretation of co-clusters
• Parsimonious

Contributions
• Graph-based block diagonal clustering
• Model-based block diagonal 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.
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
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

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

- 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
- $Q = \text{(number of intra-cluster edges)} - \text{(expected number of edges)}$

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

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

- 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 $k$
Introduction

Context

Co-clustering

Motivations

Graph-based Co-clustering

Graph Modularity

Modularity for Co-clustering

Experiments

Model-based Co-clustering

Sparse Poisson Latent Block Model (SPLBM)

Soft SPLBM-based Co-clustering Algorithm

Hard SPLBM-based Co-clustering Algorithm

Experiments

Using Co-clustering in Biomedical Text Mining Framework

The Biomedical Framework

Results and Discussions

Conclusion and Perspectives
Modularity for Co-clustering

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

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

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

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.

$$Q(A, C) = \frac{1}{a..} Trace[(A - \delta)^t ZW^t] = Q(A, ZW^t). \quad (4)$$
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) $Q(\mathbf{A}, \mathbf{C}) = \frac{1}{a_{..}} \sum_{i=1}^{n} \sum_{k=1}^{g} (a_{ik}^{W} - \frac{a_{i}.a_{.k}^{W}}{a_{..}})z_{ik} = \frac{1}{a_{..}} Trace[(\mathbf{A}^{W} - \delta^{W})^{t}\mathbf{Z}] = Q(\mathbf{A}^{W}, \mathbf{Z})$

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

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

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

Algorithm 1: Coclus

Input: positive data matrix $A$, number of co-clusters $g$

Step 1. Initialization of $W$

repeat

Step 2. Compute $A^W = AW$

Step 3. Compute $Z$ maximizing $Q(A^W, Z)$ by

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

Step 4. Compute $A^Z = Z' A$

Step 5. Compute $W$ maximizing $Q(A^Z, W)$ by

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

Step 6. Compute $Q(A, ZW^t)$

until Convergence;

Output: partition matrices $Z$ and $W$, and modularity value $Q$

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
## Experiments

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#Documents</th>
<th>#Words</th>
<th>#Clusters</th>
<th>Sparsity (%)</th>
<th>Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASSIC4</td>
<td>7095</td>
<td>5896</td>
<td>4</td>
<td>99.41</td>
<td>0.323</td>
</tr>
<tr>
<td>NG20</td>
<td>19949</td>
<td>43586</td>
<td>20</td>
<td>99.99</td>
<td>0.991</td>
</tr>
<tr>
<td>SPORTS</td>
<td>8580</td>
<td>14870</td>
<td>7</td>
<td>99.99</td>
<td>0.036</td>
</tr>
<tr>
<td>REVIEWS</td>
<td>4069</td>
<td>18483</td>
<td>5</td>
<td>99.99</td>
<td>0.098</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Type</th>
<th>References</th>
<th>Co-clustering</th>
<th>Type of implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spec</td>
<td>Positive data</td>
<td>(I. Dhillon, 2001)</td>
<td>Diagonal</td>
<td>Scikit Learn</td>
</tr>
<tr>
<td>Block</td>
<td>Binary</td>
<td>(Li, 2005)</td>
<td>Diagonal</td>
<td>Our python implementation</td>
</tr>
<tr>
<td>ITCC</td>
<td>Positive data</td>
<td>(I. S. Dhillon, Mallela, and D. S. Modha, 2003)</td>
<td>Non-diagonal</td>
<td>C++ implementation</td>
</tr>
<tr>
<td>SpecCo</td>
<td>Positive data</td>
<td>(Labiod and Nadif, 2011)</td>
<td>Diagonal</td>
<td>Our python implementation</td>
</tr>
<tr>
<td>(\chi^2)-Sim</td>
<td>Positive data</td>
<td>(Bisson and Hussain, 2008)</td>
<td>Non-diagonal</td>
<td>MATLAB implementation of the authors</td>
</tr>
<tr>
<td>FNMTF</td>
<td>Positive data</td>
<td>(Wang et al., 2011)</td>
<td>Non-diagonal</td>
<td>MATLAB implementation of the authors</td>
</tr>
</tbody>
</table>
Using Co-clustering in Biomedical Text Mining Framework

### Introduction

Graph-based Co-clustering

Model-based Co-clustering

### Graph Modularity

### Modularity for Co-clustering

### Experiments

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#Documents</th>
<th>#Words</th>
<th>#Clusters</th>
<th>Sparsity (%)</th>
<th>Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASSIC4</td>
<td>7095</td>
<td>5896</td>
<td>4</td>
<td>99.41</td>
<td>0.323</td>
</tr>
<tr>
<td>NG20</td>
<td>19949</td>
<td>43586</td>
<td>20</td>
<td>99.99</td>
<td>0.991</td>
</tr>
<tr>
<td>SPORTS</td>
<td>8580</td>
<td>14870</td>
<td>7</td>
<td>99.99</td>
<td>0.036</td>
</tr>
<tr>
<td>REVIEWS</td>
<td>4069</td>
<td>18483</td>
<td>5</td>
<td>99.99</td>
<td>0.098</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Type</th>
<th>References</th>
<th>Co-clustering</th>
<th>Type of Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spec</td>
<td>Positive data</td>
<td>(I. Dhillon, 2001)</td>
<td>Diagonal</td>
<td>Scikit Learn</td>
</tr>
<tr>
<td>Block</td>
<td>Binary</td>
<td>(Li, 2005)</td>
<td>Diagonal</td>
<td>Our python implementation</td>
</tr>
<tr>
<td>ITCC</td>
<td>Positive data</td>
<td>(I. S. Dhillon, Mallela, and D. S. Modha, 2003)</td>
<td>Non-diagonal</td>
<td>C++ implementation</td>
</tr>
<tr>
<td>SpecCo</td>
<td>Positive data</td>
<td>(Labiod and Nadif, 2011)</td>
<td>Diagonal</td>
<td>Our python implementation</td>
</tr>
<tr>
<td>(\chi)-Sim</td>
<td>Positive data</td>
<td>(Bisson and Hussain, 2008)</td>
<td>Non-diagonal</td>
<td>MATLAB implementation of the authors</td>
</tr>
<tr>
<td>FNMTF</td>
<td>Positive data</td>
<td>(Wang et al., 2011)</td>
<td>Non-diagonal</td>
<td>MATLAB implementation of the authors</td>
</tr>
</tbody>
</table>

