Social Networks & Recommendation Systems

X. Community detection algorithms.

Grzegorz Siudem

Warsaw University of Technology

European Funds
Knowledge Education Development

Warsaw University of Technology

European Union
European Social Fund

MSc program in Data Science has been developed
as a part of task 10 of the project
„NERW PW. Science - Education - Development - Cooperation”
co-funded by European Union from European Social Fund.

Before classes

Reminder

From other courses:

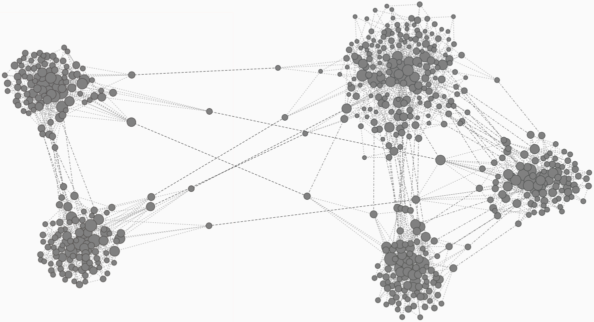
- What clustering methods in R^n do you know?

From SNARS_9:

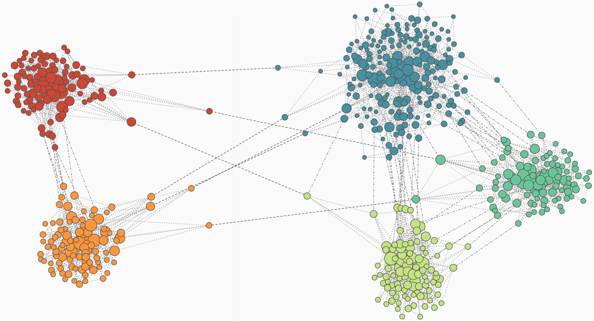
- Markov chains – random walks on graphs.

Lecture

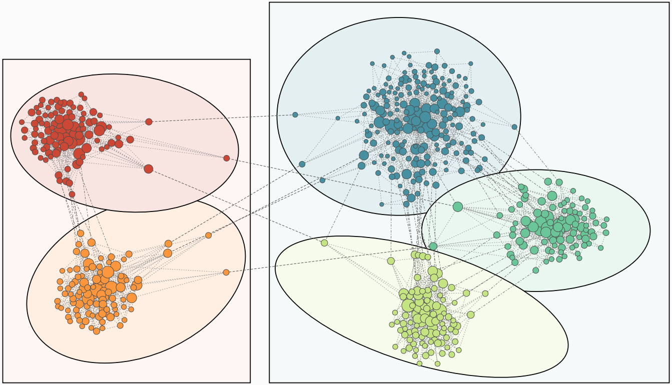
Empirical justification



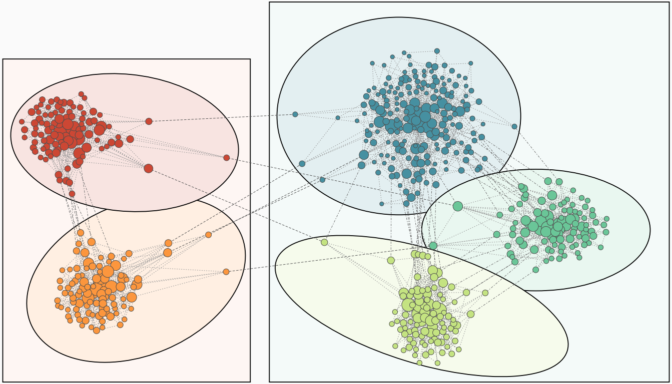
Empirical justification



Empirical justification



Empirical justification



Why we detect clusters/communities?

- we are looking for important features of the components,
- we ask for the number of these components,
- we are looking for a hierarchy in the analyzed system.

The problem with the right formulation of the problem

- there is no unambiguous and universal definition of *what* communities are,

The problem with the right formulation of the problem

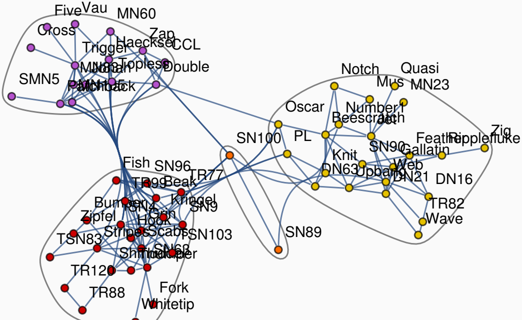
- there is no unambiguous and universal definition of *what* communities are,
- no (generally) a priori method to determine the number of communities for a given network.

The problem with the right formulation of the problem

- there is no unambiguous and universal definition of *what* communities are,
- no (generally) a priori method to determine the number of communities for a given network.
- detection method benchmarks are difficult to access (impossible in general?).

- there is no unambiguous and universal definition of *what* communities are,
- no (generally) a priori method to determine the number of communities for a given network.
- detection method benchmarks are difficult to access (impossible in general?).

And yet, intuitively, the problem is understandable



Common methods of community detection

In the following, we based on

- S. Fortunato, D. Hric, Phys. Rep., 659, 1, (2016).
- outside the academic network you can find work on arxiv: arXiv:1608.00163.

Common methods of community detection

In the following, we based on

- S. Fortunato, D. Hric, Phys. Rep., **659**, 1, (2016).
- outside the academic network you can find work on arxiv: arXiv:1608.00163.

