Topic modeling as a community-detection problem

Yuanming Tao[1] Martin Gerlach[2 4] Tiago P. Peixoto [3 5] Eduardo G. Altmann[1]

[1]School of Mathematics and Statistics, University of Sydney, 2006 New South Wales, Australia.
[2]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA.
[3]Institute for Scientific Interchange Foundation, Via Alassio 11/c, 10126 Torino, Italy.
[4]Wikimedia Research Group, Berlin, Germany.
[5]Central European University, Budapest, Hungary Vienna, Austria.

NTU Singapore, 2019

THE UNIVERSITY OF SYDNEY

Centre for Complex Systems

Photo by Alfons Morales

Topic models

- **Discover** what the books are talking about *automatically*

Photo by Alfons Morales

Topic models

- **Discover** what the books are talking about *automatically*

Photo by Alfons Morales

Topic models

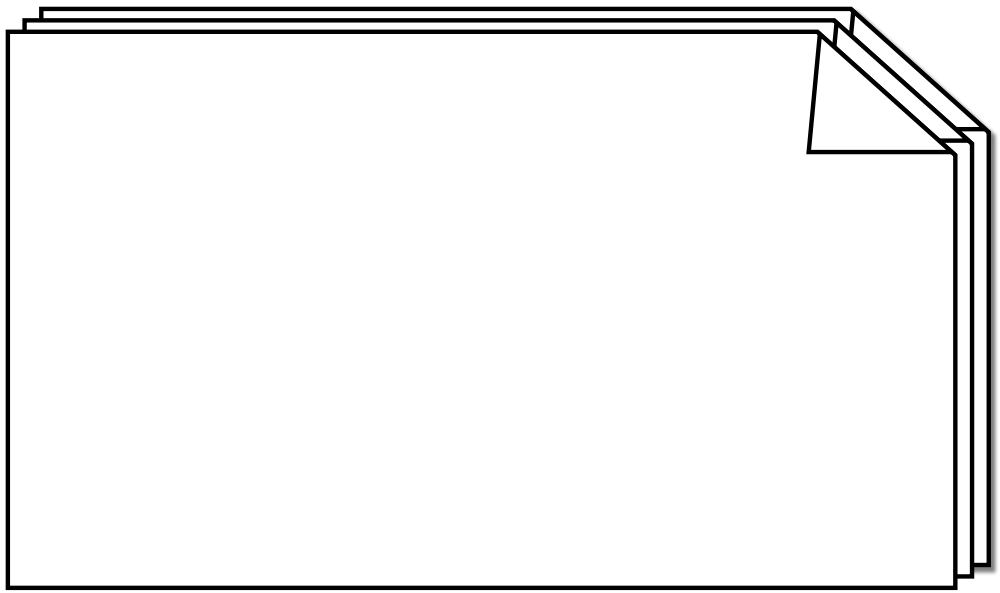
- **Discover** what the books are talking about *automatically*
- **Organize** the books onto bookshelves

Photo by Alfons Morales

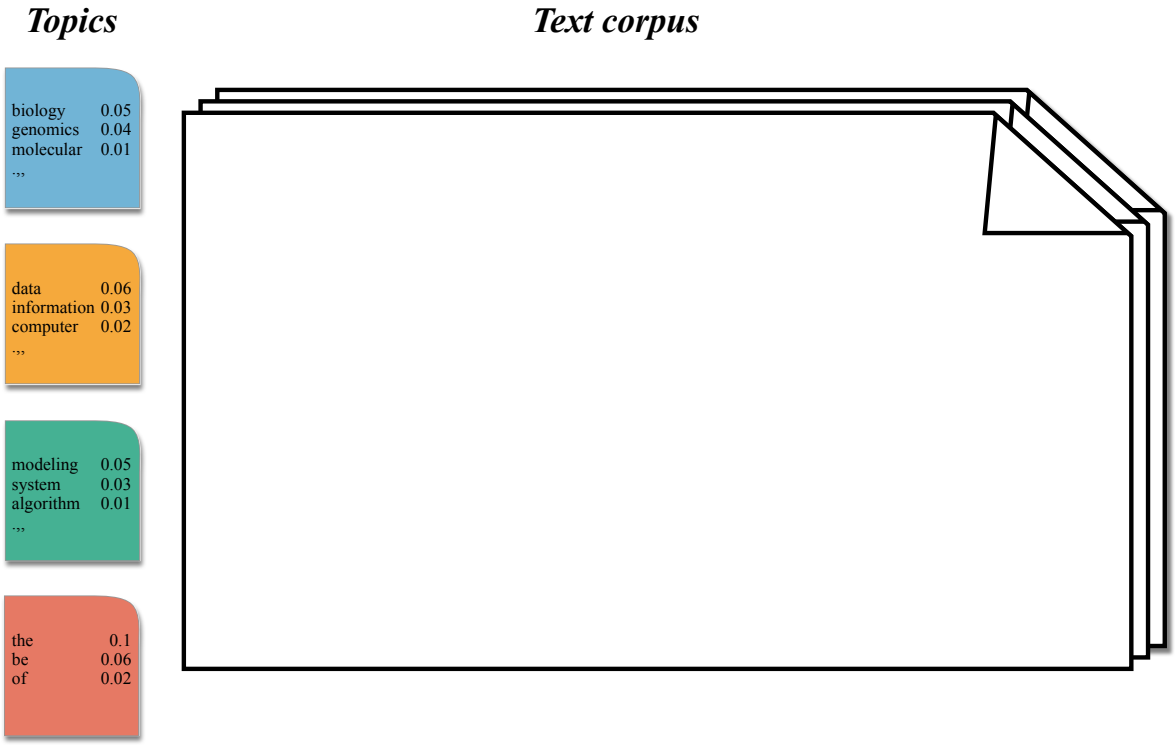
Topic modeling: A Concrete Example

Topic modeling: A Concrete Example

Text corpus

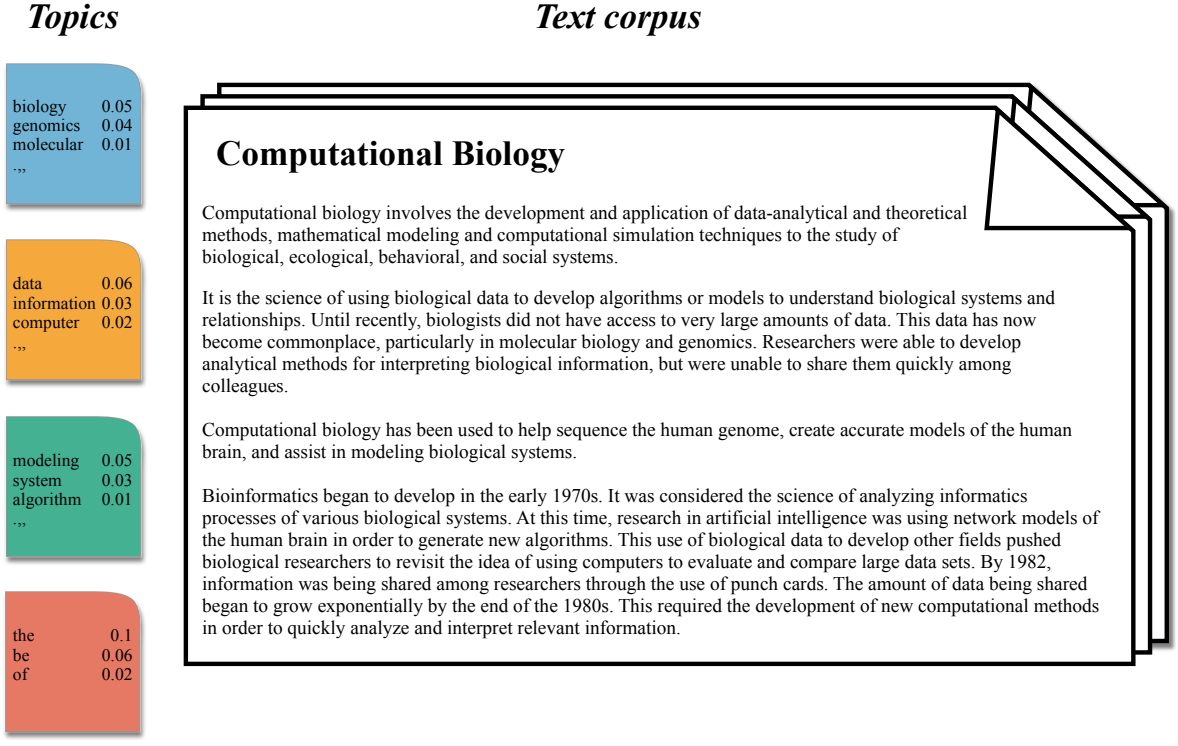


Topic modeling: A Concrete Example



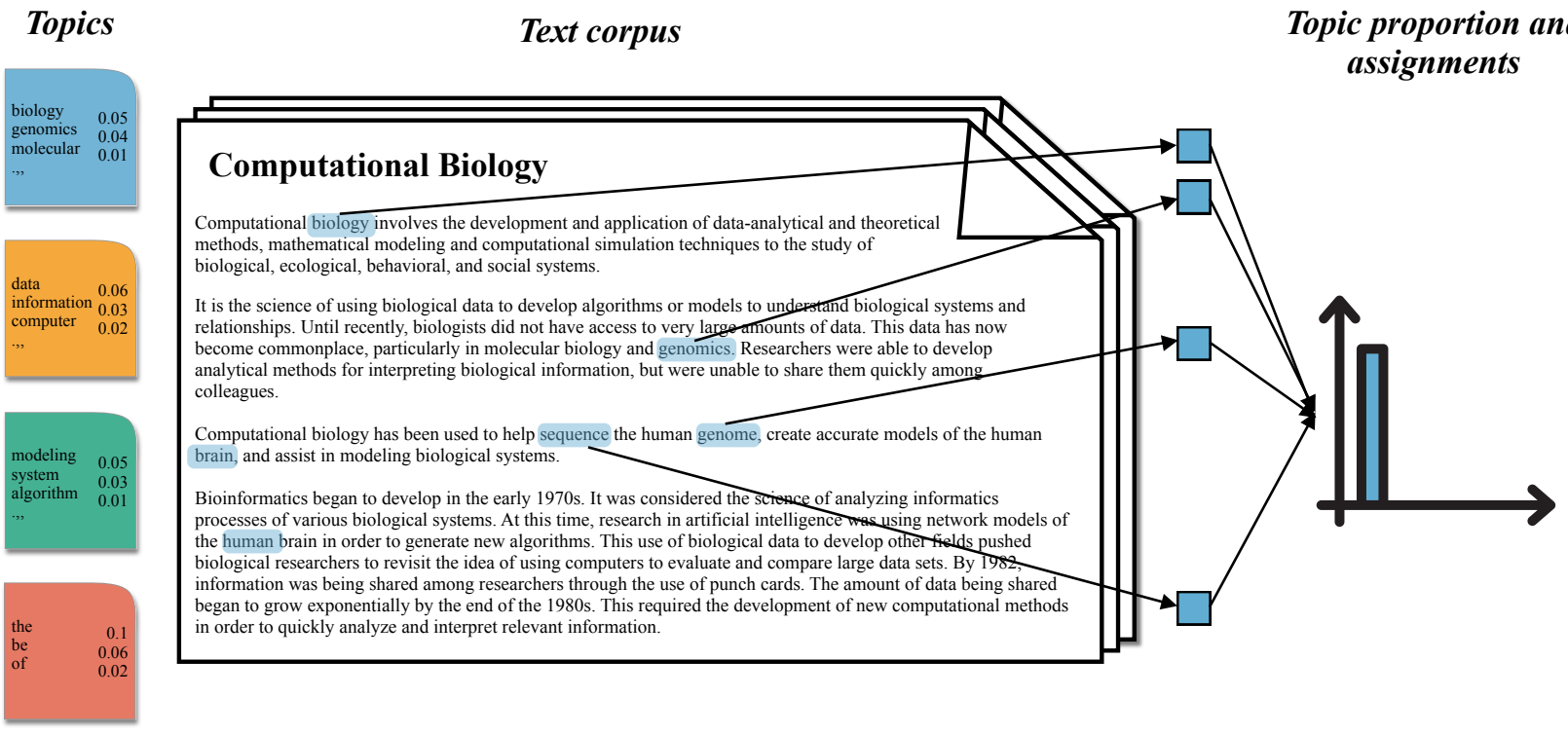
- Each **corpus-wide** topic is a cluster of words.