- 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)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>per.</th>
<th>Spec</th>
<th>ITCC</th>
<th>Block</th>
<th>SpecCo (\chi)-Sim</th>
<th>FNMTF</th>
<th>CoClus</th>
<th>Spec</th>
<th>ITCC</th>
<th>SpecCo (\chi)-Sim</th>
<th>FNMTF</th>
<th>CoClus</th>
<th>Spec</th>
<th>ITCC</th>
<th>SpecCo (\chi)-Sim</th>
<th>FNMTF</th>
<th>CoClus</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASSIC4</td>
<td>Acc</td>
<td>0.34</td>
<td>0.65</td>
<td>0.52</td>
<td>0.45</td>
<td>0.31</td>
<td>0.50</td>
<td>0.90</td>
<td>0.53</td>
<td>0.87</td>
<td>0.58</td>
<td>0.31</td>
<td>0.56</td>
<td>0.90</td>
<td>0.44</td>
<td>0.60</td>
<td>0.45</td>
</tr>
<tr>
<td>NMI</td>
<td></td>
<td>0.14</td>
<td>0.51</td>
<td>0.16</td>
<td>0.02</td>
<td>0.15</td>
<td>0.30</td>
<td>0.72</td>
<td>0.45</td>
<td>0.67</td>
<td>0.48</td>
<td>0.15</td>
<td>0.30</td>
<td>0.73</td>
<td>0.02</td>
<td>0.55</td>
<td>0.09</td>
</tr>
<tr>
<td>NG20</td>
<td>Acc</td>
<td>0.14</td>
<td>0.43</td>
<td>0.20</td>
<td>0.19</td>
<td>0.26</td>
<td>0.13</td>
<td>0.40</td>
<td>0.05</td>
<td>0.45</td>
<td>0.30</td>
<td>0.30</td>
<td>0.09</td>
<td>0.37</td>
<td>0.19</td>
<td>0.41</td>
<td>0.15</td>
</tr>
<tr>
<td>NMI</td>
<td></td>
<td>0.29</td>
<td>0.55</td>
<td>0.22</td>
<td>0.42</td>
<td>0.33</td>
<td>0.03</td>
<td>0.55</td>
<td>0.02</td>
<td>0.52</td>
<td>0.49</td>
<td>0.37</td>
<td>0.07</td>
<td>0.52</td>
<td>0.32</td>
<td>0.44</td>
<td>0.38</td>
</tr>
<tr>
<td>SPORTS</td>
<td>Acc</td>
<td>0.56</td>
<td>0.45</td>
<td>0.47</td>
<td>0.59</td>
<td>0.57</td>
<td>0.28</td>
<td>0.70</td>
<td>0.44</td>
<td>0.56</td>
<td>0.68</td>
<td>0.53</td>
<td>0.36</td>
<td>0.75</td>
<td>0.45</td>
<td>0.54</td>
<td>0.61</td>
</tr>
<tr>
<td>NMI</td>
<td></td>
<td>0.47</td>
<td>0.49</td>
<td>0.38</td>
<td>0.45</td>
<td>0.48</td>
<td>0.15</td>
<td>0.54</td>
<td>0.38</td>
<td>0.58</td>
<td>0.59</td>
<td>0.48</td>
<td>0.19</td>
<td>0.62</td>
<td>0.43</td>
<td>0.58</td>
<td>0.45</td>
</tr>
<tr>
<td>REVIEWS</td>
<td>Acc</td>
<td>0.56</td>
<td>0.58</td>
<td>0.53</td>
<td>0.59</td>
<td>0.46</td>
<td>0.34</td>
<td>0.65</td>
<td>0.50</td>
<td>0.71</td>
<td>0.45</td>
<td>0.41</td>
<td>0.38</td>
<td>0.72</td>
<td>0.35</td>
<td>0.63</td>
<td>0.46</td>
</tr>
<tr>
<td>NMI</td>
<td></td>
<td>0.36</td>
<td>0.46</td>
<td>0.42</td>
<td>0.39</td>
<td>0.31</td>
<td>0.18</td>
<td>0.54</td>
<td>0.40</td>
<td>0.57</td>
<td>0.34</td>
<td>0.23</td>
<td>0.17</td>
<td>0.58</td>
<td>0.03</td>
<td>0.51</td>
<td>0.35</td>
</tr>
</tbody>
</table>
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 its maximum value.
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 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)
**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 $X = (x_1, \ldots, x_n)$ where $x_i = (x_{i1}, \ldots x_{id}) \in \mathbb{R}^d$ is generated from a probability density function (pdf) with density:

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

The likelihood of data $X$ can be written as:

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

where

- $f_k(\cdot, \alpha_k)$ is the density of an observation $x_i$ from the $k$-th component
- $\alpha_k$'s are the corresponding class parameters
- $\pi_k$ represents the proportions of each cluster.
- Each component $k$ of the mixture represents a cluster.
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 $X$ of size $n \times d$ can be written

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

Where

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

**Algorithm 2: Generative Process of LBM**

**Input**: \( n, d, g, m, \theta = (\pi, \rho, \alpha) \)

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

For \( i = 1 \) to \( n \) do

- Generate the row label \( z_i \) according to the multinomial distribution \( \pi = (\pi_1, \ldots, \pi_g) \)

End

For \( j = 1 \) to \( d \) do

- Generate the column label \( w_j \) according to the multinomial distribution \( \rho = (\rho_1, \ldots, \rho_g) \)

End

For \( i = 1 \) to \( n \) do

  For \( j = 1 \) to \( d \) do

    - Generate a value \( x_{ij} \) according to the distribution \( f(\cdot; \alpha_{z_i, w_j}) \)

  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 $\lambda_{ij}$ 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 $\lambda_{ij}$ takes the following form:

$$
\lambda_{ij} = x_i.x.j \sum_k \sum_{\ell \neq k} z_{ik}w_{j\ell} \gamma.
$$

- 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

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

• For each diagonal block $kk$ the values $x_{ij}$ are distributed according to the Poisson distribution $P(\lambda_{ij})$ where the parameter $\lambda_{ij}$ 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 $P(\lambda_{ij})$ where the parameter $\lambda_{ij}$ takes the following form:

$$\lambda_{ij} = x_i.x_j \sum_{k, \ell \neq k} z_{ik}w_{j\ell} \gamma.$$  

• Assuming $\forall \ell \neq k$ $\gamma_{k\ell} = \gamma$ leads to suppose that all blocks outside the diagonal share the same parameter.

