# Modelowanie i analiza sieci złożonych

## X. Algorytmy wykrywania społeczności.

Grzegorz Siudem

Politechnika Warszawska

Zadanie 10 pn.
„Przygotowanie i uruchomienie nowego kierunku studiów na studiach II stopnia - Inżynieria i Analiza Danych (IAD)”

realizowane jest w ramach projektu
„NERW PW. Nauka – Edukacja – Rozwój – Współpraca”
współfinansowanego ze środków Unii Europejskiej
w ramach Europejskiego Funduszu Społecznego

# Przed zajęciami

# Przypomnienie

**Przypomnienie z innych zajęć:**

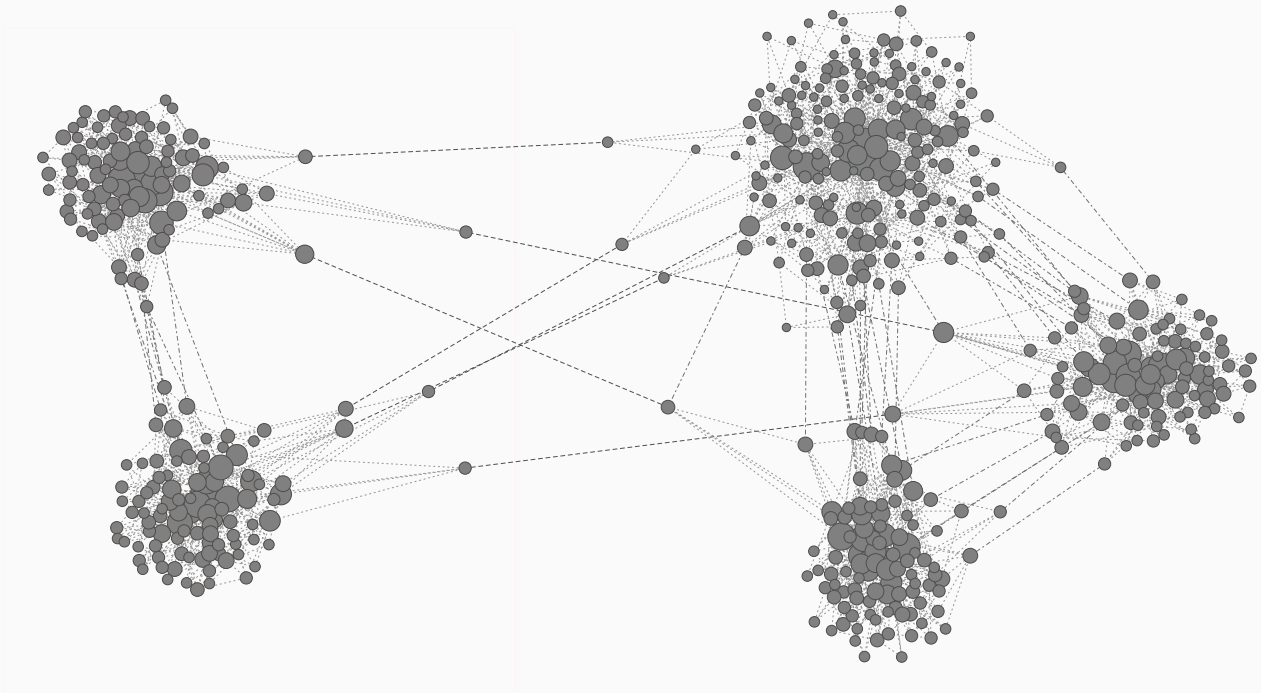
- Jak metody wykorzystuje się do wykrywania skupień w $R^n$?