Topic modeling: A Concrete Example



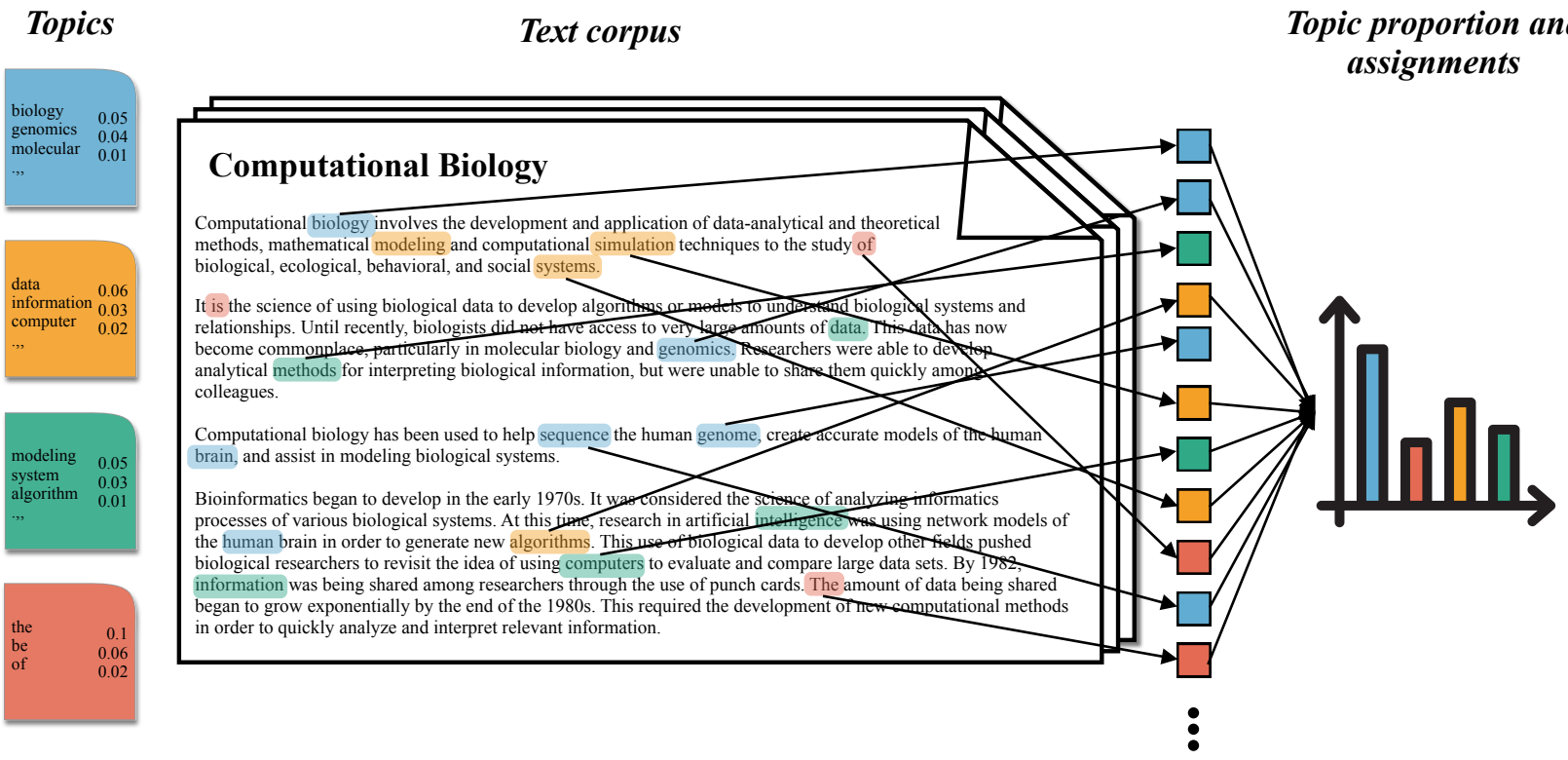
- Each **corpus-wide** topic is a cluster of words.

Topic modeling: A Concrete Example



- Each **corpus-wide** topic is a cluster of words.
- Each document is a distribution over topics.

Topic modeling: A Concrete Example



- Each **corpus-wide** topic is a cluster of words.
- Each document is a distribution over topics.

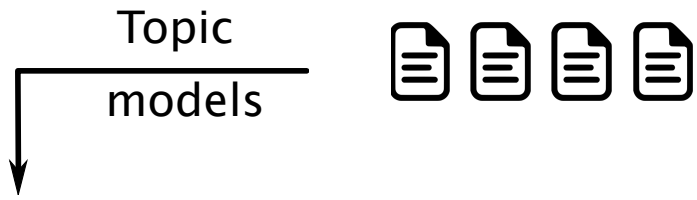
Literature Review

D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, p. 77, Apr. 2012.
M.Gerlach, T.P.Peixoto., E.G.Altmann. Sci Adv 4, 7 (2018)

Literature Review

D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, p. 77, Apr. 2012.
M.Gerlach, T.P.Peixoto., E.G.Altmann. Sci Adv 4, 7 (2018)

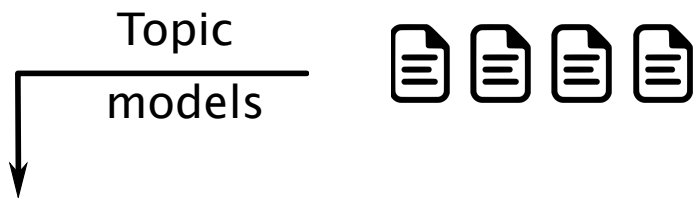
Literature Review

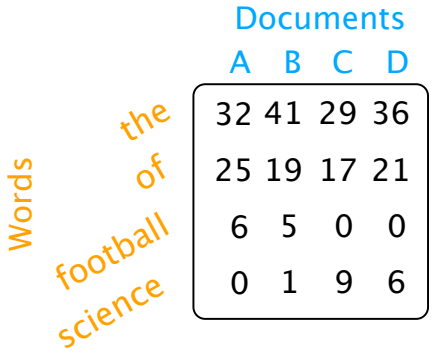


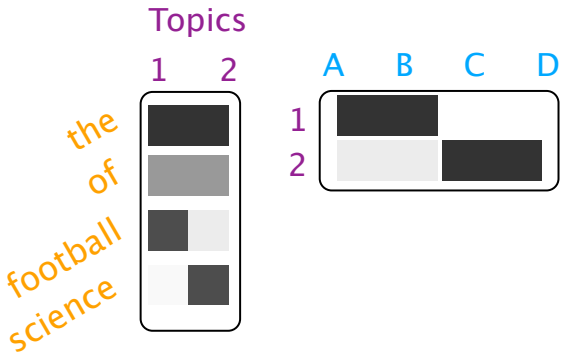
Words	Documents A	B	C	D
the	32	41	29	36
of	25	19	17	21
football	6	5	0	0
science	0	1	9	6

D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, p. 77, Apr. 2012.
M.Gerlach, T.P.Peixoto., E.G.Altmann. Sci Adv 4, 7 (2018)

Literature Review

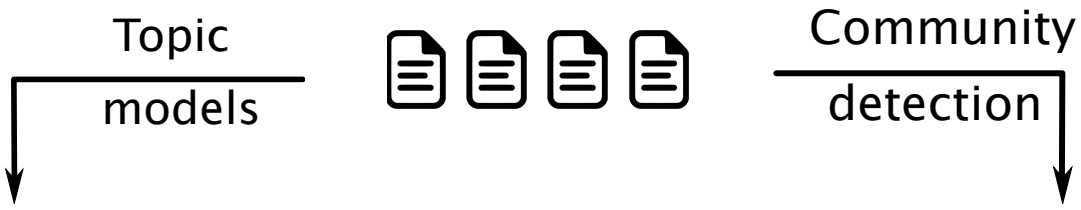


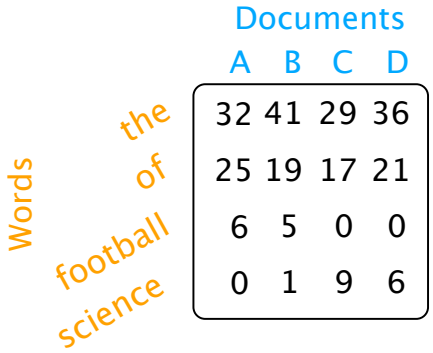


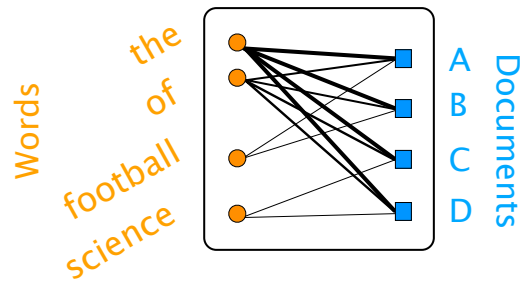


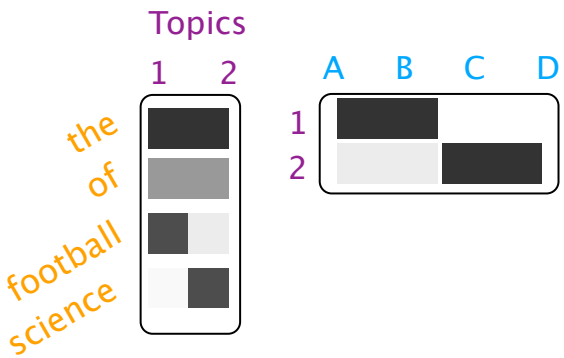
D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, p. 77, Apr. 2012.
M.Gerlach, T.P.Peixoto., E.G.Altmann. Sci Adv 4, 7 (2018)