Likelihood function

$$f(X; \theta) = \sum_{(z,w) \in Z \times W} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,k} \rho_{\ell}^{w_{jk}}$$

$$\times \prod_{i,j,k} (f(x_{ij}; \alpha_{kk}))^{z_{ik}w_{jk}} \times \prod_{i,j,\ell \neq k} (f(x_{ij}; \alpha_{k\ell}))^{z_{ik}w_{j\ell}}$$
Sparse Poisson Latent Block Model (SPLBM)

Complete Data Likelihood

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

Complete Data Log-likelihood

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

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

where \( x_{kk}^{ZW} = \sum_{ij} z_{ik} w_{jk} x_{ij}, z_{ik} = \sum_i z_{ik} \) and \( w_{jk} = \sum_j w_{jk}, x_{kk}^{Z} = \sum_i z_{ik} x_{i}, \) and \( x_{kk}^{W} = \sum_j w_{jk} x_{j} \)
Sparse Poisson Latent Block Model (SPLBM)

Complete Data Likelihood

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

Complete Data Log-likelihood

\[ L_C(\mathbf{Z}, \mathbf{W}, \theta) = \log f(\mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta) = \sum_{k=1}^{g} L_C^k \]

\[ L_C^k = z_k \log \pi_k + w_k \log \rho_k + \mathbf{x}_{kk}^{ZW} \log \left( \frac{\gamma_{kk}}{\gamma} \right) - x_k^Z \times x_k^W (\gamma_{kk} - \gamma) + \frac{N}{g} (\log(\gamma) - \gamma N) \]

where \( x_{kk}^{ZW} = \sum_{ij} z_{ik} w_{jk} x_{ij} \), \( z_k = \sum_i z_{ik} \) and \( w_k = \sum_j w_{jk} \), \( x_k^Z = \sum_i z_{ik} x_i \) and \( x_k^W = \sum_j w_{jk} x_j \)
Sparse Poisson Latent Block Model (SPLBM)

**Complete Data Likelihood**

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

**Complete Data Log-likelihood**

\[
L_C(Z, W, \theta) = \log f(X, Z, W; \theta) = \sum_{k=1}^{g} L_C^k
\]

\[
L_C^k = z.k \log \pi_k + w.k \log \rho_k + x_{kk}^{ZW} \log \left( \frac{\gamma_{kk}}{\gamma} \right) - x_{k}^{Z} x_{k}^{W} (\gamma_{kk} - \gamma) + \frac{N}{g} \left( \log(\gamma) - \gamma N \right)
\]

where \( x_{kk}^{ZW} = \sum_{ij} z_{ik} w_{jk} x_{ij} \), \( z.k = \sum_{i} z_{ik} \) and \( w.k = \sum_{j} w_{jk} \), \( x_{k}^{Z} = \sum_{i} z_{ik} x_{i} \) and \( x_{k}^{W} = \sum_{j} w_{jk} x_{j} \)

(a) Traditional LBM - 64 parameters  
(b) Sparse PLBM - 9 parameters
Sparse Poisson Latent Block Model (SPLBM)

- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm

Experiments

Figure: SPLBM-based co-clustering algorithms


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 $\theta$, $\tilde{Z}$ and $\tilde{W}$

• We rely on the Expectation-Maximization (EM) algorithm that consists in maximizing the expectation of the complete data likelihood $L_C(Z, W, \theta)$ given by:

$$
\mathbb{E}\left( L_C(Z, W, \theta) | \theta^{(t)}, X \right) = \sum_{i,k} \tilde{z}_{ik}^{(t)} \log \pi_k + \sum_{j,k} \tilde{w}_{jk}^{(t)} \log \rho_k
$$

$$
+ \sum_{i,j,k} \tilde{e}_{ijk}^{(t)} (x_{ij} \log(\gamma_{kk}) - x_{i..} x_{..j} \gamma_{kk})
$$

$$
+ \sum_{i,j,k,\ell \neq k} \tilde{e}_{ikj\ell}^{(t)} (x_{ij} \log(\gamma) - x_{i..} x_{..j} \gamma)
$$

where $\tilde{z}_{ik}^{(t)} = \mathbb{E}(z_{ik} = 1|X, \theta^{(t)})$, $\tilde{w}_{j\ell}^{(t)} = \mathbb{E}(w_{j\ell}^{(t)} = 1|X, \theta^{(t)})$,

$\tilde{e}_{ikj\ell}^{(t)} = \mathbb{E}(e_{ikj\ell}^{(t)} = 1|X, \theta^{(t)}) = \mathbb{E}(z_{ik} w_{j\ell} = 1|X, \theta^{(t)})$ and $\tilde{e}_{ijk}^{(t)} = \mathbb{E}(z_{ik} w_{jk} = 1|X, \theta^{(t)})$. 
Soft SPLBM-based Co-clustering Algorithm

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

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

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

The coupling of $Z$ and $W$ 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:

\[
F_C(\tilde{Z}, \tilde{W}, \theta) = L_C(\tilde{Z}, \tilde{W}, \theta) + H(\tilde{Z}) + H(\tilde{W}),
\]

- **where** \(H(\tilde{Z}) = -\sum_{i,k} \tilde{z}_{ik} \log \tilde{z}_{ik}\) and \(H(\tilde{W}) = -\sum_{j,k} \tilde{w}_{jk} \log \tilde{w}_{jk}\) are respectively the entropy of the missing variables \(\tilde{Z}\) and \(\tilde{W}\).

