Summary of Papers Read

Zhang Mingyuan

# 1. Topic modelling

With the large amount data, we wish to have an algorithm to sort this documents base on their topics.

## Paper 1: Introduction to Probabilistic Topic Models + Latent Dirichlet Allocation

Authors (1): David M. Blei

Authors (2): David M. Blei; Andrew Y. Ng; Michael I. Jordan

This paper introduces Latent Dirichlet allocation (LDA). In Latent Dirichlet Allocation (LDA), the “generative process” is a conceptual framework rather than a literal process of generating documents. This framework is used to model how documents in a corpus might have been generated in terms of underlying topics. The goal of LDA is to reverse-engineer this process: to infer the latent topics that likely generated the observed documents. You do not actually generate the documents.

A diagram of a triangle

Description automatically generated

First, to understand Dirichlet distributions, if alpha is equals to 1, the points scatter uniformly in the triangle. If less than one, scatter closer to the edges. If greater than 1, concentrated in the centers.

**There are a few things to take note of for this model:**

1. The number of topic ‘k’ must be predetermined and there is no intrinsic way to determine this ‘k’ (at least for the simplest model of LDA).
2. Each documents are assumed to be a mixture of topics.
3. LDA assumes the bag-of-word approach, which means that the order of the word is not important.
4. In generative probabilistic modeling, data is treated as arising from a generative process that includes hidden random variables. The hidden variable in this case is the topic proportions. Only the word, which is our data is known.
5. The 2nd paper also made comparison with other techniques such as unigram, mixture of unigram and PLSI. Notably, this method is better than PLSI in terms of overcoming the overfitting issue.

**The generative process is done in two steps:**

1. Randomly choose a distribution over topics
2. For each word in the document
   1. Randomly choose a topic from the distribution over topics in step 1
   2. Randomly choose a word from the corresponding distribution over the vocabulary.

A diagram of a diagram

Description automatically generatedYou can also understand as:

1. A document-topic distribution is first sampled (alpha Dirichlet)
2. A topic is chosen from the multinomial distribution samples from step 1(theta)
3. Based on the result of step 2, a multinomial distribution is sampled from the Dirichlet distribution of parameter beta.
4. A word is sampled from the multinomial distribution obtained in step 3.

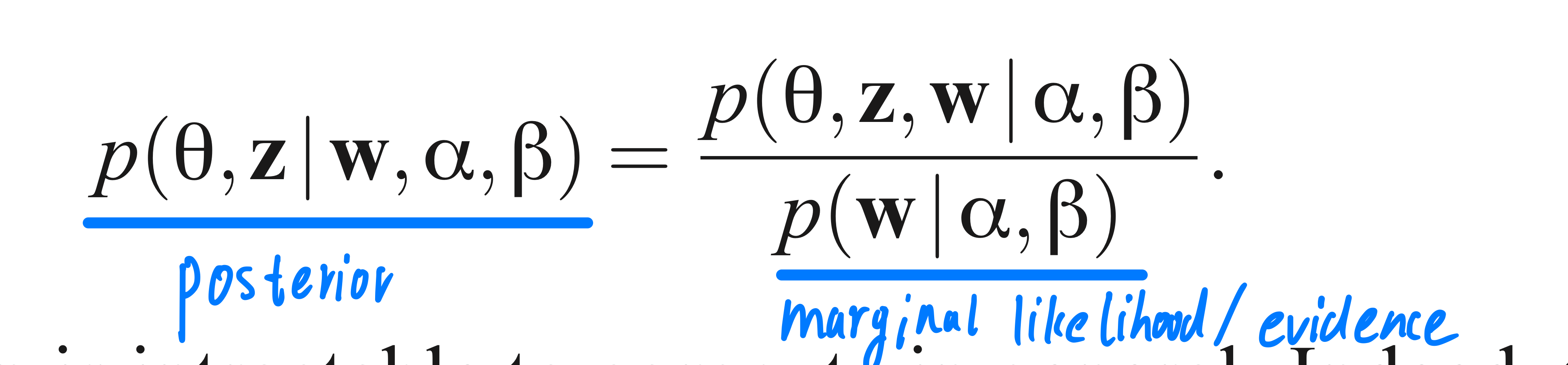
A diagram of different types of geometric shapes