If you are interested, please read

- rich bibliography *ibid.*
- with particular emphasis on work `https://arxiv.org/abs/0906.0612`

Common methods of community detection

In the following, we based on

- S. Fortunato, D. Hric, Phys. Rep., **659**, 1, (2016).
- outside the academic network you can find work on arxiv: arXiv:1608.00163.

If you are interested, please read

- rich bibliography *ibid.*
- with particular emphasis on work `https://arxiv.org/abs/0906.0612`

If you are very interested, please read

- the whole *community* of publications which cite this monograph.

What are communities?

What are communities?

- classic: division a set of vertices.

What are communities?

- classic: division a set of vertices.
- sometimes, however, we allow for subsets overlapping.

What are communities?

- classic: division a set of vertices.
- sometimes, however, we allow for subsets overlapping.
- practically: sets where *inside* connections are more dense than *outside*.

What are communities?

- classic: division a set of vertices.
- sometimes, however, we allow for subsets overlapping.
- practically: sets where *inside* connections are more dense than *outside*.

Reminder – simple model with community structure

Generalization of the Erdös-Rényi graphs to stochastic block model

$$\begin{pmatrix} p_{11} & p_{12} & \dots & p_{1K} \\ p_{21} & p_{22} & \dots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K1} & p_{K2} & \dots & p_{KK} \end{pmatrix}$$

- K - number of communities,
- $N > K$ number of vertices.

Stochastic block model (from Fortunato and Hric)

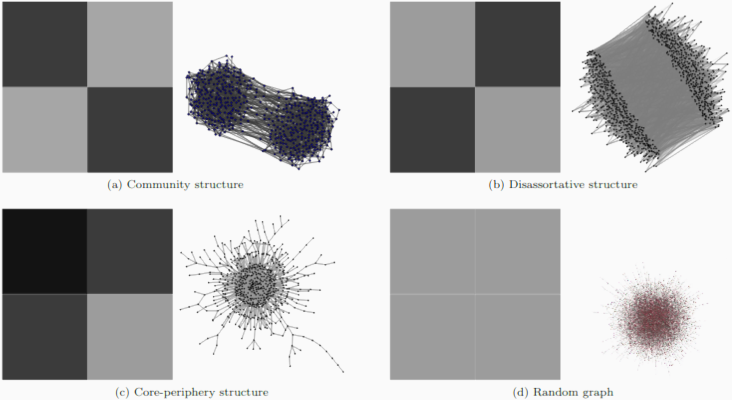


FIG. 8 Stochastic block model. We show the schematic adjacency matrices of network realisations produced by the model for special choices of the edge probabilities, along with one representative realisation for each case. For simplicity we show the case of two blocks of equal size. Darker blocks indicate higher edge probabilities and consequently a larger density of edges inside the block. Figure 8a illustrates community (or assortative) structure: the probabilities (link densities) are much higher inside the diagonal blocks than elsewhere. Figure 8b shows the opposite situation (disassortative structure). Figure 8c illustrates a core-periphery structure. Figure 8d shows a random graph à la Erdős and Rényi: all edge probabilities are identical, inside and between the blocks, so there are no actual groups. Adapted figure with permission from (Jeub *et al.*, 2015). © 2015, by the American Physical Society.

Spectral methods

General description

- We are looking for the eigenvalues of the adjacency matrix (or other related ones).
- We search for clusters of these eigenvalues in $\mathbb{C} = \mathbb{R}^2$.
- The eigenvectors corresponding to these clusters *should* define the division into clusters in the graph.

Spectral methods

General description

- We are looking for the eigenvalues of the adjacency matrix (or other related ones).
- We search for clusters of these eigenvalues in $\mathbb{C} = \mathbb{R}^2$.
- The eigenvectors corresponding to these clusters *should* define the division into clusters in the graph.

Cons:

- the method fails for sparse graphs.

I recommend read chapter VII in `https://arxiv.org/pdf/0906.0612`.

Methods based on the statistical inference

General description

- We assume that the considered network can be described with a stochastic block model.
- We are looking for the maximum likelihood estimator for the model parameters.

I recommend to read: `https://arxiv.org/abs/1008.3926`.

Methods based on the statistical inference

General description

- We assume that the considered network can be described with a stochastic block model.
- We are looking for the maximum likelihood estimator for the model parameters.

Cons:

- the method requires knowledge of the number of communities.

I recommend to read: `https://arxiv.org/abs/1008.3926`.

Diffusion-based methods

General description

- We generate a random walk trajectory on a given network.
- We try to do it optimally, which is equivalent to community detection.

I recommend to read: `https://arxiv.org/pdf/0707.0609.pdf`.

Diffusion-based methods

General description

- We generate a random walk trajectory on a given network.
- We try to do it optimally, which is equivalent to community detection.

Cons:

- it requires for random walker visiting each node of the network.

I recommend to read:
`https://arxiv.org/pdf/0707.0609.pdf`.

Different methods

- methods based on the spin dynamics,

Different methods

- methods based on the spin dynamics,
- optimization methods,

Different methods

- methods based on the spin dynamics,
- optimization methods,
- each of the methods presented has many variations!

Thank you for your attention!

MSc program in Data Science has been developed
as a part of task 10 of the project
„NERW PW. Science - Education - Development - Cooperation”
co-funded by European Union from European Social Fund.