- **\(L_C(\tilde{Z}, \tilde{W}, \theta)\)** is the soft complete data likelihood defined as follows:

\[
L_C(\tilde{Z}, \tilde{W}, \theta) = \sum_{i,k} \tilde{z}_{ik} \log \pi_k + \sum_{j,k} \tilde{w}_{jk} \log \rho_k + \sum_{i,j,k} \tilde{z}_{ik} \tilde{w}_{jk} x_{ij} \log \left( \frac{\gamma_{kk}}{\gamma} \right)
- \sum_{k} x_{k} x_{k} \gamma_{kk} + \gamma \sum_{k} x_{k} x_{k} + N(\log(\gamma) - \gamma N)
\]

- The SPLBvem algorithm consists of the expectation and maximization steps.
M-step

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

\[
\hat{\gamma}_{kk} = \frac{x_{kk} \bar{Z} \bar{W}}{x_k \bar{Z} \bar{W}_k}.
\]

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

\[
\hat{\gamma} = \frac{N - \sum_k x_{kk} \bar{Z} \bar{W}}{N^2 - \sum_k x_k \bar{Z} \bar{W}_k}.
\]

- **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 \( F_C \) are respectively given by \( \pi_k = \frac{\bar{Z}_k}{n} \) and \( \rho_k = \frac{\bar{W}_k}{d} \).
Model Fitting Using the Variational EM Algorithm

**E-step**

- The E-step consists in computing the posterior probabilities $\tilde{z}_{ik}$ and $\tilde{w}_{jk}$ maximizing $F_C$

  Plug the estimation of $\gamma_{kk}$'s and $\gamma$ (explicitly in some terms of $F_C$) we obtain

  \[
  F_C(\tilde{Z}, \tilde{W}, \hat{\theta}) = \sum_{i,k} \tilde{z}_{ik} \log \hat{\pi}_k + \sum_{j,k} \tilde{w}_{jk} \log \hat{\rho}_k + \sum_{i,j,k} \tilde{z}_{ik} \tilde{w}_{jk} x_{ij} \log \left( \frac{\hat{\gamma}_{kk}}{\hat{\gamma}} \right) + N(\log(\hat{\gamma}) - 1) - \sum_{i,k} \tilde{z}_{ik} \log \tilde{z}_{ik} - \sum_{j,k} \tilde{w}_{jk} \log \tilde{w}_{jk}.
  \]

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

  \[
  \tilde{z}_{ik} \propto \pi_k \exp (x_{ik} \log \frac{\gamma_{kk}}{\gamma}).
  \]

  \[
  \tilde{w}_{jk} \propto \rho_k \exp (x_{kj} \log \frac{\gamma_{kk}}{\gamma}).
  \]
**The SPLBvem Algorithm**

**Algorithm 3:** SPLBvem

**Input:** $X, g$

**Initialization:** $\tilde{Z}, \tilde{W}, \pi_k, \rho_k, \gamma_{kk}, \gamma$

repeat

\[ x_{ik}^\tilde{W} = \sum_j \tilde{w}_{jk} x_{ij} \]

**step 1.** $\tilde{z}_{ik} \propto \pi_k \exp(x_{ik}^\tilde{W} \log \gamma_{kk})$

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

\[ x_{kj}^\tilde{Z} = \sum_i \tilde{z}_{ik} x_{ij} \]

**step 3.** $\tilde{w}_{jk} \propto \rho_k \exp(x_{kj}^\tilde{Z} \log \gamma_{kk})$

**step 4.** $\rho_k = \frac{\tilde{w}_{jk}}{n}$, $\gamma_{kk} = \frac{\sum_j \tilde{w}_{jk} x_{kj}^\tilde{Z}}{x_{kj}^\tilde{Z} x_{kj}} = \frac{x_{kk}^\tilde{W}}{x_{kj}^\tilde{Z} x_{kj}}$, $\gamma = \frac{N=\sum_k x_{kk}^\tilde{Z} x_{kk}}{N^2-\sum_k x_{kk}^\tilde{Z} x_{kk}}$

until Convergence;

**Output:** $\tilde{Z}, \tilde{W}, \pi_k, \rho_k, \gamma_{kk}, \gamma$
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:** $X, g$
**Initialization:** $Z, W, \pi_k, \rho_k, \gamma_{kk}, \gamma$

repeat

| Step 1 | $\bar{z}_{ik} \propto \pi_k \exp(x_{ik} \log \frac{\gamma_{kk}}{\gamma})$
|---|---|
| $z_{ik} = \arg \max_k \bar{z}_{ik}$
| Step 2 | $\pi_k = \frac{\bar{z}_k}{n}, \gamma_{kk} = \frac{\sum_i \bar{z}_{ik} x_{ik}}{x_k \cdot k} = \frac{\bar{Z} W}{x_k \cdot k}$
| $\gamma = \frac{N - \sum_k x_{kk} \bar{Z} W}{N^2 - \sum_k \bar{Z}^2 \bar{W}}$

| $\bar{Z} x_{kj} = \sum_i \bar{z}_{ik} x_{ij}$
| Step 3 | $w_{jk} \propto \rho_k \exp(x_{kj} \log \frac{\gamma_{kk}}{\gamma})$
| $w_{jk} = \arg \max_k \bar{w}_{jk}$
| $\rho_k = \frac{\bar{w}_k}{d}, \gamma_{kk} = \frac{\sum_j \bar{w}_{jk} \bar{Z} x_{kj}}{x_k \cdot k \cdot j} = \frac{\bar{W} \bar{Z} W}{x_k \cdot k}$
| $\gamma = \frac{N - \sum_k x_{kk} \bar{W} \bar{Z} W}{N^2 - \sum_k \bar{Z}^2 \bar{W}}$

until Convergence;