Description automatically generated

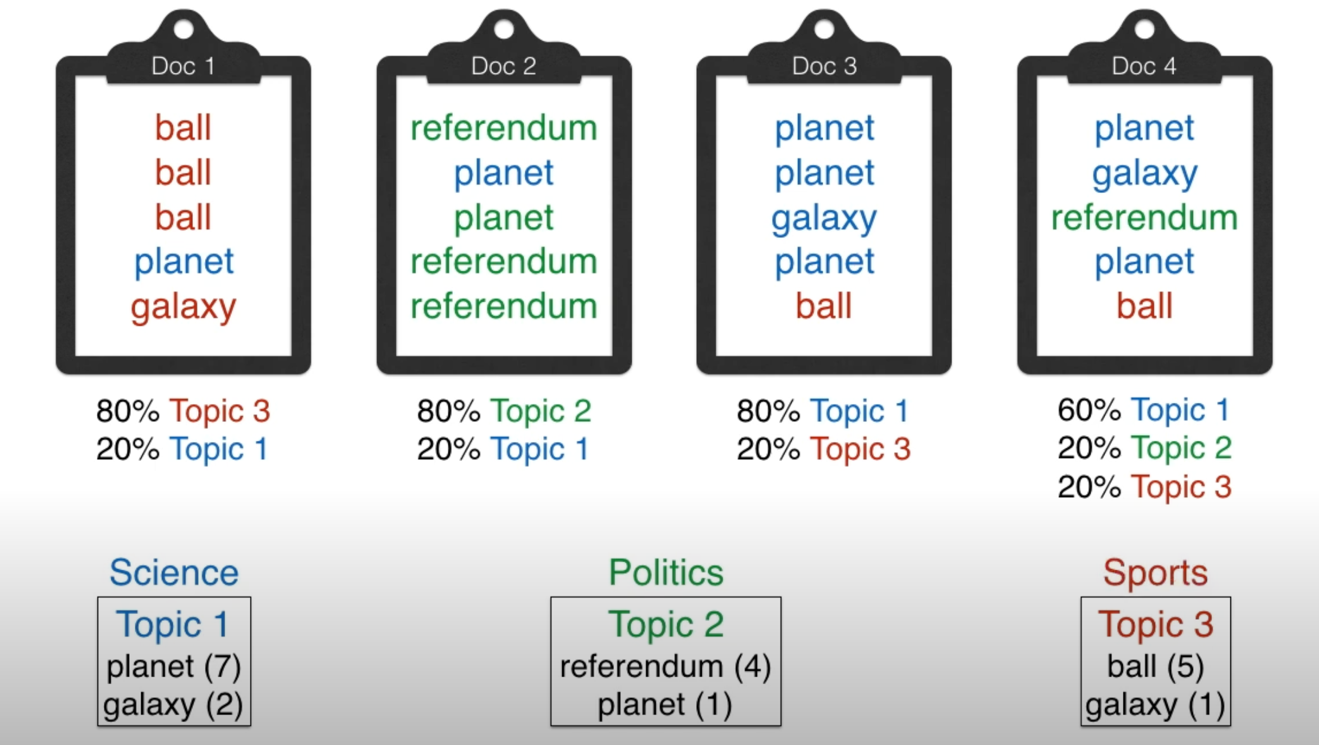
**Inference and parameter estimation:**

The key inferential problem that we need to solve to use LDA is that of computing the posterior distribution of the hidden variables given a document. The posterior distribution tells us the probability of the latent variables (the topics assigned to words in documents and the topic distributions for each document) given the observed data (the words in the documents).

If I am not wrong, alpha and beta is determined based on you prior knowledge of the document or simply use grid-search.



There are usually 2 ways to do this, variational inference (optimization problems) and Gibbs sampling.



**In essence,** the main computational challenge in Latent Dirichlet Allocation (LDA) lies in estimating the posterior distribution, which involves determining the most likely topic distribution for each document and the most likely word distribution for each topic based on the observed data (the words in the documents). Once this estimation process is complete, LDA provides several key outputs that can be used for various analytical tasks:

1. Topic Distribution per Document (θd)
   1. For each document in the corpus, you get a probability distribution over the topics. This tells you how much each topic contributes to that document.
2. Word Distribution per Topic ( φk):
   1. For each topic, you get a probability distribution over the vocabulary. This tells you which words are most strongly associated with each topic.
3. Topic Assignments for Words ( zd,n ):
   1. LDA also provides a topic assignment for each word in each document. This indicates which topic the model believes a particular word in a document is most likely associated with.

## Paper 2: Probabilistic latent Semantic Indexing(PLSI)

Author: Thomas Hofmann

This paper builds on the Latent sematic indexing(LSI) which uses Singular value decomposition to carry out topic modelling. LSI lacks a strong statistical foundation and PLSA aims to solve that.