**Przypomnienie z MASZ_9:**

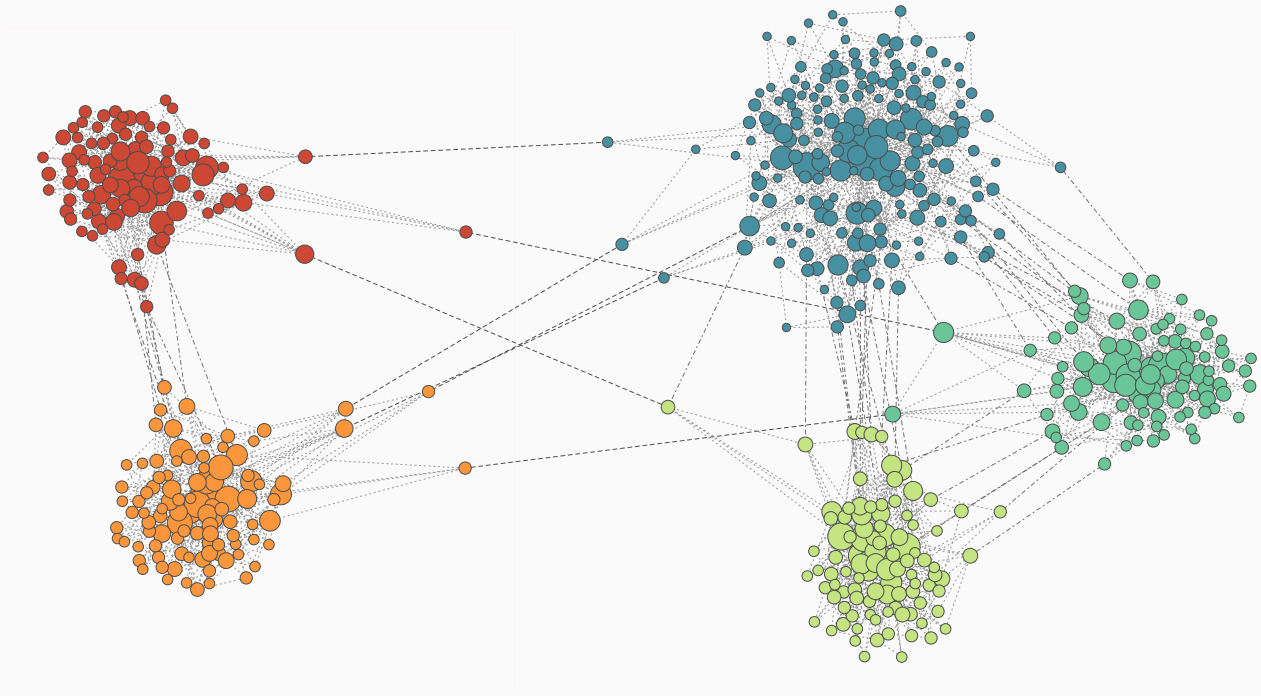
- Procesy Markowa - błądzenie na grafach.