Literature Review



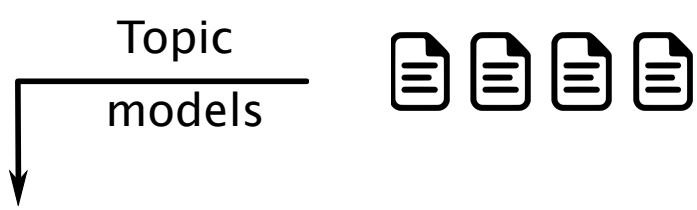


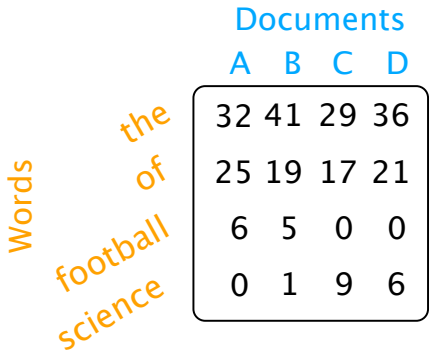


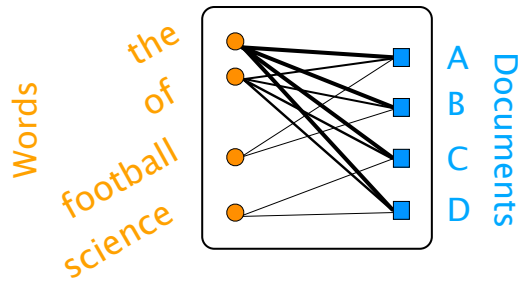


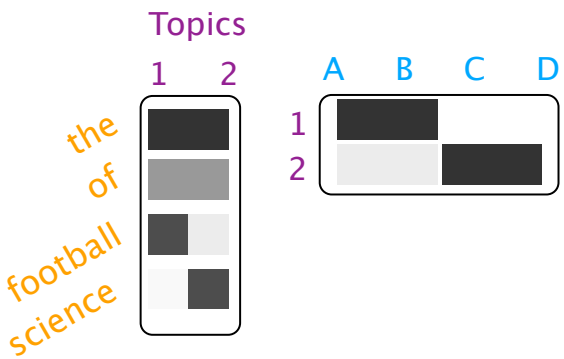
D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, p. 77, Apr. 2012.
M.Gerlach, T.P.Peixoto., E.G.Altmann. Sci Adv 4, 7 (2018)

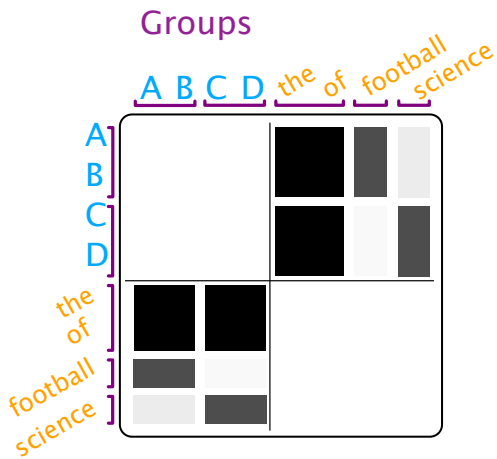
Literature Review











D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, p. 77, Apr. 2012.
M.Gerlach, T.P.Peixoto., E.G.Altmann. Sci Adv 4, 7 (2018)

Limitations: the above two models are *only* based on **word frequencies**.

The Research Problem

Can we incorporate **further information** about documents?

Example: A Wikipedia Article

Example: A Wikipedia Article

Computational biology

From Wikipedia, the free encyclopedia

Computational biology involves the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, ecological, behavioral, and social systems.[1] The field is broadly defined and includes foundations in biology, applied mathematics, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, computer science and evolution.[2]

Computational biology is different from biological computing, which is a subfield of computer science and computer engineering using bioengineering and biology to build computers, but is similar to bioinformatics, which is an interdisciplinary science using computers to store and process biological data.

Introduction

Computational Biology, which includes many aspects of bioinformatics, is the science of using biological data to develop algorithms or models to understand biological systems and relationships. Until recently, biologists did not have access to very large amounts of data. This data has now become commonplace, particularly in molecular biology and genomics. Researchers were able to develop analytical methods for interpreting biological information, but were unable to share them quickly among colleagues.[3]

Bioinformatics began to develop in the early 1970s. It was considered the science of analyzing informatics processes of various biological systems. At this time, research in artificial intelligence was using network models of the human brain in order to generate new algorithms. This use of biological data to develop other fields pushed biological researchers to revisit the idea of using computers to evaluate and compare large data sets. By 1982, information was being shared among researchers through the use of punch cards. The amount of data being shared began to grow exponentially by the end of the 1980s. This required the development of new computational methods in order to quickly analyze and interpret relevant information.[3]

Example: A Wikipedia Article

Computational biology

Connected to: Genetics | Molecular biology | Chemistry

From Wikipedia, the free encyclopedia

Computational biology involves the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, ecological, behavioral, and social systems.[1] The field is broadly defined and includes foundations in biology, applied mathematics, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, computer science and evolution.[2]

Computational biology is different from biological computing, which is a subfield of computer science and computer engineering using bioengineering and biology to build computers, but is similar to bioinformatics, which is an interdisciplinary science using computers to store and process biological data.

Introduction

Computational Biology, which includes many aspects of bioinformatics, is the science of using biological data to develop algorithms or models to understand biological systems and relationships. Until recently, biologists did not have access to very large amounts of data. This data has now become commonplace, particularly in molecular biology and genomics. Researchers were able to develop analytical methods for interpreting biological information, but were unable to share them quickly among colleagues.[3]

Bioinformatics began to develop in the early 1970s. It was considered the science of analyzing informatics processes of various biological systems. At this time, research in artificial intelligence was using network models of the human brain in order to generate new algorithms. This use of biological data to develop other fields pushed biological researchers to revisit the idea of using computers to evaluate and compare large data sets. By 1982, information was being shared among researchers through the use of punch cards. The amount of data being shared began to grow exponentially by the end of the 1980s. This required the development of new computational methods in order to quickly analyze and interpret relevant information.[3]

Articles can have user-generated labels, called **metadata tags**.

Example: A Wikipedia Article

Computational biology

Connected to: Genetics | Molecular biology | Chemistry

From Wikipedia, the free encyclopedia

Computational biology involves the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, ecological, behavioral, and social systems.[1] The field is broadly defined and includes foundations in biology, applied mathematics, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, computer science and evolution.[2]

Computationa… …eld of computer science and computer engineering u… …r to bioinformatics, which is an interdisciplina…

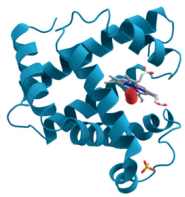
Introd…

Computationa… …science of using biological data to develop algorithms or… …l recently, biologists did not have access to very large a… …ularly in molecular biology and genomics. Researchers v… …l information, but were unable to share them quickly…

Bioinformatic… …ce of analyzing informatics processes of various biolog… …sing network models of the human brain in order to ge… …er fields pushed biological researchers to revisit the idea of using computers to evaluate and compare large data sets. By 1982, information was being shared among researchers through the use of punch cards. The amount of data being shared began to grow exponentially by the end of the 1980s. This required the development of new computational methods in order to quickly analyze and interpret relevant information.[3]

Molecular biology

Molecular biology is a branch of biology that concerns the molecular basis of biological activity between biomolecules in the various systems of a cell, including the interactions between DNA, RNA, proteins and their biosynthesis, as well as the regulation of these interactions.Writing in Nature in 1961, William Astbury described molecular biology as: ...not so much a technique as an approach, an approach from the viewpoint of the so-called basic sciences with the leading idea of searching below the large-scale manifestations of classical biology for the corresponding molecular plan. It is concerned particularly with the forms of biological molecules and [...] is predominantly three-dimensional and structural – which does not mean, however, that it is merely a refinement of morphology.



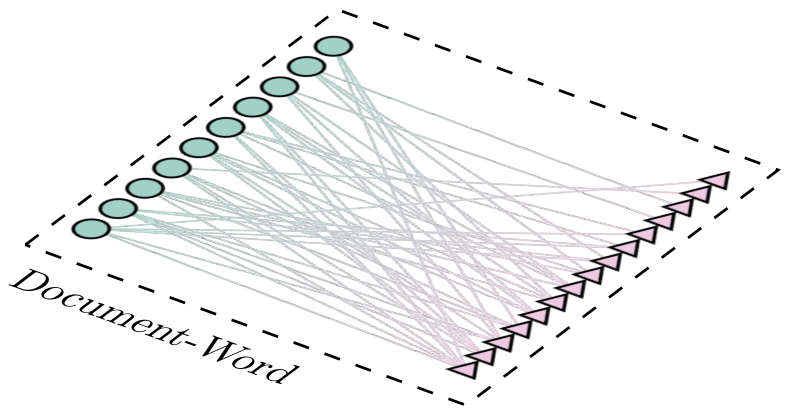
Pro tip: Ctrl+Click any word on any website for quick definition

Articles can contain external links to other articles, e.g. **hyperlinks**.

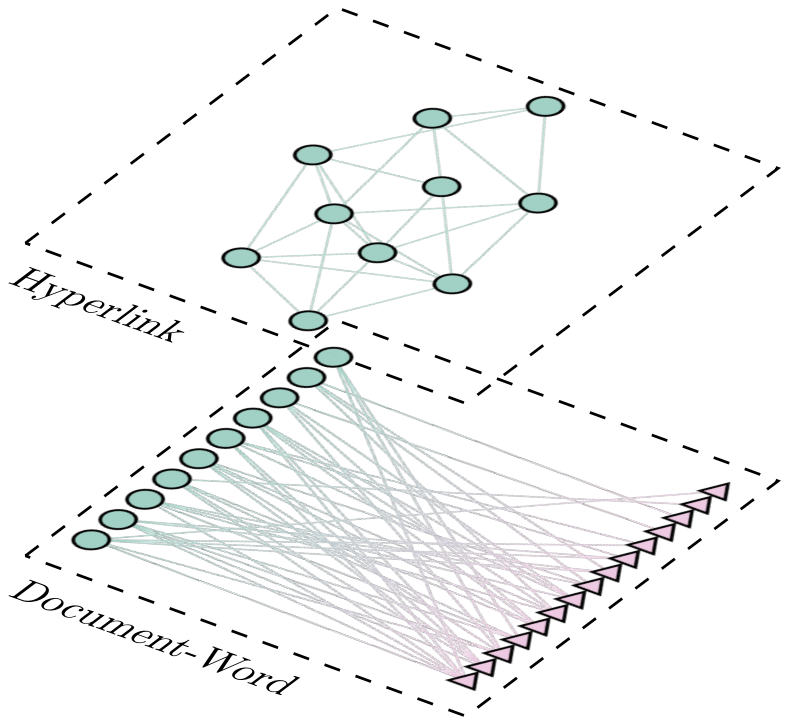
How do we incorporate **metadata tags** and **hyperlinks** in the model?