**Output:** $Z, W, \pi_k, \rho_k, \gamma_{kk}, \gamma$

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

<table>
<thead>
<tr>
<th>Input :</th>
<th>X, g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization :</td>
<td>Z, W, π_k, ρ_k, γ_kk, γ</td>
</tr>
<tr>
<td>repeat</td>
<td></td>
</tr>
<tr>
<td>[ x_{ik} = \sum_j \tilde{w}<em>{jk} x</em>{ij} ]</td>
<td></td>
</tr>
<tr>
<td>step 1.</td>
<td>[ \tilde{z}<em>{ik} \propto \pi_k \exp(x</em>{ik} \log \frac{\gamma_k}{\gamma}) ]</td>
</tr>
<tr>
<td>step 1’.</td>
<td>[ z_{ik} = \operatorname{arg\ max}<em>k \tilde{z}</em>{ik} ]</td>
</tr>
<tr>
<td>step 2.</td>
<td>[ \pi_k = \frac{\tilde{z}<em>k}{n}, \gamma_k = \frac{\sum_i \tilde{z}</em>{ik} \tilde{w}<em>{ik}}{x_k \bar{Z} \bar{W} \gamma_k \gamma</em>{kk}}, \gamma = \frac{N - \sum_k x_{kk}}{N^2 - \sum_k x_k \bar{Z} \bar{W} \gamma_k \gamma_{kk}} ]</td>
</tr>
<tr>
<td>[ \bar{Z} x_{kj} = \sum_i \tilde{z}<em>{ik} x</em>{ij} ]</td>
<td></td>
</tr>
<tr>
<td>step 3.</td>
<td>[ \tilde{w}<em>{jk} \propto \rho_k \exp(x</em>{kj} \log \frac{\gamma_k}{\gamma}) ]</td>
</tr>
<tr>
<td>step 3’.</td>
<td>[ w_{jk} = \operatorname{arg\ max}<em>k \tilde{w}</em>{jk} ]</td>
</tr>
<tr>
<td>step 4.</td>
<td>[ \rho_k = \frac{\tilde{w}<em>k}{d}, \gamma_k = \frac{\sum_j \tilde{w}</em>{jk} \bar{Z} x_{kj}}{x_k \bar{Z} \bar{W} \gamma_k \gamma_{kk}}, \gamma = \frac{N - \sum_k x_{kk}}{N^2 - \sum_k x_k \bar{Z} \bar{W} \gamma_k \gamma_{kk}} ]</td>
</tr>
<tr>
<td>until Convergence;</td>
<td></td>
</tr>
<tr>
<td>Output :</td>
<td>Z, W, π_k, ρ_k, γ_kk, γ</td>
</tr>
</tbody>
</table>

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)

**SPLBsem and SPLBcem are very dependant on their starting points!**

**Algorithm 5: SPLBsem**

**Input:** $X, g$

**Initialization:** $	ilde{Z}, 	ilde{W}, \pi_k, \rho_k, \gamma_{kk}, \gamma$

repeat

step 1. $\tilde{z}_{ik} \propto \pi_k \exp(\tilde{w}_{ik} \log \gamma_{kk})$

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

step 2. $\pi_k = \frac{\tilde{z}_{ik}}{n}$, $\gamma_{kk} = \frac{\sum\tilde{z}_{ik}\tilde{w}_{ik}}{\tilde{Z}_{ik} \tilde{W}_{ik}}$, $\gamma = \frac{N - \sum_k \tilde{z}_{ik}\tilde{W}_{ik}}{N^2 - \sum_k \tilde{Z}_{ik} \tilde{W}_{ik}}$

step 3. $\tilde{w}_{jk} \propto \rho_k \exp(\tilde{z}_{kj} \log \gamma_{kk})$

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

step 4. $\rho_k = \frac{\tilde{w}_{jk}}{d}$, $\gamma_{kk} = \frac{\sum\tilde{w}_{jk}\tilde{z}_{kj}}{\tilde{Z}_{kj} \tilde{W}_{kj}}$, $\gamma = \frac{N - \sum_k \tilde{z}_{ik}\tilde{W}_{ik}}{N^2 - \sum_k \tilde{Z}_{ik} \tilde{W}_{ik}}$

until Convergence;

**Output:** $Z, W, \pi_k, \rho_k, \gamma_{kk}, \gamma$
The stochastic SPLBM-based co-clustering algorithm (SPLBsem)

Algorithm 5: SPLBsem

Input : $X, g$
Initialization : $\tilde{Z}, \tilde{W}, \pi_k, \rho_k, \gamma_{kk}, \gamma$
repeat
$x_{ik} \tilde{W} = \sum_j \tilde{w}_{jk} x_{ij}$

step 1. $\tilde{z}_{ik} \propto \pi_k \exp(x_{ik} \log \gamma_{kk})$
step 1’. simulation of $z_i$ according to $M(\tilde{z}_{i1}, \ldots, \tilde{z}_{ig})$

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

$x_{kj} \tilde{Z} = \sum_i \tilde{z}_{ik} x_{ij}$

step 3. $\tilde{w}_{jk} \propto \rho_k \exp(x_{kj} \log \gamma_{kk})$
step 3’. simulation of $w_j$ according to $M(\tilde{w}_{j1}, \ldots, \tilde{w}_{jg})$

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

until Convergence;
Output : $Z, W, \pi_k, \rho_k, \gamma_{kk}, \gamma$
The stochastic SPLBM-based co-clustering algorithm (SPLBsem)

**Algorithm 5: SPLBsem**

*Input*: $X, g$

*Initialization*: $\tilde{Z}, \tilde{W}, \pi_k, \rho_k, \gamma_{kk}, \gamma$

repeat

$x_{ik} = \sum_j \tilde{w}_{jk} x_{ij}$

**step 1.** $\tilde{z}_{ik} \propto \pi_k \exp(x_{ik} \log \gamma_{kk} / \gamma)$

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

$\pi_k = \frac{\tilde{z}_{.k}}{n}, \gamma_{kk} = \frac{\sum_i \tilde{z}_{ik} x_{ik}}{x_{.k} \tilde{w}_{.k} \tilde{w}} = \frac{x_{kk}}{x_{.k} \tilde{w}_{.k}}, \gamma = \frac{N - \sum_k x_{kk}}{N^2 - \sum_k x_{.k} \tilde{w}_{.k}}$

