Contextual Stochastic Block Models

Yash Deshpande

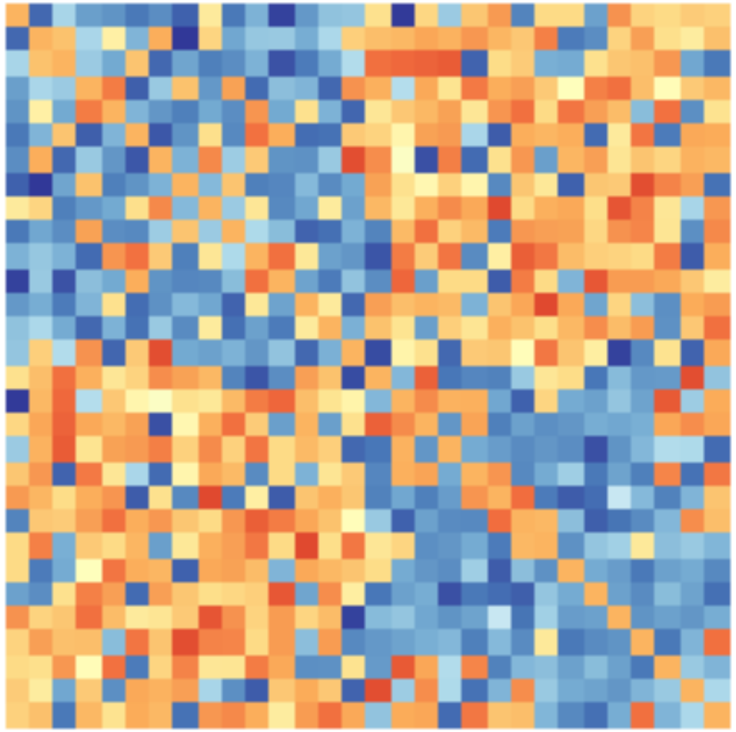
Andrea Montanari

Elchanan Mossel

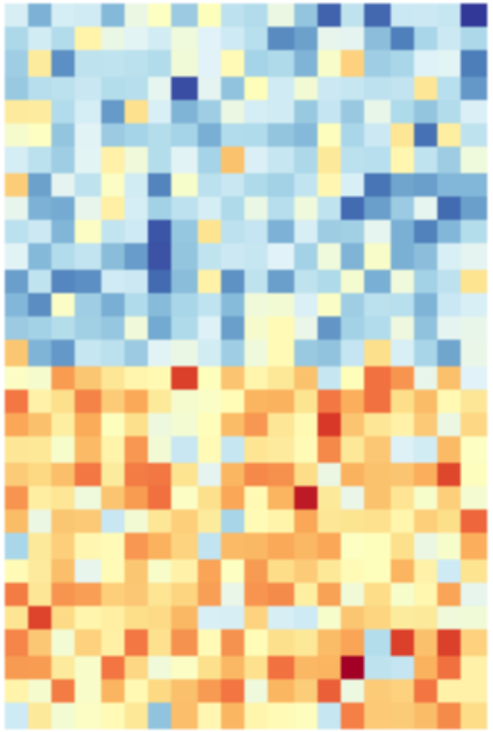
Subhabrata Sen

Two paradigms for clustering

Similarity-based

$A =$ 

Feature-based

$B =$ 

What if we have both?

- Ecological networks: covariates on species (mass, feed,…)
- Citation networks: covariates from article (keyword, journal,…)
- …

A statistical model

Two latent clusters, encoded as $v \in \{\pm 1\}^n$

Graph similarity

$$\mathbb{P}\{A_{ij} = 1\} = \begin{cases} \frac{c_{\text{in}}}{n} & \text{if } v_i = v_j \\ \frac{c_{\text{out}}}{n} & \text{otherwise.} \end{cases}$$

$$c_{\text{in}} =: d + \lambda d$$

$$c_{\text{out}} =: d - \lambda d$$

Gaussian mixture covariates

$$b_i = \sqrt{\frac{\mu}{n}} u v_i + z_i,$$

$$u, z_i \sim \mathrm{N}(0, \mathrm{I}_p)$$

Each individually

Graph similarity

Theorem (MNS13, 15, Mas14):

v recoverable from similarity graph if and only if:

$$\lambda^2 > 1$$

Gaussian mixture covariates

Theorem (BBAP05, OMH13):

v recoverable from covariates if and only if:

$$\mu^2 > \gamma =: \frac{n}{p}$$

Our result combines two phase transitions

Informal theorem (D, Montanari, Mossel, Sen)

In the limit of large degree d, v recoverable from graph and covariate data if and only if:

$$\lambda^2 + \frac{\mu^2}{\gamma} > 1$$

Thank you!

Room 210, Poster # 79

5pm – 7pm