# Wykład

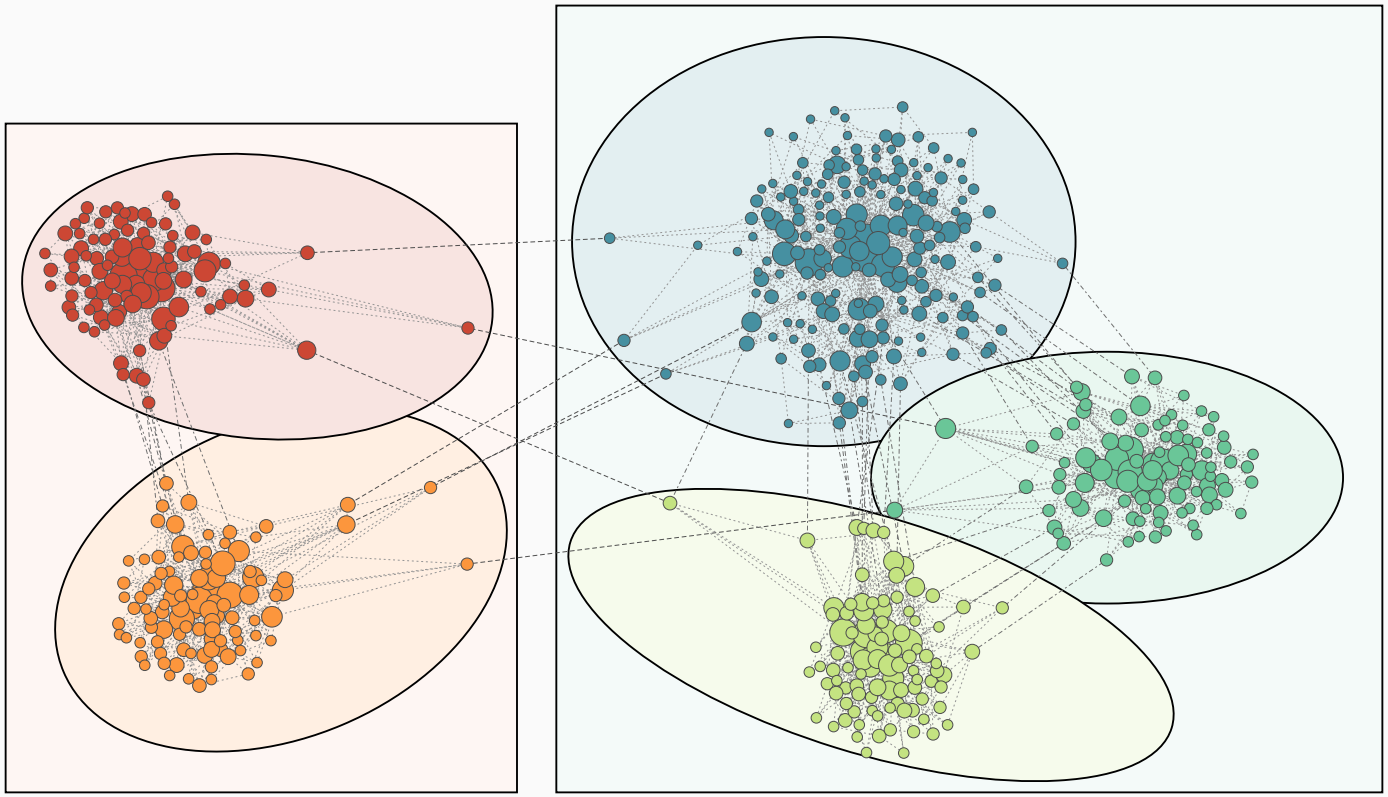
# Empiryczne uzasadnienie



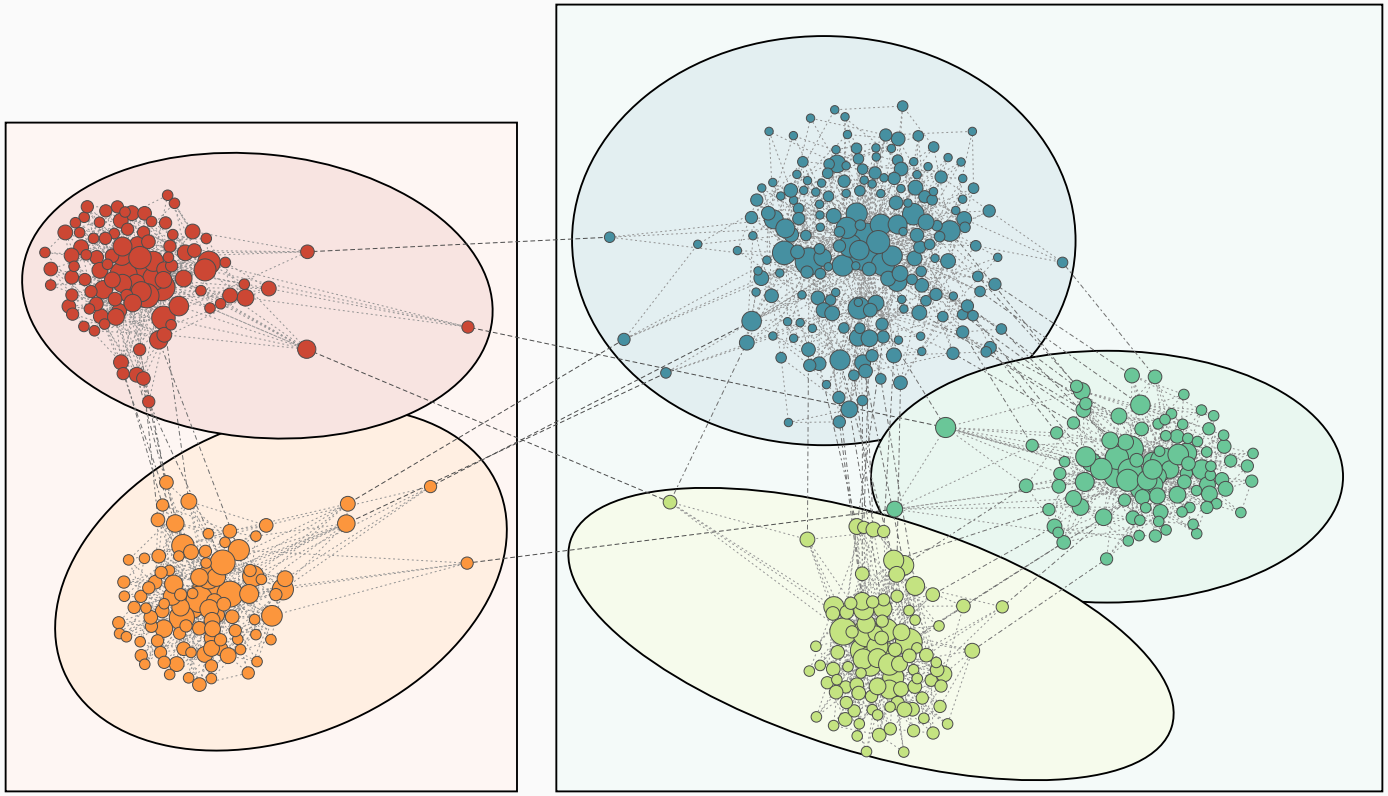
# Empiryczne uzasadnienie



# Empiryczne uzasadnienie



# Empiryczne uzasadnienie



**Dlaczego wykrywamy skupienia/społeczności?**

- poszukujemy istotnych cech elementów składowych,
- pytamy o liczbę tych składowych,
- szukamy hierarchii w analizowanym układzie.