$x_{kj} = \sum_i \tilde{z}_{ik} x_{ij}$

**step 2.** $\tilde{w}_{jk} \propto \rho_k \exp(x_{kj} \log \gamma_{kk} / \gamma)$

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

$\rho_k = \frac{\tilde{w}_{.k}}{d}, \gamma_{kk} = \frac{\sum_j \tilde{w}_{jk} x_{kj}}{x_{.k} \tilde{w}_{.k} \tilde{w}} = \frac{x_{kk}}{x_{.k} \tilde{w}_{.k}}, \gamma = \frac{N - \sum_k x_{kk}}{N^2 - \sum_k x_{.k} \tilde{w}_{.k}}$

until Convergence;

*Output*: $Z, W, \pi_k, \rho_k, \gamma_{kk}, \gamma$

**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:** $X, g$

**Initialization:** $\bar{Z}, \bar{W}, \pi_k, \rho_k, \gamma_{kk}, \gamma$

repeat

$x_{ik}^{\bar{W}} = \sum_j \bar{W}_{jk} x_{ij}$

**step 1.** $z_{ik} \sim \pi_k \exp\left(x_{ik}^{\bar{W}} \log \frac{\gamma_{kk}}{\gamma}\right)$

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

$\pi_k = \frac{z_{ik}}{n}, \gamma_{kk} = \frac{\sum_i z_{ik} x_{ik}}{x_{k.} x_{.k}} = \frac{x_{kk}}{x_{k.} x_{.k}}, \gamma = \frac{N - \sum_k x_{kk}}{N^2 - \sum_k x_{k.} x_{.k}}$

$x_{kj} = \sum_i z_{ik} x_{ij}$

**step 2.** $\bar{w}_{jk} \sim \rho_k \exp\left(\bar{z}_{kj} \log \frac{\gamma_{kk}}{\gamma}\right)$

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

$\rho_k = \frac{w_{jk}}{d}, \gamma_{kk} = \frac{\sum_j w_{jk} x_{kj}}{x_{k.} x_{.k}} = \frac{x_{kk}}{x_{k.} x_{.k}}, \gamma = \frac{N - \sum_k x_{kk}}{N^2 - \sum_k x_{k.} x_{.k}}$

until Convergence;

**Output:** $Z, W, \pi_k, \rho_k, \gamma_{kk}, \gamma$

**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: \( X, g \)

Initialization: \( \tilde{Z}, \tilde{W}, \pi_k, \rho_k, \gamma_{kk}, \gamma \)

repeat

\[
\tilde{x}_{ik} = \sum_j \tilde{w}_{jk} x_{ij}
\]

**step 1.** \( \tilde{z}_{ik} \propto \pi_k \exp (\tilde{x}_{ik} \log \gamma_{kk}) \)

**step 1’.** simulation of \( z_i \) according to \( M(\tilde{z}_{i1}, \ldots, \tilde{z}_{ig}) \)

\[
\pi_k = \frac{\tilde{z}_{i1} \cdot \ldots \cdot \tilde{z}_{ig}}{n}, \quad \gamma_{kk} = \frac{\sum_i \tilde{z}_{ik} \cdot \tilde{x}_{ik}}{\tilde{Z} \cdot \tilde{W} \cdot x_{1.} \cdot k} = \frac{\tilde{x}_{kk}}{\tilde{Z} \cdot \tilde{W} \cdot x_{1.} \cdot k}, \quad \gamma = \frac{N - \sum_k \tilde{x}_{kk}}{N^2 - \sum_k \tilde{x}_{1.} \cdot k \cdot \tilde{x}_{k.} \cdot k}
\]

\[
x_{kj} = \sum_i \tilde{z}_{ik} x_{ij}
\]

**step 2.** \( \tilde{w}_{jk} \propto \rho_k \exp (\tilde{z}_{kj} \log \gamma_{kk}) \)

**step 3’.** simulation of \( w_j \) according to \( M(\tilde{w}_{j1}, \ldots, \tilde{w}_{jg}) \)

\[
\rho_k = \frac{\tilde{w}_{j1} \cdot \ldots \cdot \tilde{w}_{jg}}{d}, \quad \gamma_{kk} = \frac{\sum_j \tilde{w}_{jk} \cdot \tilde{Z} \cdot \tilde{W} \cdot x_{1.} \cdot k}{\tilde{Z} \cdot \tilde{W} \cdot x_{1.} \cdot k} = \frac{\tilde{x}_{kk}}{\tilde{Z} \cdot \tilde{W} \cdot x_{1.} \cdot k}, \quad \gamma = \frac{N - \sum_k \tilde{x}_{kk}}{N^2 - \sum_k \tilde{x}_{1.} \cdot k \cdot \tilde{x}_{k.} \cdot k}
\]

until Convergence;

Output: \( Z, W, \pi_k, \rho_k, \gamma_{kk}, \gamma \)

**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 ⇒ SPLBvem
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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#Documents</th>
<th>#Words</th>
<th>#Clusters</th>
<th>Sparsity (%)</th>
<th>Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPORTS</td>
<td>8580</td>
<td>14870</td>
<td>7</td>
<td>99.14</td>
<td>0.036</td>
</tr>
<tr>
<td>TDT2</td>
<td>9394</td>
<td>36771</td>
<td>30</td>
<td>99.64</td>
<td>0.028</td>
</tr>
<tr>
<td>Yahoo_K1B</td>
<td>2340</td>
<td>21839</td>
<td>6</td>
<td>99.41</td>
<td>0.043</td>
</tr>
<tr>
<td>Reuters40</td>
<td>8203</td>
<td>18914</td>
<td>40</td>
<td>99.75</td>
<td>0.003</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#Documents</th>
<th>#Words</th>
<th>#Clusters</th>
<th>Sparsity (%)</th>
<th>Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPORTS</td>
<td>8580</td>
<td>14870</td>
<td>7</td>
<td>99.14</td>
<td>0.036</td>
</tr>
<tr>
<td>TDT2</td>
<td>9394</td>
<td>36771</td>
<td>30</td>
<td>99.64</td>
<td>0.028</td>
</tr>
<tr>
<td>Yahoo_K1B</td>
<td>2340</td>
<td>21839</td>
<td>6</td>
<td>99.41</td>
<td>0.043</td>
</tr>
<tr>
<td>Reuters40</td>
<td>8203</td>
<td>18914</td>
<td>40</td>
<td>99.75</td>
<td>0.003</td>
</tr>
</tbody>
</table>