Modeling Texts with auxiliary information using multilayered network

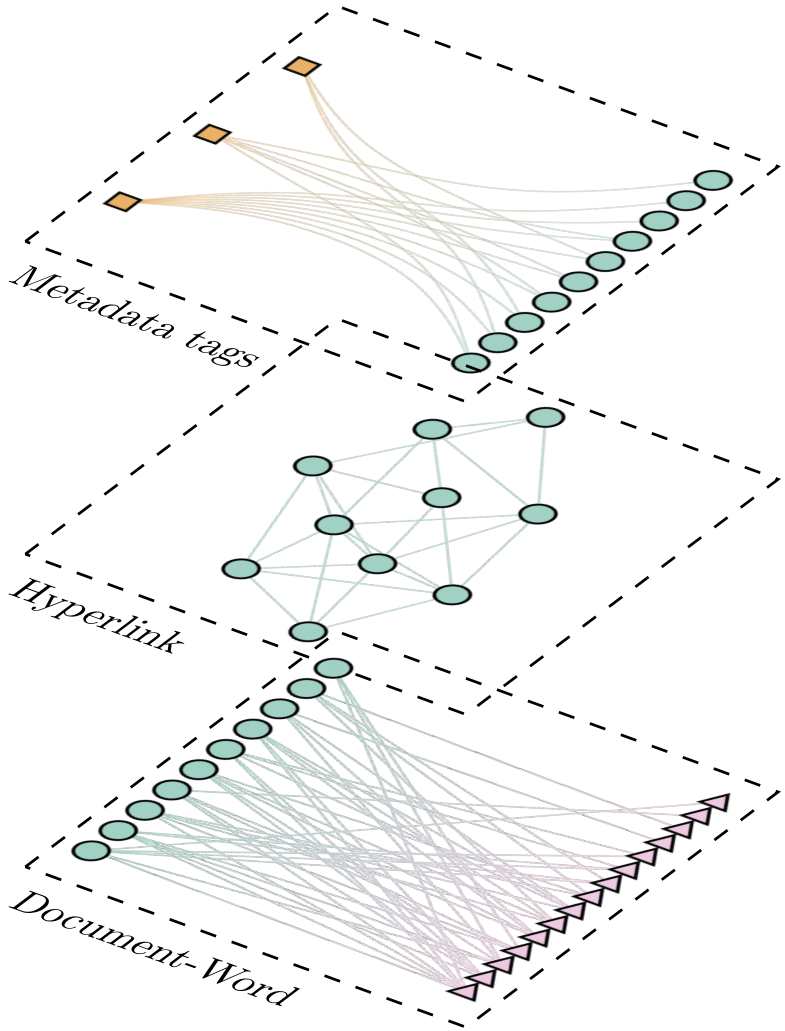
Modeling Texts with auxiliary information using multilayered network



Modeling Texts with auxiliary information using multilayered network



Modeling Texts with auxiliary information using multilayered network



Part II

What is the model for the formation of the above network structures?

The Stochastic Block Models (SBM)

P. W. Holland et al., Soc Networks 5, 109 (1983)

The Stochastic Block Models (SBM)

Planted partition: N nodes divided into B blocks.

P. W. Holland et al., Soc Networks 5, 109 (1983)

The Stochastic Block Models (SBM)

Planted partition: N nodes divided into B blocks.

Parameters:

$b_i \rightarrow$ block membership of node i

$p_{rs} \rightarrow$ probability of an edge between nodes of groups r and s

P. W. Holland et al., Soc Networks 5, 109 (1983)

The Stochastic Block Models (SBM)

Planted partition: N nodes divided into B blocks.

Parameters:

$b_i \rightarrow$ block membership of node i

$p_{rs} \rightarrow$ probability of an edge between nodes of groups r and s

P. W. Holland et al., Soc Networks 5, 109 (1983)

The Stochastic Block Models (SBM)

Planted partition: N nodes divided into B blocks.

Parameters:

$b_i \rightarrow$ block membership of node i

$p_{rs} \rightarrow$ probability of an edge between nodes of groups r and s

Likelihood: $P(A_{ij}|b_i, p_{b_i,b_j}) = \text{Poisson}\left(C \times p_{b_i,b_j}\right)$, where C is a constant.

P. W. Holland et al., Soc Networks 5, 109 (1983)

The Stochastic Block Models (SBM)

Planted partition: N nodes divided into B blocks.

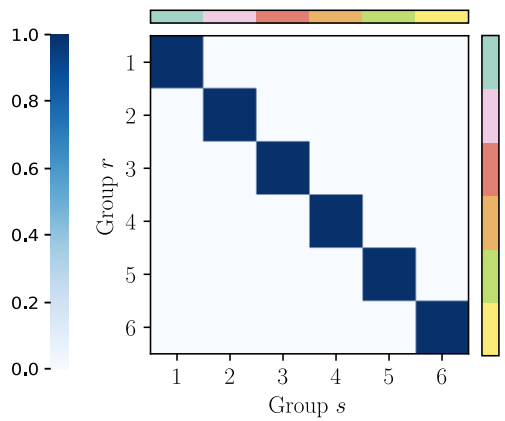
Parameters:

$b_i \rightarrow$ block membership of node i

$p_{rs} \rightarrow$ probability of an edge between nodes of groups r and s

Likelihood: $P(A_{ij}|b_i, p_{b_i,b_j}) = \text{Poisson}\left(C \times p_{b_i,b_j}\right)$, where C is a constant.

Example:



P. W. Holland et al., Soc Networks 5, 109 (1983)

The Stochastic Block Models (SBM)

Planted partition: N nodes divided into B blocks.

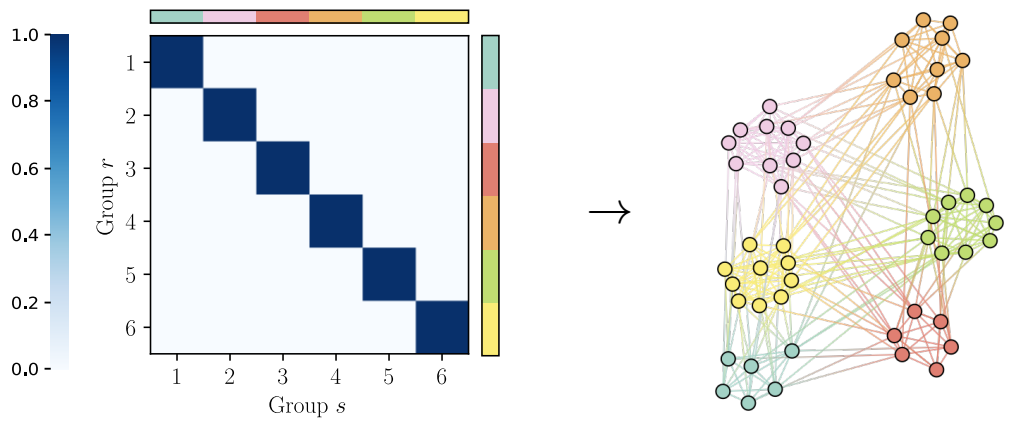
Parameters:

$b_i \rightarrow$ block membership of node i

$p_{rs} \rightarrow$ probability of an edge between nodes of groups r and s

Likelihood: $P(A_{ij}|b_i, p_{b_i,b_j}) = \text{Poisson}\left(C \times p_{b_i,b_j}\right)$, where C is a constant.

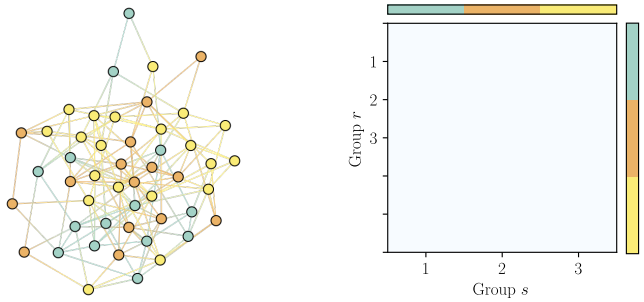
Example:



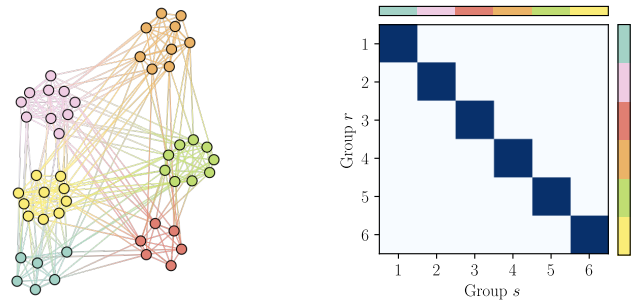
P. W. Holland et al., Soc Networks 5, 109 (1983)

The SBM is not restricted to assortative structures

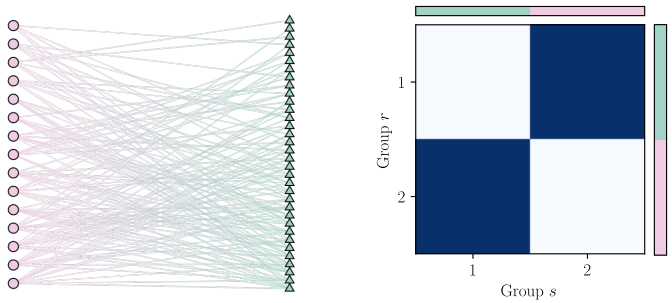
The SBM is not restricted to assortative structures



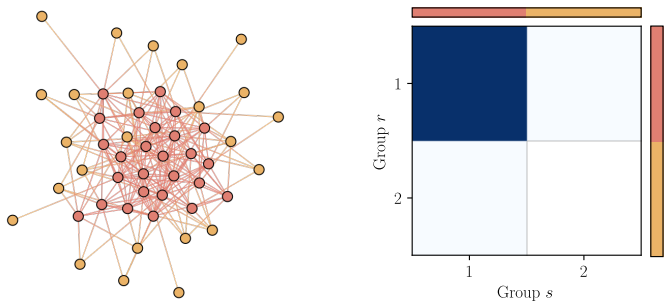
(a) The random network



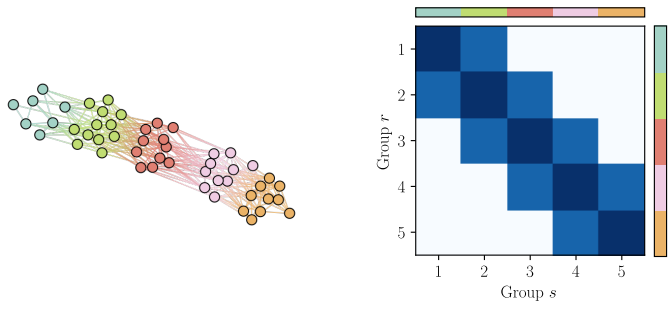
(b) The assortative structure



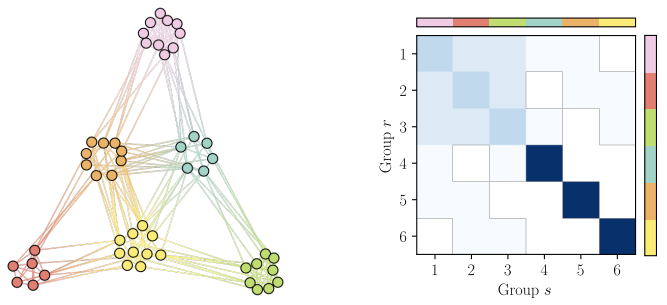
(c) The bipartite structure



(d) The core-periphery structure



(e) The ordered structure



(f) The mixed pattern

SBM with independent layers

Parameter:

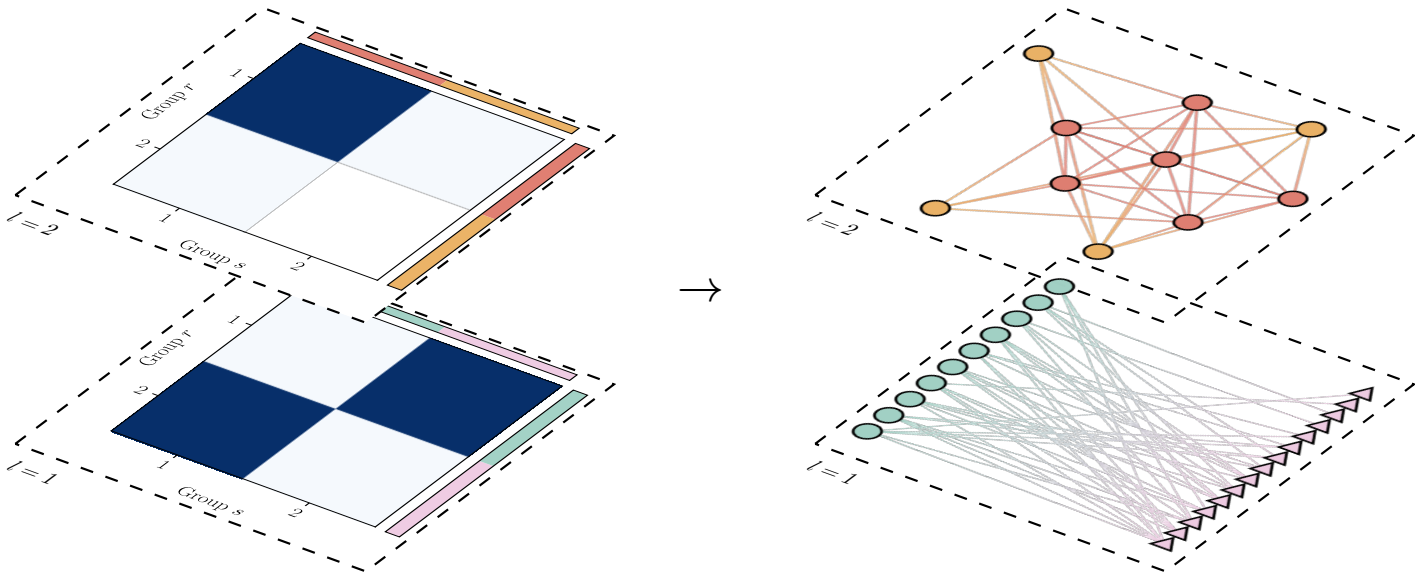
$z_{il} \rightarrow$ binary values to indicate whether a node i appears on layer l

SBM with independent layers

Parameter:

$z_{il} \rightarrow$ binary values to indicate whether a node i appears on layer l

Example:

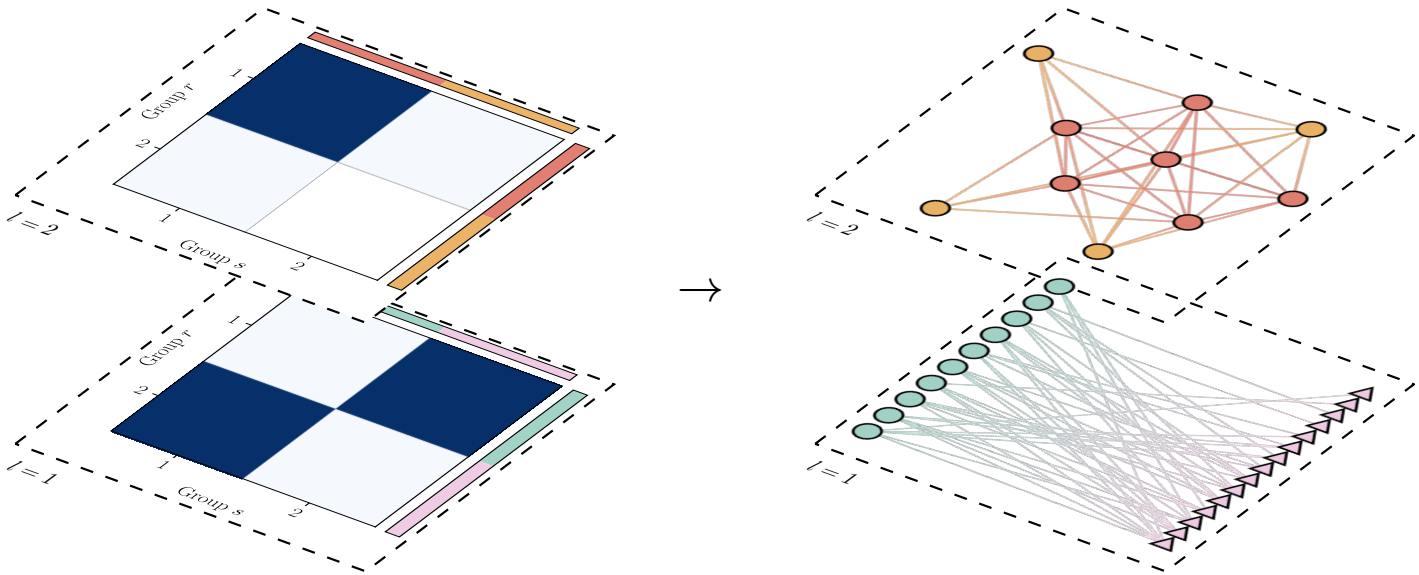


SBM with independent layers

Parameter:

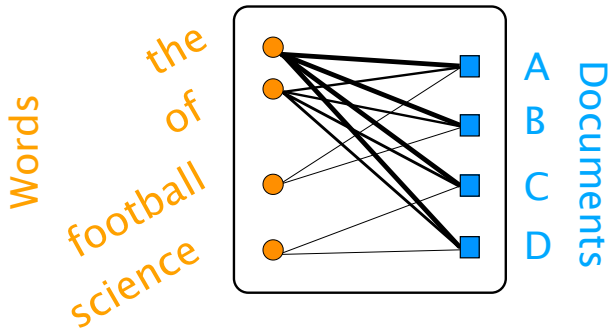
$z_{il} \rightarrow$ binary values to indicate whether a node i appears on layer l

Example:

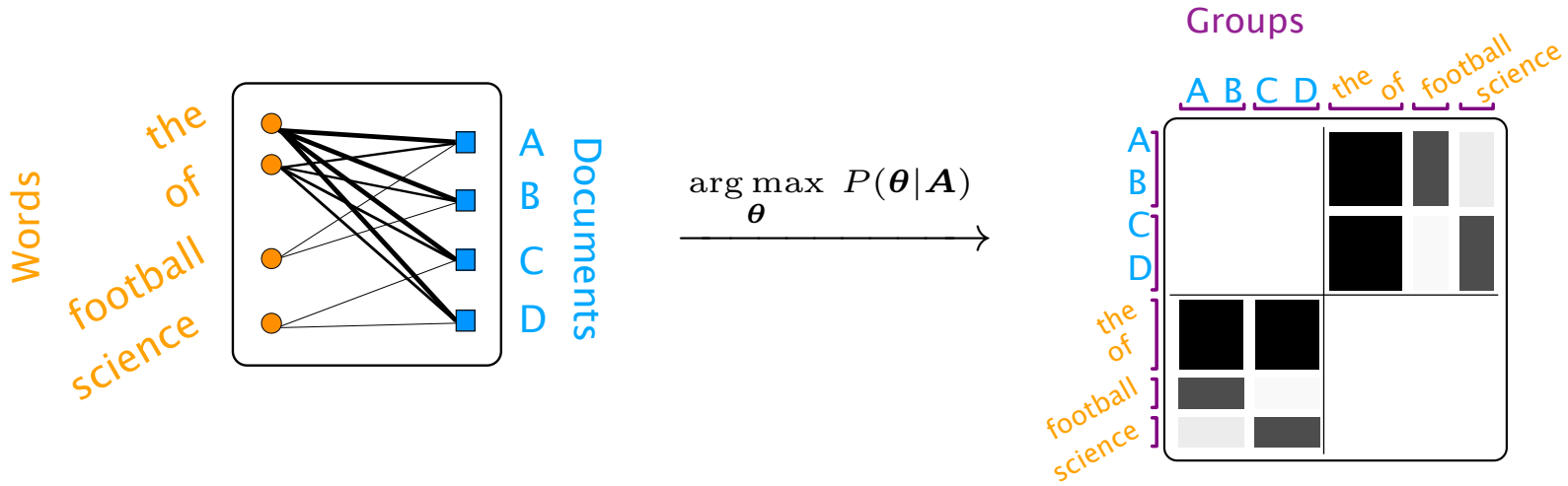


Important assumptions:

Degree distribution varies across different layers.

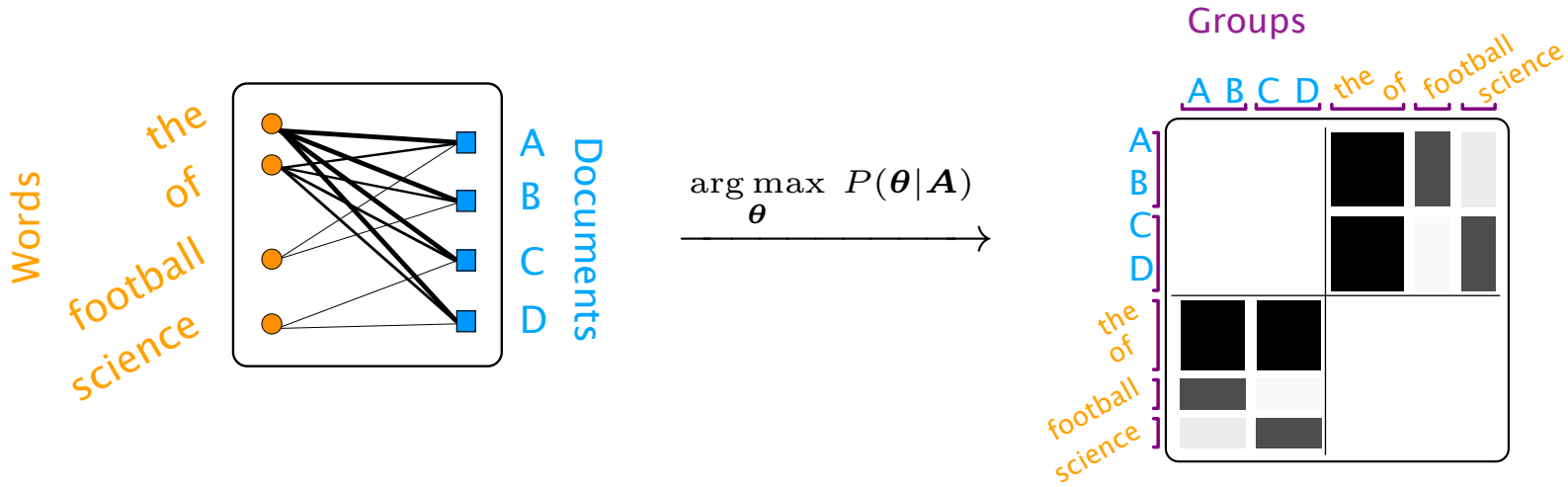

Words
the
of
football
science
A
B
C
D
Documents

The Bayesian Inference of the SBM



Grünwald (The Minimum Description Length Principle, 2007)