# Problem z właściwym postawieniem problemu

- brak jednoznacznej i uniwersalnej definicji *czym* są społeczności,

# Problem z właściwym postawieniem problemu

- brak jednoznacznej i uniwersalnej definicji *czym* są społeczności,
- brak (w ogólności) apriorycznej metody ustalenia liczby społeczności dla danej sieci.

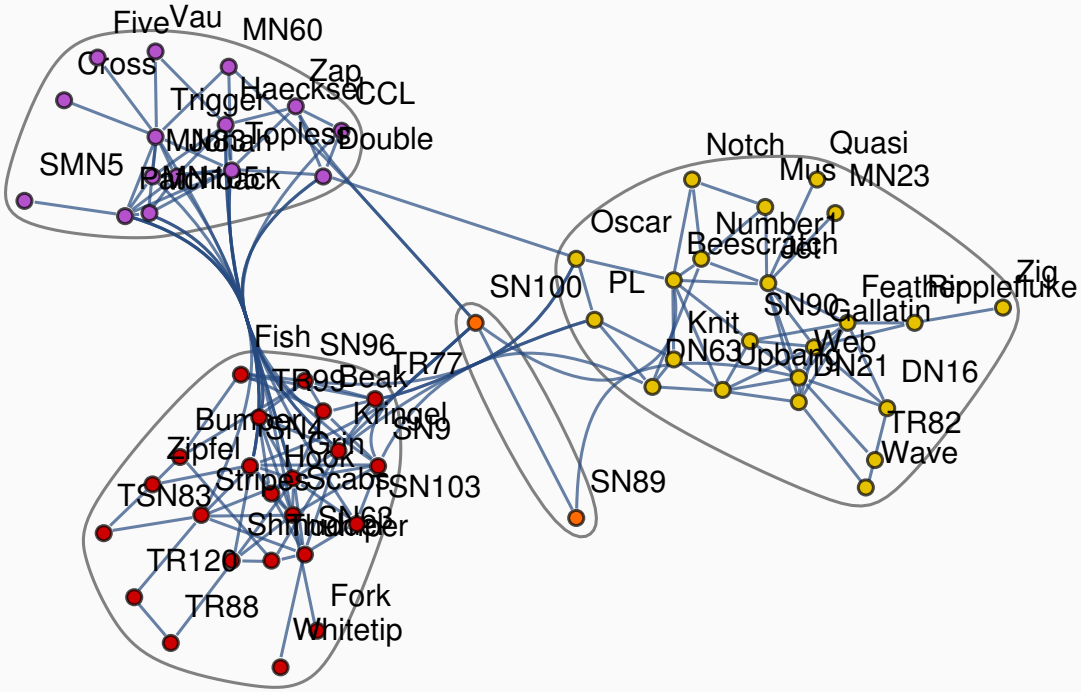
# Problem z właściwym postawieniem problemu