### 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)

### Data
- Data: contingency tables

### Evaluation measures
- Acc
- NMI (Strehl and Ghosh, 2003)
- ARI (Rand, 1971)

### Table: Global Performance Comparison - Document Clustering

<table>
<thead>
<tr>
<th>Datasets</th>
<th>per.</th>
<th>Skmeans</th>
<th>ITCC</th>
<th>LDA</th>
<th>PLBvem</th>
<th>CoClus</th>
<th>SPLBcem</th>
<th>SPLBvem</th>
<th>SPLBsem</th>
<th>SPLBsvem</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPORTS</td>
<td>Acc</td>
<td>0.49</td>
<td>0.53</td>
<td>0.53</td>
<td>0.47</td>
<td>0.75</td>
<td>0.85</td>
<td>0.85</td>
<td>0.86</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.50</td>
<td>0.60</td>
<td>0.54</td>
<td>0.64</td>
<td>0.62</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>ARI</td>
<td>0.30</td>
<td>0.44</td>
<td>0.33</td>
<td>0.49</td>
<td>0.55</td>
<td>0.76</td>
<td>0.75</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>TDT2</td>
<td>Acc</td>
<td>0.57</td>
<td>0.59</td>
<td>0.60</td>
<td>0.59</td>
<td>0.87</td>
<td>0.83</td>
<td>0.84</td>
<td>0.84</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.76</td>
<td>0.78</td>
<td>0.73</td>
<td>0.76</td>
<td>0.84</td>
<td>0.81</td>
<td>0.82</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>ARI</td>
<td>0.46</td>
<td>0.52</td>
<td>0.49</td>
<td>0.51</td>
<td>0.85</td>
<td>0.81</td>
<td>0.80</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Yahoo_K1B</td>
<td>Acc</td>
<td>0.57</td>
<td>0.61</td>
<td>0.62</td>
<td>0.58</td>
<td>0.60</td>
<td>0.79</td>
<td>0.84</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.64</td>
<td>0.58</td>
<td>0.58</td>
<td>0.62</td>
<td>0.54</td>
<td>0.66</td>
<td>0.69</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>ARI</td>
<td>0.39</td>
<td>0.40</td>
<td>0.37</td>
<td>0.38</td>
<td>0.31</td>
<td>0.60</td>
<td>0.72</td>
<td>0.76</td>
<td>0.79</td>
</tr>
<tr>
<td>Reuters40</td>
<td>Acc</td>
<td>0.26</td>
<td>0.27</td>
<td>0.47</td>
<td>0.25</td>
<td>0.61</td>
<td>0.73</td>
<td>0.74</td>
<td>0.73</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.50</td>
<td>0.52</td>
<td>0.51</td>
<td>0.52</td>
<td>0.54</td>
<td>0.57</td>
<td>0.58</td>
<td>0.57</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>ARI</td>
<td>0.11</td>
<td>0.18</td>
<td>0.42</td>
<td>0.15</td>
<td>0.51</td>
<td>0.71</td>
<td>0.75</td>
<td>0.73</td>
<td>0.76</td>
</tr>
</tbody>
</table>

- 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

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

- 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 $\gamma_{kk}$’s (left) and $\gamma$ (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
• 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
• 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
**Comparison of the standard deviation in cluster size (SDCS) of clusters obtained by each method**

SDCS = \( \left( \frac{1}{g-1} \sum_{k=1}^{g} \left( z_{g,k} - \frac{n}{g} \right)^2 \right)^{0.5} \)

• The SDCS values of the clusters obtained with SPLBcem are the closest to the real SDCS of the datasets.

<table>
<thead>
<tr>
<th>Data</th>
<th>Clustering</th>
<th>Co-clustering</th>
<th></th>
<th>Diagonal</th>
<th>Real SDCS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Non-diagonal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skmeans</td>
<td>ITCC</td>
<td>LDA</td>
<td>PLBcem</td>
</tr>
<tr>
<td>REUTERS40</td>
<td>112.638</td>
<td>144.195</td>
<td>362.102</td>
<td>201.162</td>
<td></td>
</tr>
<tr>
<td>REUTERS30</td>
<td>161.797</td>
<td>238.353</td>
<td>414.568</td>
<td>261.291</td>
<td></td>
</tr>
<tr>
<td>K1B</td>
<td>154.3684</td>
<td>198.828</td>
<td>261.765</td>
<td>189.849</td>
<td>336.555</td>
</tr>
<tr>
<td>TDT2</td>
<td>154.143</td>
<td>216.152</td>
<td>189.609</td>
<td>235.698</td>
<td>516.685</td>
</tr>
<tr>
<td>SPORTS</td>
<td>760.099</td>
<td>346.066</td>
<td>482.714</td>
<td>393.510</td>
<td>1359.321</td>
</tr>
</tbody>
</table>

**Conclusion and Perspectives**

Using Co-clustering in Biomedical Text Mining Framework.

**Experiments**

- Graph-based Co-clustering
- Model-based Co-clustering
- Using Co-clustering in Biomedical Text Mining Framework

**References**

**Sparse Poisson Latent Block Model (SPLBM)**

- Soft SPLBM-based Co-clustering Algorithm
- Hard SPLBM-based Co-clustering Algorithm
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 $12648 \times 19518$ that contains documents about 5 different diseases.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migraine</td>
<td>3703</td>
</tr>
<tr>
<td>Age-related Macular Degeneration</td>
<td>3283</td>
</tr>
<tr>
<td>Otitis</td>
<td>2596</td>
</tr>
<tr>
<td>Kidney Calculi</td>
<td>1549</td>
</tr>
<tr>
<td>Hay Fever</td>
<td>1517</td>
</tr>
</tbody>
</table>
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 $12648 \times 19518$ that contains documents about 5 different diseases.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migraine</td>
<td>3703</td>
</tr>
<tr>
<td>Age-related Macular Degeneration</td>
<td>3283</td>
</tr>
<tr>
<td>Otitis</td>
<td>2596</td>
</tr>
<tr>
<td>Kidney Calculi</td>
<td>1549</td>
</tr>
<tr>
<td>Hay Fever</td>
<td>1517</td>
</tr>
</tbody>
</table>
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 $12648 \times 19518$ that contains documents about 5 different diseases.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migraine</td>
<td>3703</td>
</tr>
<tr>
<td>Age-related Macular Degeneration</td>
<td>3283</td>
</tr>
<tr>
<td>Otitis</td>
<td>2596</td>
</tr>
<tr>
<td>Kidney Calculi</td>
<td>1549</td>
</tr>
<tr>
<td>Hay Fever</td>
<td>1517</td>
</tr>
</tbody>
</table>
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 $S_c = X_c^{\text{norm}} X_c^{\text{norm}}$ 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".
Sparse Poisson Latent Block Model (SPLBM)
Soft SPLBM-based Co-clustering Algorithm
Hard SPLBM-based Co-clustering Algorithm
Experiments