LSA solves the problem by trying to map documents as well as terms to a representation in the so-called latent semantic space. It applies dimensionality reduction by using SVD.

**The aspect model:**

There are 2 independent assumptions of this model:

1. Observation pairs are assumed to be generated independently(‘bag-of-word)
2. The conditional independence assumption is made that conditioned on the latent class ‘z’, words ‘w’ are generated independently of the specific document identity d.

**The generative model is defined as above:**

A black and white text with black text

Description automatically generated

**Inference:**

The primary challenge in PLSI, like in LDA, is to estimate the posterior distributions:

1. P(z|d): The probability distribution of topics for each document.
2. P(w|z): The probability distribution of words for each topic.

Using tempered EM.

**Outputs of PLSI**

Once the EM algorithm converges, PLSI provides the following outputs:

1. Topic Distribution per Document ( P(z|d) ):
   1. For each document, PLSI provides a distribution over topics. This tells you the proportion of each topic in the document.
2. Word Distribution per Topic ( P(w|z) ):
   1. For each topic, PLSI provides a distribution over the vocabulary, indicating which words are most strongly associated with each topic.
3. Topic Assignments for Words:
   1. Similar to LDA, PLSI can also infer the topic assignment for each word in each document, though this is typically done as part of the E-step.

## Paper 3: A high-reproducibility and high-accuracy method for automated topic classification

Author: Andrea Lancichinetti, M. Irmak Sirer, Jane X. Wang

This paper aims to improve the current method in topic modelling through a network approach. Corpus is first treated as a bipartite networks of words which. Are connected if they co-appear in a document. A topic matching is done first to do a hard clustering(each document can has only one topic). The PLSI comes in to do a soft clustering, allowing mixture of topics.

## Paper 4: A network approach to topic models

Authors: Martin Gerlach, Tiago P. Peixoto

This paper aim to solve topic modelling using a network approach. It draws connections and similarities between the techniques used in topic modelling and community detection. Namely, it is the mathematical equivalence between SBM and PLSI and the parallelism between LDA and hSBM. The aims is to overcome the problem of overfitting in the current topic modelling methods. In addition, for LDA, there is no justification of using the Dirichlet distribution, which will limit the type of mixtures of topics and is not designed to be compatible with well-known properties of real text , such as Zipf’s law.