The Bayesian Inference of the SBM

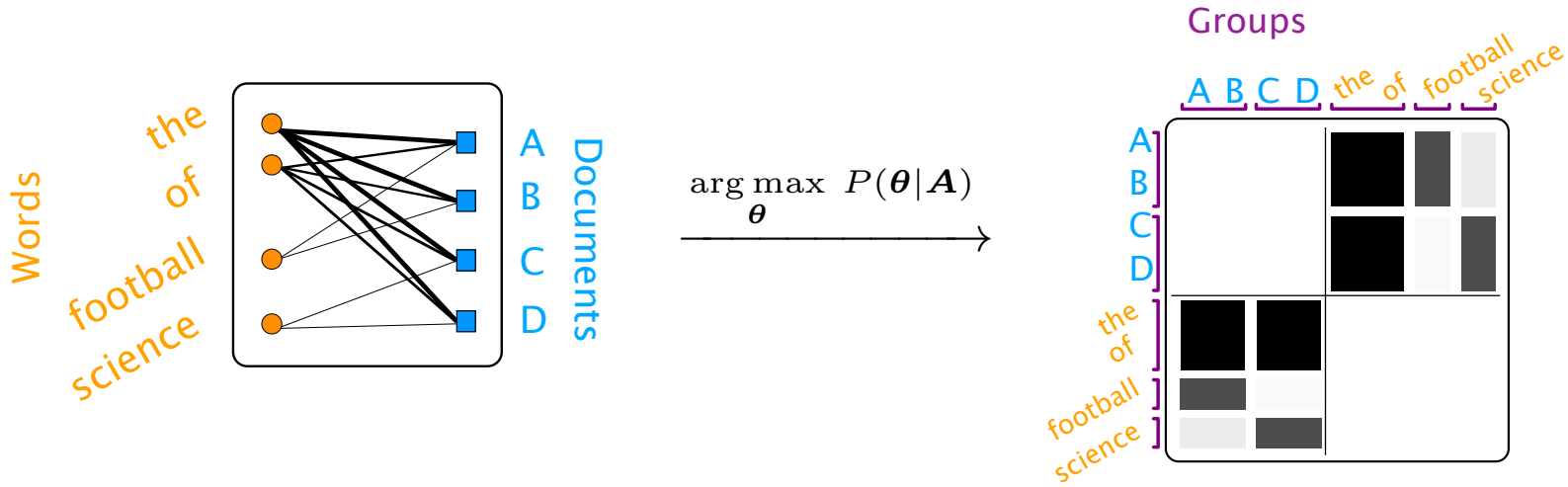


$$\Sigma = -\ln P(\boldsymbol{\theta}|\boldsymbol{A})$$

$\Sigma \rightarrow$ Total information necessary to describe the data.

Grünwald (The Minimum Description Length Principle, 2007)

The Bayesian Inference of the SBM



$$\Sigma = -\ln P(\boldsymbol{\theta}|\boldsymbol{A}) = \underbrace{-\ln P(\boldsymbol{A}|\boldsymbol{\theta})}_{\text{data}|\text{model},\mathcal{S}} \underbrace{-\ln P(\boldsymbol{\theta})}_{\text{model},\mathcal{L}}$$

$\Sigma \rightarrow$ Total information necessary to describe the data.

$\mathcal{S} \rightarrow$ Information required to describe the network $\boldsymbol{A}$, when the model is known.

$\mathcal{L} \rightarrow$ Information required to describe the model parameters $\boldsymbol{\theta}$.

Grünwald (The Minimum Description Length Principle, 2007)

Part III

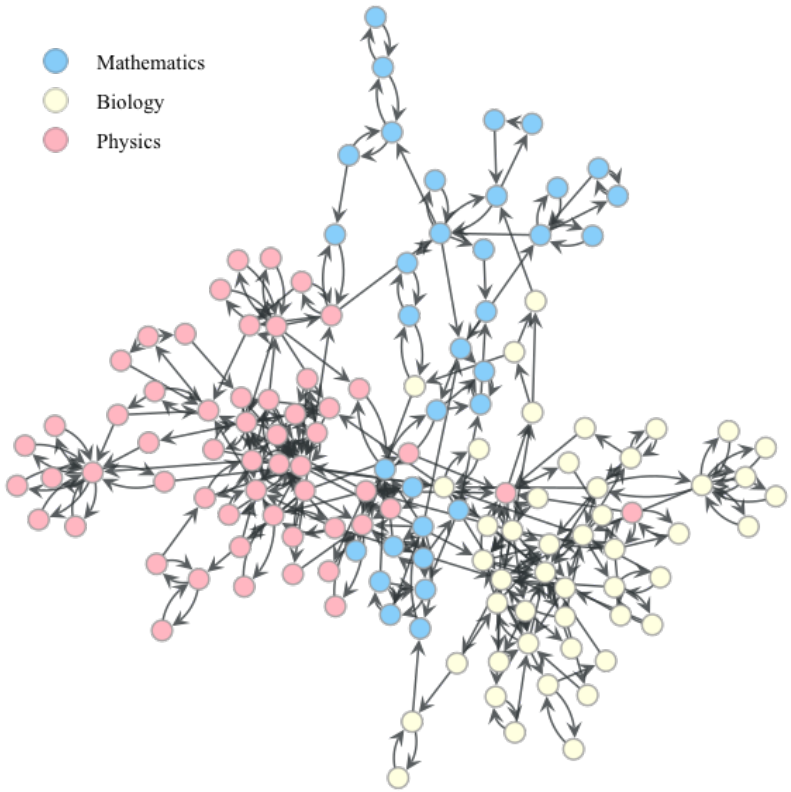
Will incorporating additional information available about the documents help us to **better** classify them?

Case Study: The Wikipedia Articles

Case Study: The Wikipedia Articles

- 138 Wikipedia articles with three categories:

  mathematics: 34,
  physics: 56,
  biology: 48

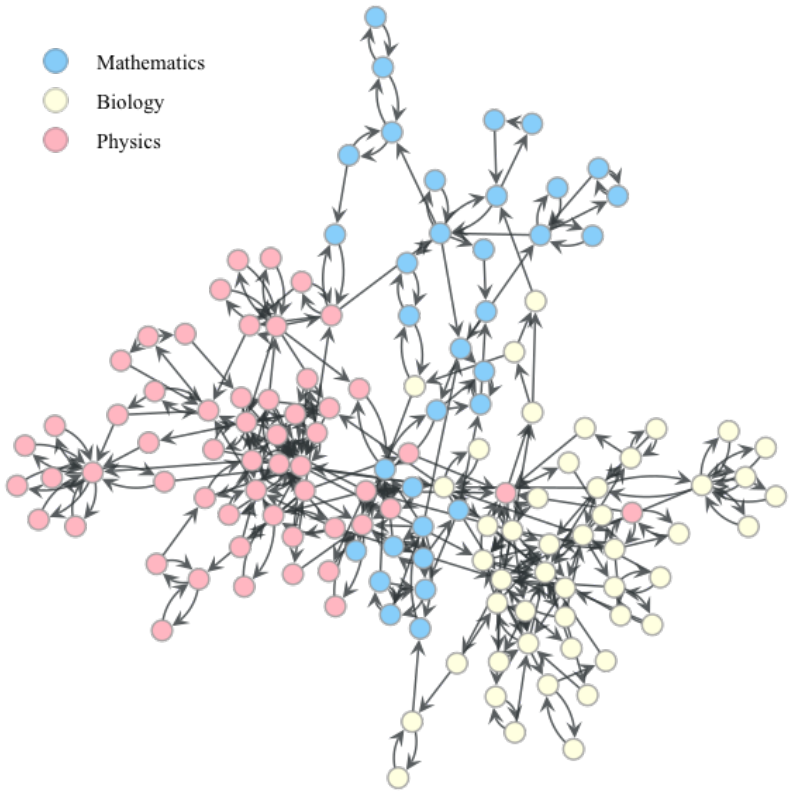


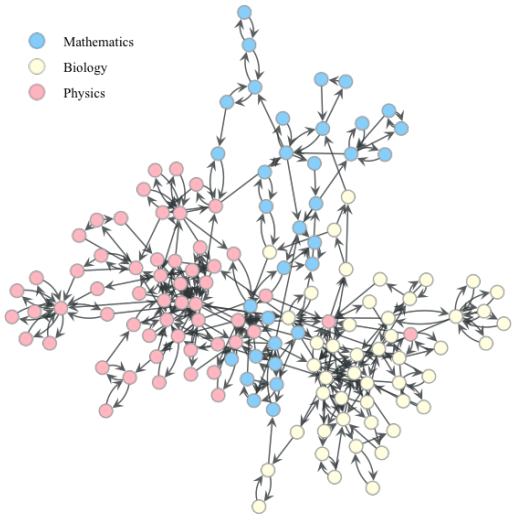
Case Study: The Wikipedia Articles

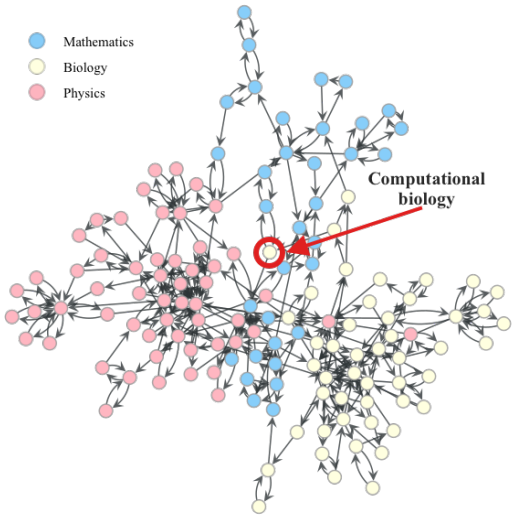
- 138 Wikipedia articles with three categories:

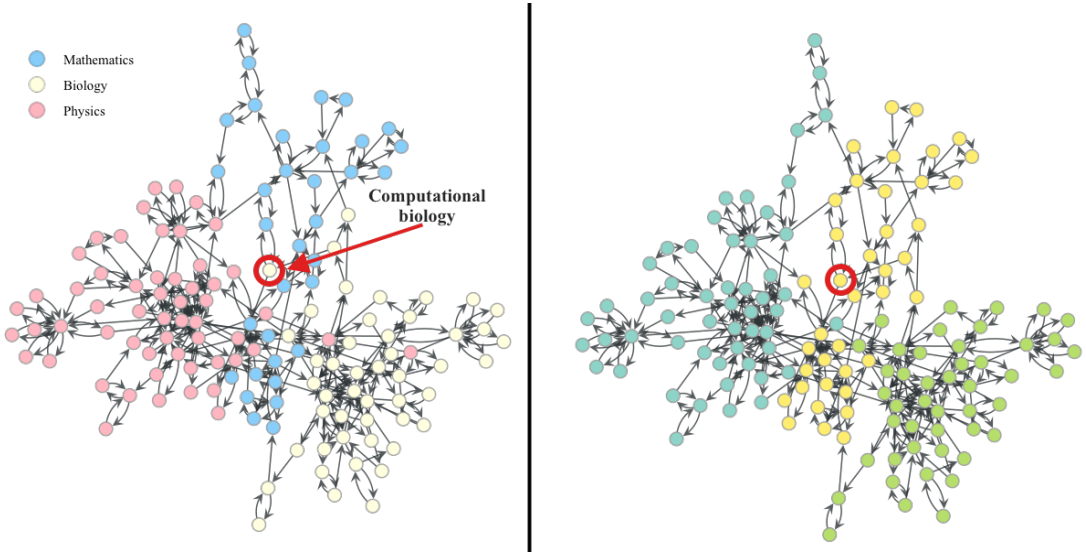
  mathematics: 34,
  physics: 56,
  biology: 48