47 / 66
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 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.
Sparse Poisson Latent Block Model (SPLBM)
Soft SPLBM-based Co-clustering Algorithm
Hard SPLBM-based Co-clustering Algorithm
Experiments

(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

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

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

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

Input: set of 10 asthma genes \((G)\) and 100 sets of random genes \(\{R_1, \ldots, R_{100}\}\) 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 \((G + R_i)\) (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 (G) and 100 sets of random genes \( \{R_1, \ldots, R_{100}\} \) 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 \((G + R_i)\) (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 (G) and 100 sets of random genes \(\{R_1, \ldots, R_{100}\}\) selected randomly from the human genome.

2. **Assess the strength of association between asthma-associated genes**
   - Use the PUBMED database to create a \(\text{gene} \times \text{term}\) 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 \(\text{gene} \times \text{term}\) matrix for each set \((G + R_i)\) (100 matrices).
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem.

4. **New candidate genes for asthma**
Input: set of 10 asthma genes (G) and 100 sets of random genes \{R_1, \ldots, 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 × 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 × term matrix for each set (G + R_i) (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem

4. New candidate genes for asthma
Input: set of 10 asthma genes (G) and 100 sets of random genes \( \{R_1, \ldots, R_{100}\} \) selected randomly from the human genome.

2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a gene × 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 × term matrix for each set \( (G + R_i) \) (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 (G) and 100 sets of random genes \( \{ R_1, \ldots, R_{100} \} \) 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 \((G + R_i)\) (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 (G) and 100 sets of random genes \{R_1, \ldots, 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 × 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 × term matrix for each set (G + R_i) (100 matrices)
   - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem

4. New candidate genes for asthma
Input: set of 10 asthma genes (G) and 100 sets of random genes \( \{R_1, \ldots, R_{100}\} \) selected randomly from the human genome

2. Assess the strength of association between asthma-associated genes
   - Use the PUBMED database to create a gene × 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 × term matrix for each set \((G + R_i)\) (100 matrices)
     - Clustering (Zhao and K, 2002) and Co-clustering with Coclus and SPLBcem

4. New candidate genes for asthma
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

<table>
<thead>
<tr>
<th>Smoking</th>
<th>diabetes</th>
<th>th2</th>
<th>environmental</th>
</tr>
</thead>
<tbody>
<tr>
<td>immune-mediated</td>
<td>chronic</td>
<td>enterotoxin</td>
<td>proinflammatory</td>
</tr>
<tr>
<td>child</td>
<td>microenvironment</td>
<td>cytokine</td>
<td>autoimmune</td>
</tr>
<tr>
<td>immunohistochemistry</td>
<td>childhood</td>
<td>influenza</td>
<td>asthma</td>
</tr>
<tr>
<td>drug</td>
<td>inflammation</td>
<td>crohn</td>
<td>necrosis</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Smoking</th>
<th>diabetes</th>
<th>th2</th>
<th>environmental</th>
</tr>
</thead>
<tbody>
<tr>
<td>immune-mediated</td>
<td>chronic</td>
<td>enterotoxin</td>
<td>proinflammatory</td>
</tr>
<tr>
<td>child</td>
<td>microenvironment</td>
<td>cytokine</td>
<td>autoimmune</td>
</tr>
<tr>
<td>immunohistochemistry</td>
<td>childhood</td>
<td>influenza</td>
<td>asthma</td>
</tr>
<tr>
<td>drug</td>
<td>inflammation</td>
<td>crohn</td>
<td>necrosis</td>
</tr>
</tbody>
</table>
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 (G)
- Study the Top 20 genes
- Use the biomedical literature and experts to validate the results

<table>
<thead>
<tr>
<th>IL1RL1</th>
<th>RAG1</th>
<th>CLEC1B</th>
<th>IL23R</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAT6</td>
<td>EFNA3</td>
<td>S1PR5</td>
<td>TGFBR1</td>
</tr>
<tr>
<td>FCMR</td>
<td>CXCL8/IL8</td>
<td>CHRN5B4</td>
<td>NFKB1</td>
</tr>
<tr>
<td>TNFRSF1A</td>
<td>TMED1</td>
<td>NOD2</td>
<td>TSLP</td>
</tr>
<tr>
<td>NLRP10</td>
<td>POMP</td>
<td>SPINK1</td>
<td>PTGES</td>
</tr>
</tbody>
</table>

- 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 (G).
- Study the Top 20 genes.
- Use the biomedical literature and experts to validate the results.

<table>
<thead>
<tr>
<th>IL1RL1</th>
<th>RAG1</th>
<th>CLEC1B</th>
<th>IL23R</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAT6</td>
<td>EFNA3</td>
<td>S1PR5</td>
<td>TGFBR1</td>
</tr>
<tr>
<td>FCMR</td>
<td>CXCL8/IL8</td>
<td>CHRN B4</td>
<td>NFKB1</td>
</tr>
<tr>
<td>TNFRSF1A</td>
<td>TMED1</td>
<td>NOD2</td>
<td>TSLP</td>
</tr>
<tr>
<td>NLRP10</td>
<td>POMP</td>
<td>SPINK1</td>
<td>PTGES</td>
</tr>
</tbody>
</table>

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

Graph-based Co-clustering

Model-based Co-clustering

Using Co-clustering in Biomedical Text Mining Framework

Conclusion and Perspectives

References
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/coclusl)
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/coclus)
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.

---


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


References III


