Research Report

Zhang Mingyuan AY 24/25 S1

Aim: Topic modelling using community networks with covariates(corpus)

Data:

1. A community network of ‘n’ nodes. Adjacency matrix
2. A corpus of size n×p, where p is the number of words

Propose a 2-step algorithm:

1. Community detection: find the communities in the network.
   1. People in the same communities may share similar topics.
   2. The algorithm used and hence the output should facilitate the ease of topic modelling that is carried out in the next step.
2. Topic modelling:
   1. Usually, algorithms such as LDA uses a corpus directly, now you have additional network data with communities found, which algorithm will be good to utilise this additional information?
   2. In addition, you aim to obtain a topic distribution for each community.

## 1. Community detection:

**Objective**: Detect communities within the network that will be leveraged for more effective topic modelling.

Can it uncover information which can be used to model the latent variables in the next step? 🡺 interpreting and understanding the community.

Assuming that each node belongs to one community only.

**Potential Algorithms**:

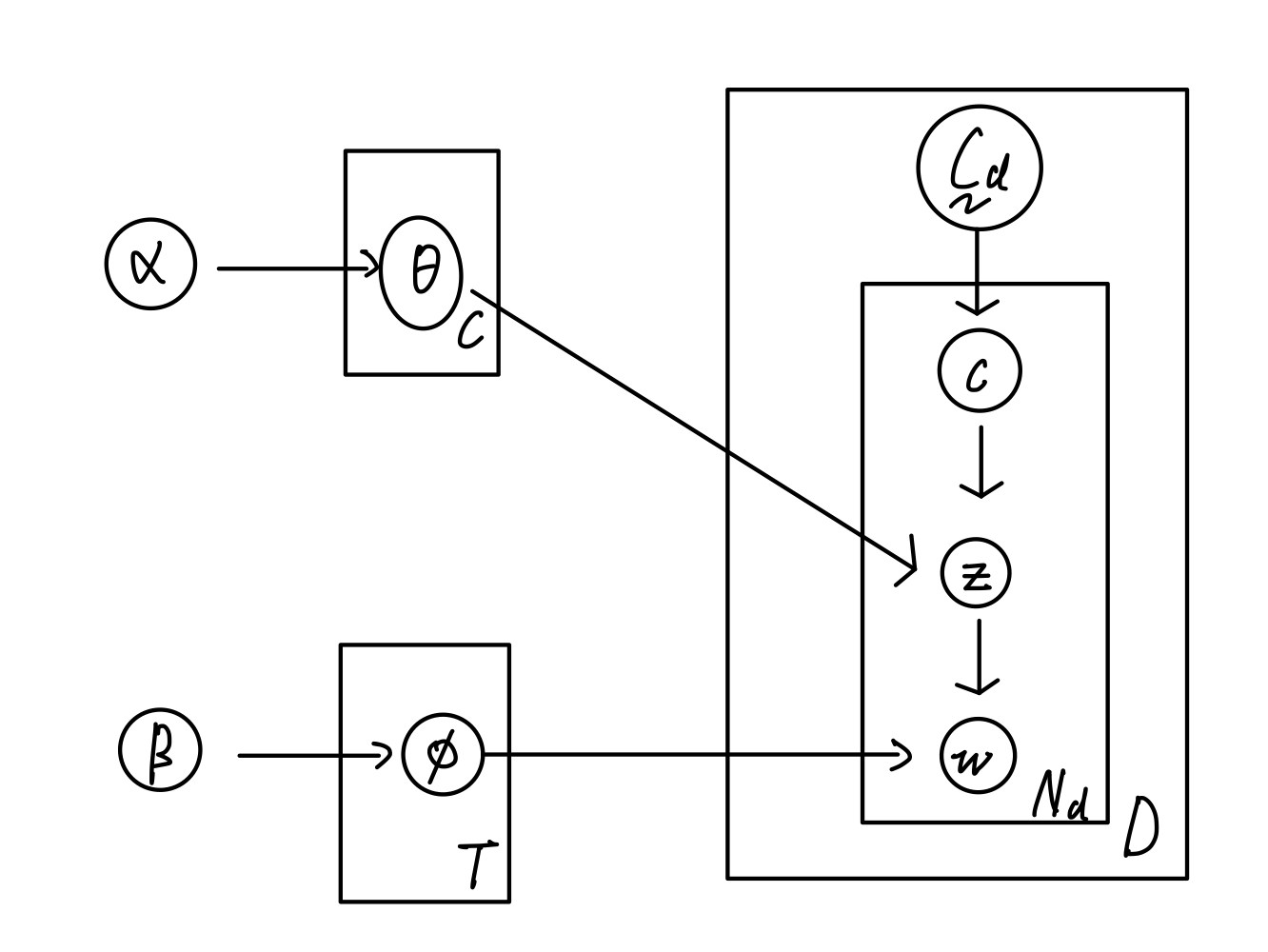
1. **Stochastic Block Model (SBM)**
2. **Girvan-Newman**

## 2. Topic modelling:

**Objective**: Find out common topics specific to each of the community (topic distribution for each of the communities).

**Algorithms**:

1. Introduce community as a latent variable into the LDA model.



However, there must be a way in which the chosen ‘c’ affects the ‘z’ based on the parameter alpha.

In addition, how should the ‘c’ be chosen? Since we have the corpus, we know which community the person and hence the word belongs to.

Assuming exchangeability of communities? Such that we can treat the communities as being generated from a common distribution without worrying about the specific order in which authors are processed.

Next up:

1. Interpretation of community
   1. Choose an algorithm accordingly