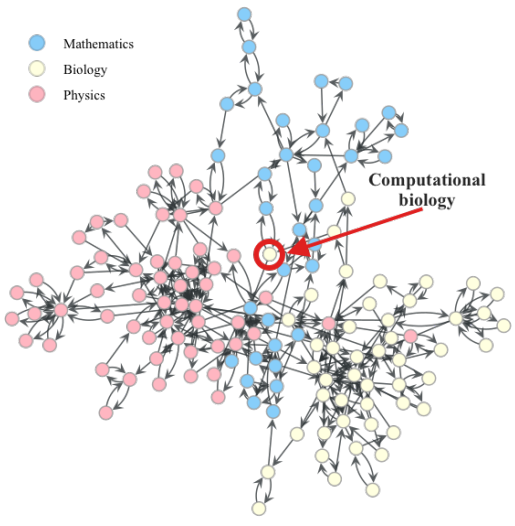
- Number of hyperlinks: 341
- Number of distinct words: 16,378
- Number of words: 351,710

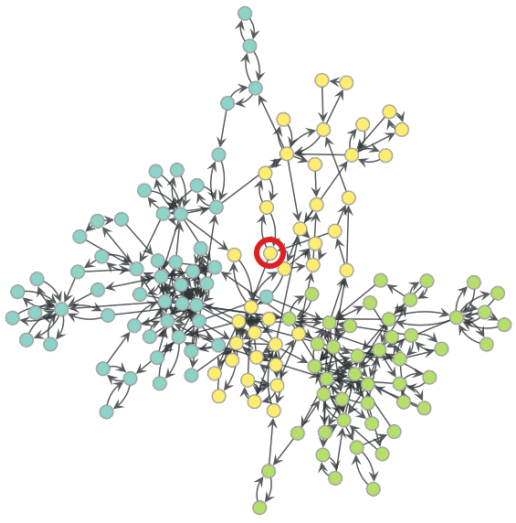


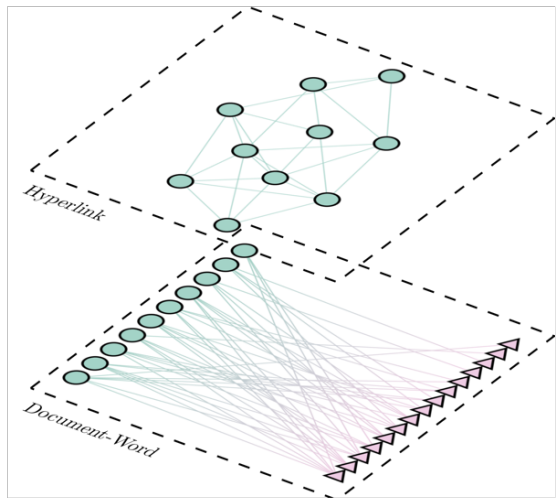

Mathematics
Biology
Physics

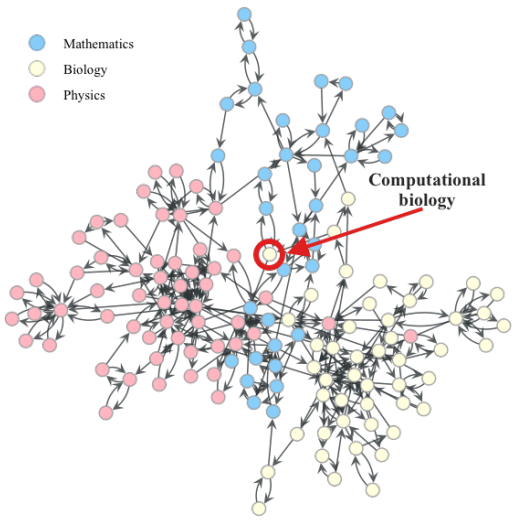

Mathematics
Biology
Physics
Computational biology

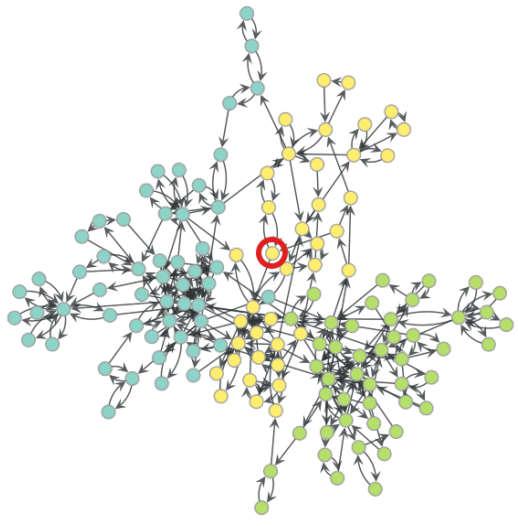

Mathematics
Biology
Physics
Computational biology

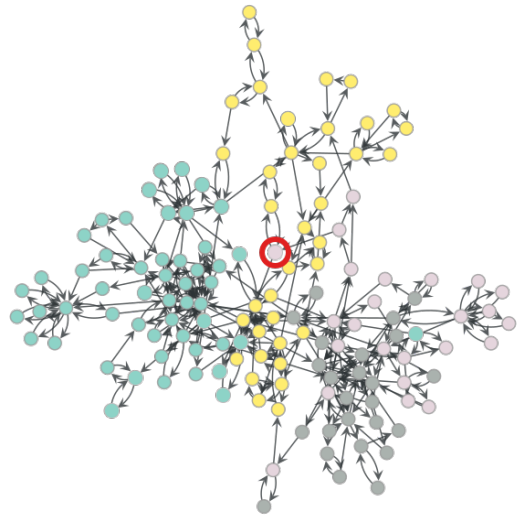

Mathematics
Biology
Physics
Computational biology



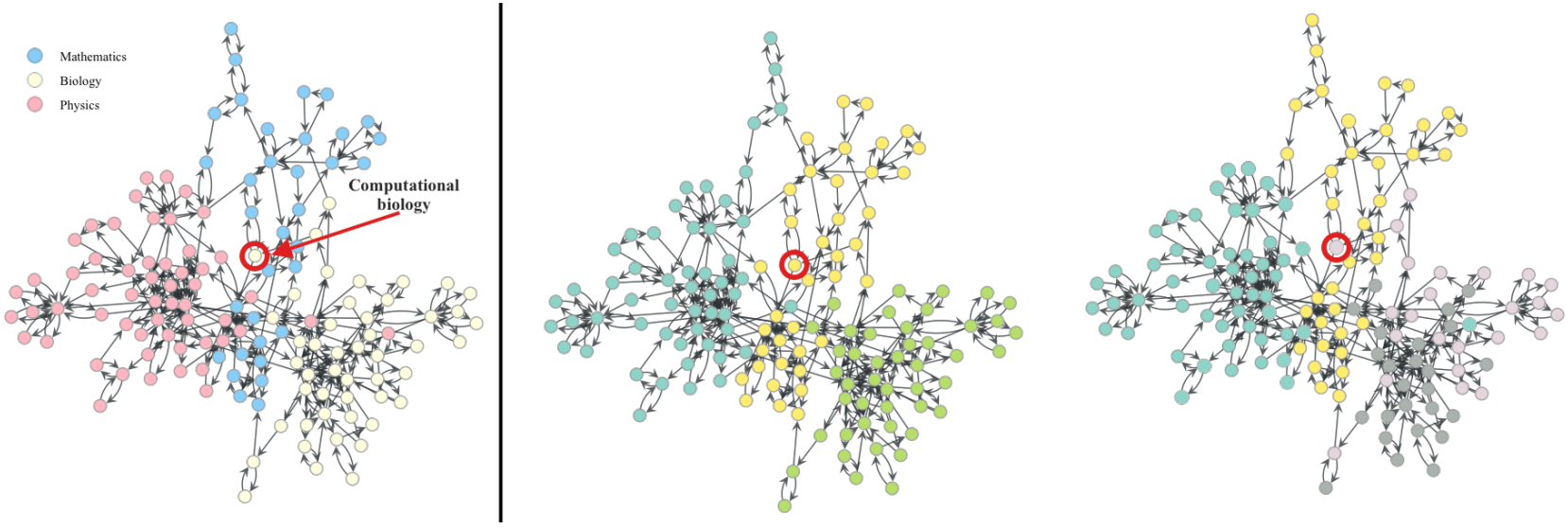

Hyperlink
Document-Word


Mathematics
Biology
Physics
Computational biology

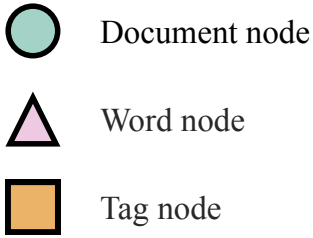


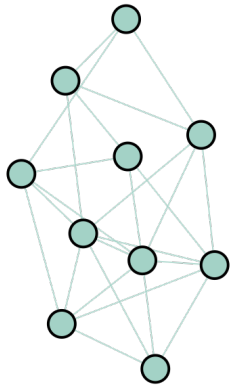


Measuring Partition similarity: Normalized Mutual information (NMI)

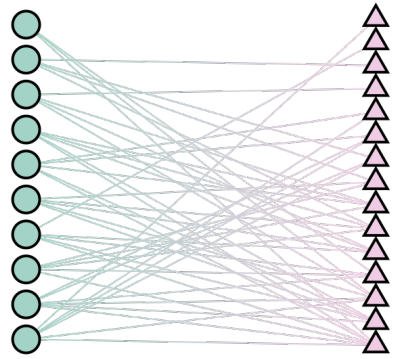


- NMI $\in [0, 1]$.
- Larger NMI values indicate better agreement.

