Clustering in machine learning literature II

Model-based graph clustering

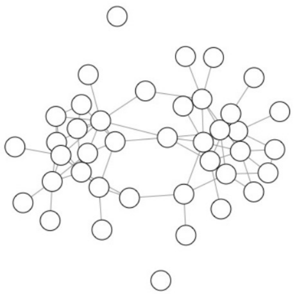
- introduce a probabilistic generative model for the networks
- recast the graph comparison task as a problem of estimating and comparing the probabilistic models
- for graphs with shared nodes: Stanley et al. (2016); Mukherjee et al. (2017)
- for graphs without any node correspondence: Sabanayagam et al. (2022)

New network clustering approach

Our approach

- is a statistical one
- introduce a **statistical model** for each graph
- perform **model-based clustering** (like a classical mixture model)
- hierarchical **agglomerative clustering algorithm**
- interpretable output

Modelling I

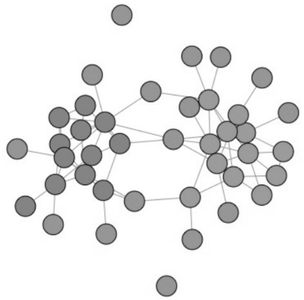


$\Longleftrightarrow$

```
> adjMatrix
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
 [1,]    0    0    0    0    0    0    0    0    0     0     1     0     1
 [2,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [3,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [4,]    0    0    0    0    0    0    0    0    1     0     0     0     0
 [5,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [6,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [7,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [8,]    0    0    0    0    0    0    0    0    0     1     0     0     0
 [9,]    0    0    0    1    0    0    0    0    0     0     0     0     0
[10,]    0    0    0    0    0    0    0    1    0     0     1     0     0
[11,]    1    0    0    0    0    0    0    0    0     1     0     0     0
[12,]    0    0    0    0    0    0    0    0    0     0     0     0     0
[13,]    1    0    0    0    0    0    0    0    0     0     0     0     0
```

Adjacency matrix

Modelling II



Stochastic block model (SBM)

- **Block memberships** For node i, $Z_i \in \{1, \ldots, K\}$ is drawn independently with probabilities

$$\mathbb{P}(Z_i = \bullet) = \pi_{\bullet}$$

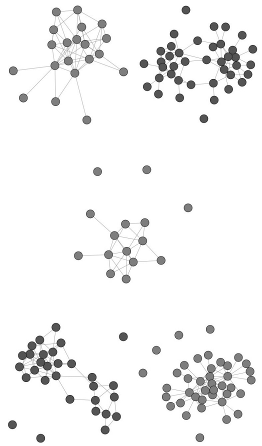
- **Edges** Conditionally on $Z_1, \ldots, Z_n$, $A_{i,j}$ are drawn independently

$$A_{i,j} | (Z_i = \bullet, Z_j = \bullet) \sim \text{Bernoulli}(\gamma_{\bullet\bullet})$$

- **Model parameters** of the SBM:

$$\theta^{\mathrm{SBM}} = ((\pi_k)_{1 \le k \le K}, (\gamma_{k,l})_{1 \le k,l \le K})$$

Modelling III



Mixture model of SBMs

- **Cluster membership** For network m, $U_m \in \{1, \dots, C\}$ is drawn independently with probabilities

$$\mathbb{P}(U_m = c) = p_c$$

- C different SBM parameters $\theta_c^{\mathrm{SBM}}, c = 1, \dots, C$
- Conditionally on $U_1, \dots, U_M$, the adjacency matrix $A^{(m)}$ is drawn from a SBM:

$$A^{(m)} | (U_m = c) \sim \mathrm{SBM}(\theta_c^{\mathrm{SBM}})$$

Estimation I

Estimation in the simple SBM

- MCMC (Nowicki and Snijders, 2001; Peixoto, 2014)
- variational EM (Daudin et al., 2008)
- spectral clustering (Rohe et al., 2011)
- pseudo-likelihood (Amini et al., 2013)
- ICL maximization (Côme and Latouche, 2015)
- VAE (Mehta et al., 2019)