A diagram of a diagram

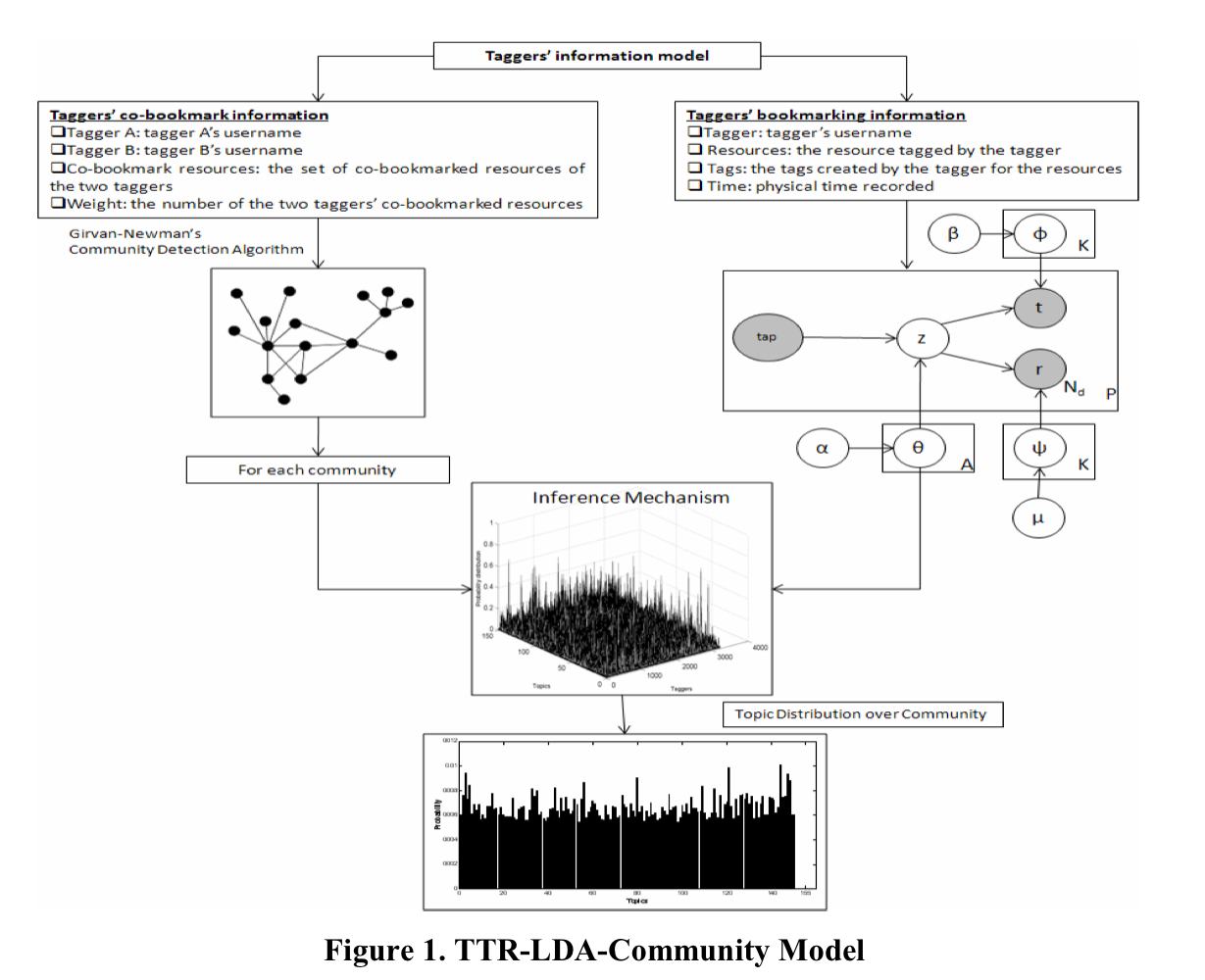
Description automatically generated

To do it, it transformed the corpus into a bipartite graph and solve the problem using SBM.

## Paper 5: Community-based topic modelling for social tagging

Author: Daifeng Li, Bing He, Ying Ding…

Whats their data, whats their goal, what are they estimating, whats the correlation between their work and ours.



The community detection is done using Girvan-Newman algorithm while the topic modelling is done using Tagger-Tag-Resource LDA(TTR-LDA) which is a 3 layer Bayesian model. The first layer is the tagger, the second is the latent topics while the last is the resources and tags. Each layer has their respective information profile.

The flow of the entire mechanism is:

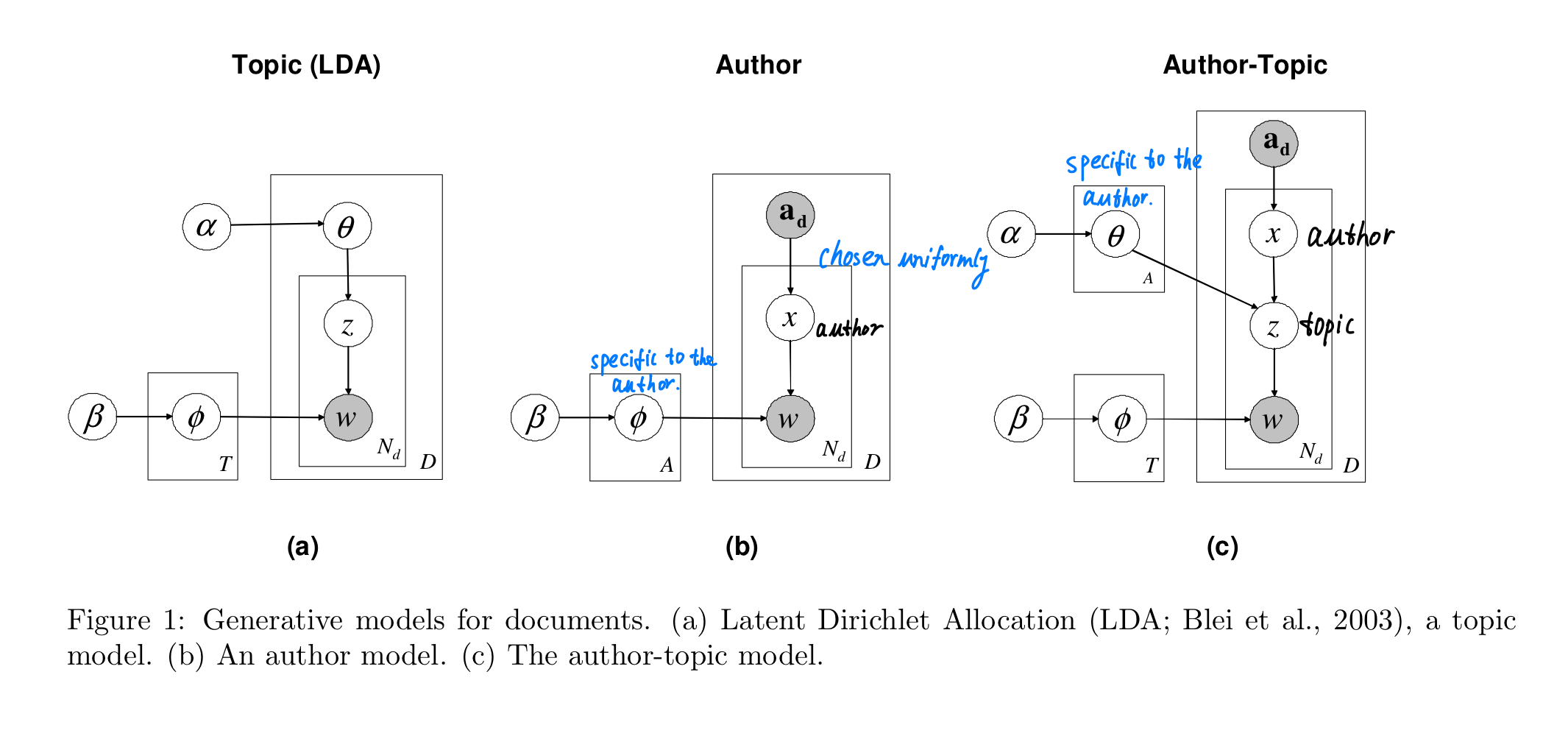
1. Create an information profile for each of the taggers in the sample.
2. Topic modelling using TTR-LDA and Community detection using Girvan-Newman algorithm.
3. Integrate the results from the TTR-LDA and community detection model.