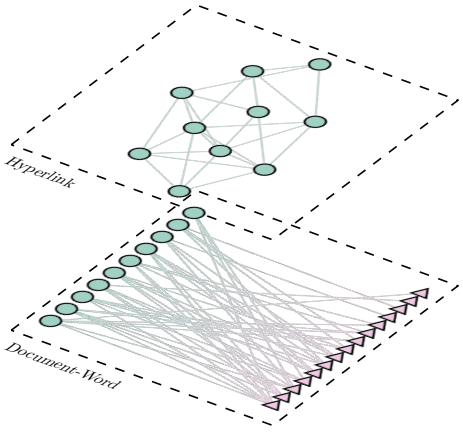




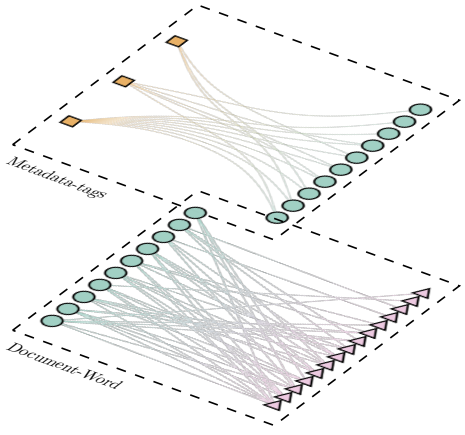
$\mathcal{D}_1$



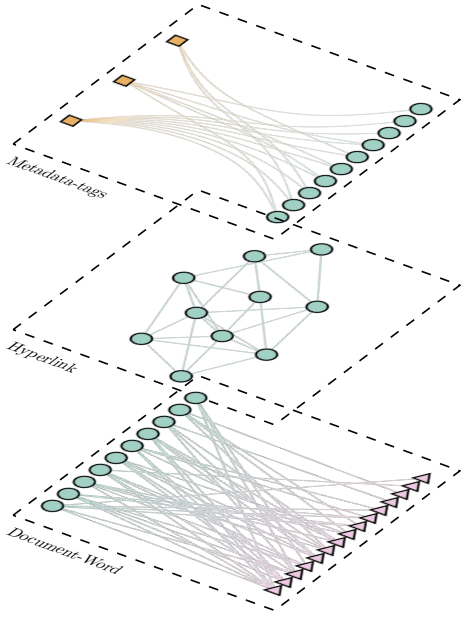
$\mathcal{D}_2$



$\mathcal{D}_3$



$\mathcal{D}_4$



$\mathcal{D}_5$

In each dataset $\mathcal{D}_1, \ldots, \mathcal{D}_5$

- Run the MCMC algorithm for multiple times.
- Obtain the inferred partition after annealing.

Comparison of partition similarities of documents

Wikipedia Labels	1
$\mathcal{D}_1$	0.51

In each dataset $\mathcal{D}_1, \ldots, \mathcal{D}_5$

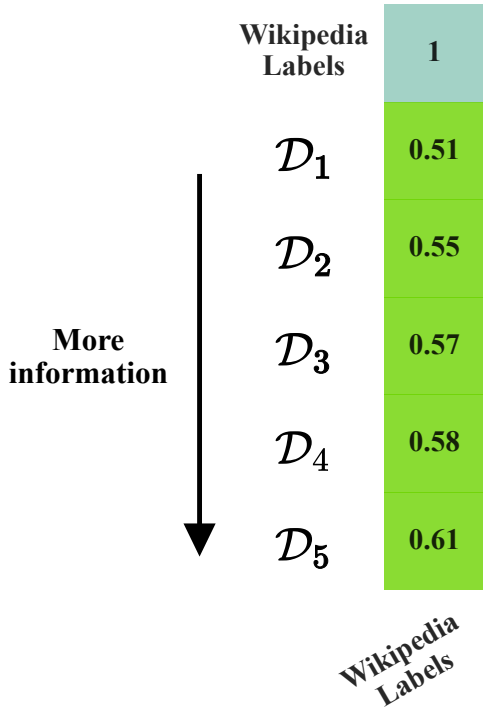
- Run the MCMC algorithm for multiple times.
- Obtain the inferred partition after annealing.

$\mathcal{D}_1$: **without** word nodes.

Comparison of partition similarities of documents

In each dataset $\mathcal{D}_1, \ldots, \mathcal{D}_5$

- Run the MCMC algorithm for multiple times.
- Obtain the inferred partition after annealing.



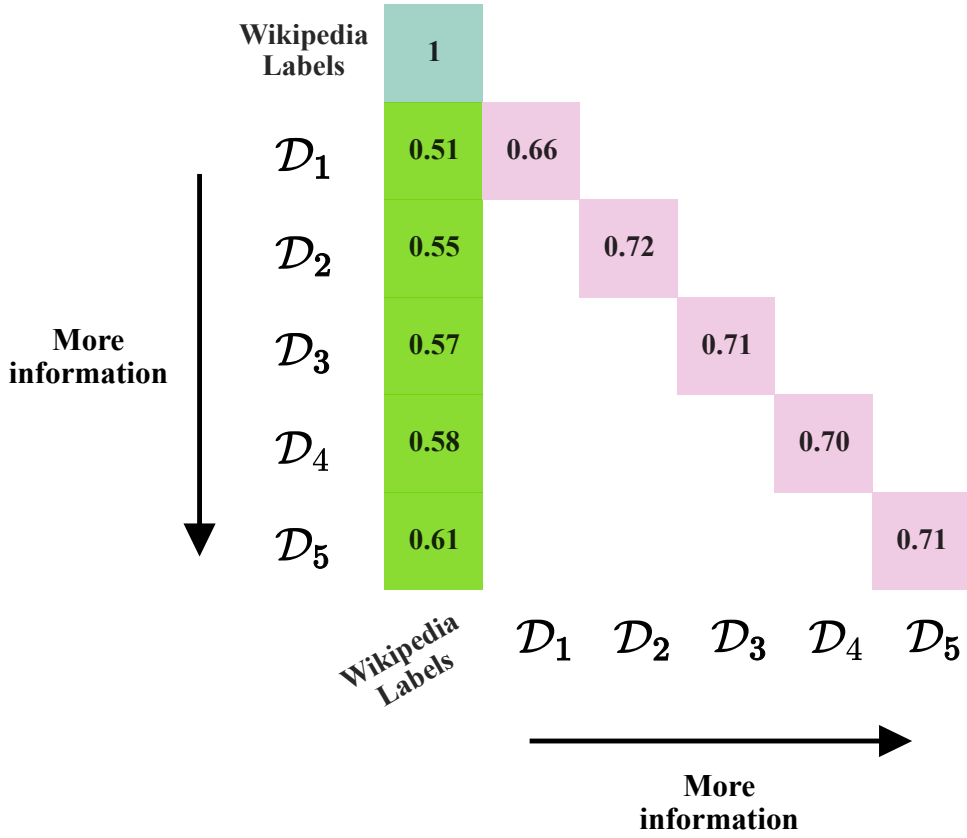
$\mathcal{D}_1$: **without** word nodes.

$\mathcal{D}_2 \sim \mathcal{D}_5$: **with** word nodes.

Comparison of partition similarities of documents

In each dataset $\mathcal{D}_1, \ldots, \mathcal{D}_5$

- Run the MCMC algorithm for multiple times.
- Obtain the inferred partition after annealing.



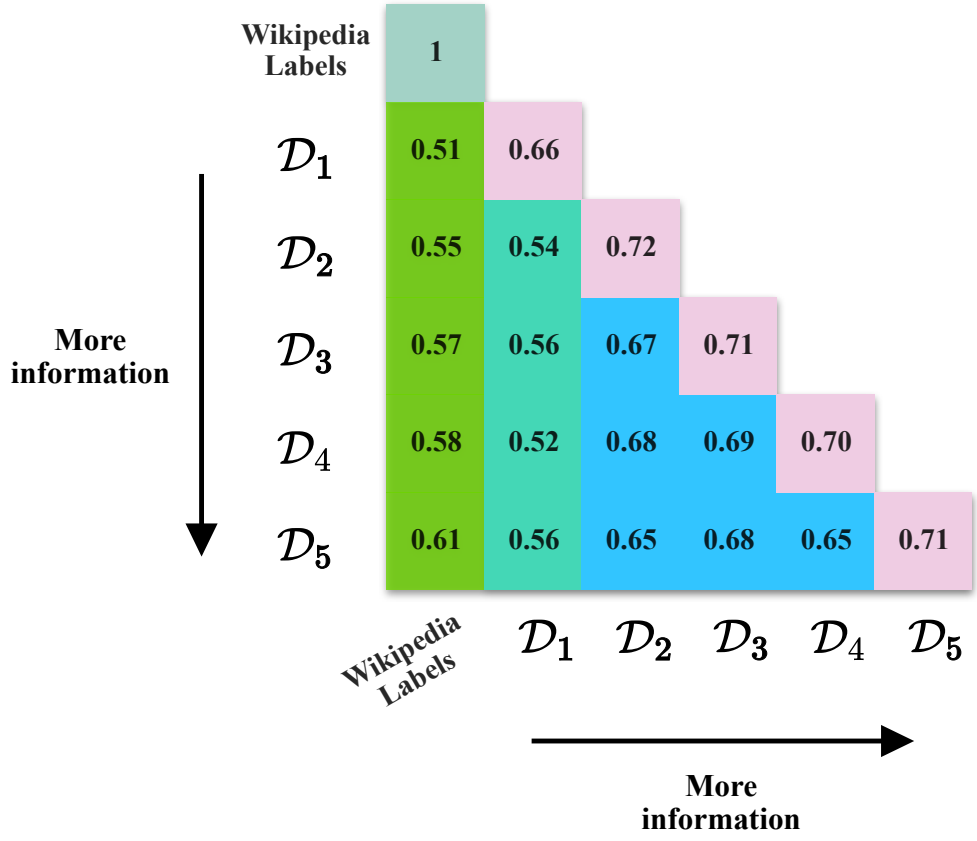
$\mathcal{D}_1$: **without** word nodes.

$\mathcal{D}_2 \sim \mathcal{D}_5$: **with** word nodes.

Comparison of partition similarities of documents

In each dataset $\mathcal{D}_1, \ldots, \mathcal{D}_5$

- Run the MCMC algorithm for multiple times.
- Obtain the inferred partition after annealing.



$\mathcal{D}_1$: **without** word nodes.

$\mathcal{D}_2 \sim \mathcal{D}_5$: **with** word nodes.

Main message:

Main message:

- We extend the previous work by incorporating additional information available about documents as additional layers in the same SBM framework.

Main message:

- We extend the previous work by incorporating additional information available about documents as additional layers in the same SBM framework.
- Compared to Wikipedia labeling of articles, incorporating more information will lead to better agreement.

Main message:

- We extend the previous work by incorporating additional information available about documents as additional layers in the same SBM framework.
- Compared to Wikipedia labeling of articles, incorporating more information will lead to better agreement.
- Document-word layer is dominating in the inference.

Main message:

- We extend the previous work by incorporating additional information available about documents as additional layers in the same SBM framework.
- Compared to Wikipedia labeling of articles, incorporating more information will lead to better agreement.
- Document-word layer is dominating in the inference.
- Partitions of documents are more similar when incorporating words.

Main message:

- We extend the previous work by incorporating additional information available about documents as additional layers in the same SBM framework.
- Compared to Wikipedia labeling of articles, incorporating more information will lead to better agreement.
- Document-word layer is dominating in the inference.
- Partitions of documents are more similar when incorporating words.

References:

1. Tiago P. Peixoto. "Bayesian stochastic blockmodeling." arXiv preprint arXiv:1705.10225 (2017).
2. Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. "A network approach to topic models." Science advances 4, no. 7 (2018): eaaq1360.
3. The graph-tool package `https://graph-tool.skewed.de`.