This is done diachronically as contrast to another paper which does it synchronically (Paper 7).

The output will be a topic distribution for each community at a certain time. Read section 3.2 in detail.

## Paper 6: The Author-Topic Model for Authors and Documents

Author: Michal Rosen-Zvi; Thomas Griffiths



This paper introduces 2 modified LDA. The Author-Topic LDA subsumes the two other models, meaning that they are special cases of Author-Topic LDA.

As shown the figure (c), Author-Topic LDA introduces another latent variables, authors. For each word, an author is chosen uniformly. For each author, he has a specific distribution of topics which is sampled from the Dirichlet distribution with parameter alpha. Then, a topic is sampled from the multinomial distribution with parameter theta. The rest is the same as in LDA.

This paper estimates the parameters using Gibbs Sampling.

## Paper 7: Topic-Link LDA: Joint Models of Topic and Author Community

Author: Yan Liu…

A diagram of a diagram of a diagram

Description automatically generated

This paper considers community as a latent variable. With the assumption of exchangeability of authors, they are modelled into a mixture distribution.

Other than the topic modelling, the algorithm also tries to find out if a link exists. This is done by considering the community of the author as well as τ which represents parameters that influence the probability of a link (or citation) between documents. It captures how much the community and topic similarities influence the likelihood of a link.

# 2. Community detection

## Paper 1: Community Detection and Stochastic Block Models: Recent Developments

Author: Emmanuel Abbe

**SBM(n,p,W)**

Where n is the number of nodes,

p is a probability vector of dimension,

W is a k \* k symmetric matrix with entries in [0,1]

Input data will be an observed network G.

**Inference:**

Parameters that require to be estimated will be the community membership z and probability matrix p. MLE/EM/Variational inference can be used.

The output will be assignment of community to each node.

**Distinction between weak recovery and strong recovery:**

Strong recovery meaning to recover all the labels and they has to be correct. Note that as the number of n increases, the probability tend to 1.

Weak recovery only ask for reconstruction of the nodes that is positively correlated. Detection is a even weaker recovery where you try and see if the graph is an Erodos-Renyi graph or not. i.e. if there exist any communities.

**Phase transition phenomena:**

Kesten-Stigum threshold

A white board with blue writing

Description automatically generated

Agreement…

## Paper 2: Computational Social Networks – Section 4

Author: V.Labatut and J.-M. Balasque

### Choosing an appropriate algorithm

**Rubrics in choosing:**

1. The type of information that algorithm is able process.
   1. Weights, directions.
   2. Node attributes
   3. Different classes of links
2. The kind of community structure the algorithm produces.
   1. Mutually exclusive vs overlapping
3. Nature of communities

**Categories of community detection algorithm**

1. Density
2. Pattern
3. Node similarity
4. Link centrality
   1. Girvan-Newman
5. Others

A table of information

Description automatically generated with medium confidence

### Interpretation of community

1. Topology of the community structure
2. Attribute-Based Interpretation