- brak jednoznacznej i uniwersalnej definicji *czym* są społeczności,
- brak (w ogólności) apriorycznej metody ustalenia liczby społeczności dla danej sieci.
- trudnodostępne (niemożliwe w ogólności?) są benchmarki metod detekcji.

# Problem z właściwym postawieniem problemu

- brak jednoznacznej i uniwersalnej definicji *czym* są społeczności,
- brak (w ogólności) apriorycznej metody ustalenia liczby społeczności dla danej sieci.
- trudnodostępne (niemożliwe w ogólności?) są benchmarki metod detekcji.

A jednak intuicyjnie problem jest zrozumiały



# Typowe metody detekcji społeczności

W dalszej części korzystam z

- S. Fortunato, D. Hric, Phys. Rep., **659**, 1, (2016).
- poza siecią akademicką pracę można znaleźć na arxiv-ie: arXiv:1608.00163.

# Typowe metody detekcji społeczności

W dalszej części korzystam z

- S. Fortunato, D. Hric, Phys. Rep., **659**, 1, (2016).
- poza siecią akademicką pracę można znaleźć na arxiv-ie: arXiv:1608.00163.

Osoby zainteresowane zachęcam do

- przejrzenia obfitej bibliografi *ibid.*
- ze szczególnym uwzględnieniem pracy `https://arxiv.org/abs/0906.0612`

# Typowe metody detekcji społeczności

W dalszej części korzystam z

- S. Fortunato, D. Hric, Phys. Rep., **659**, 1, (2016).
- poza siecią akademicką pracę można znaleźć na arxiv-ie: arXiv:1608.00163.

Osoby zainteresowane zachęcam do

- przejrzenia obfitej bibliografi *ibid.*
- ze szczególnym uwzględnieniem pracy `https://arxiv.org/abs/0906.0612`

Osobom bardzo zainteresowanym proponuję

- lekturę *społeczności* prac cytujących te monografie.

# Czym są społeczności?

# Czym są społeczności?

- klasycznie: rozbiciem zbioru wierzchołków.

# Czym są społeczności?

- klasycznie: rozbiciem zbioru wierzchołków.
- czasami dopuszczamy jednak przykrywanie się zbiorów.

# Czym są społeczności?

- klasycznie: rozbiciem zbioru wierzchołków.
- czasami dopuszczamy jednak przykrywanie się zbiorów.
- praktycznie: zbiorami, w których połączenia *do wewnątrz* są liczniejsze niż *na zewnątrz*.

# Czym są społeczności?

- klasycznie: rozbiciem zbioru wierzchołków.
- czasami dopuszczamy jednak przykrywanie się zbiorów.
- praktycznie: zbiorami, w których połączenia *do wewnątrz* są liczniejsze niż *na zewnątrz*.

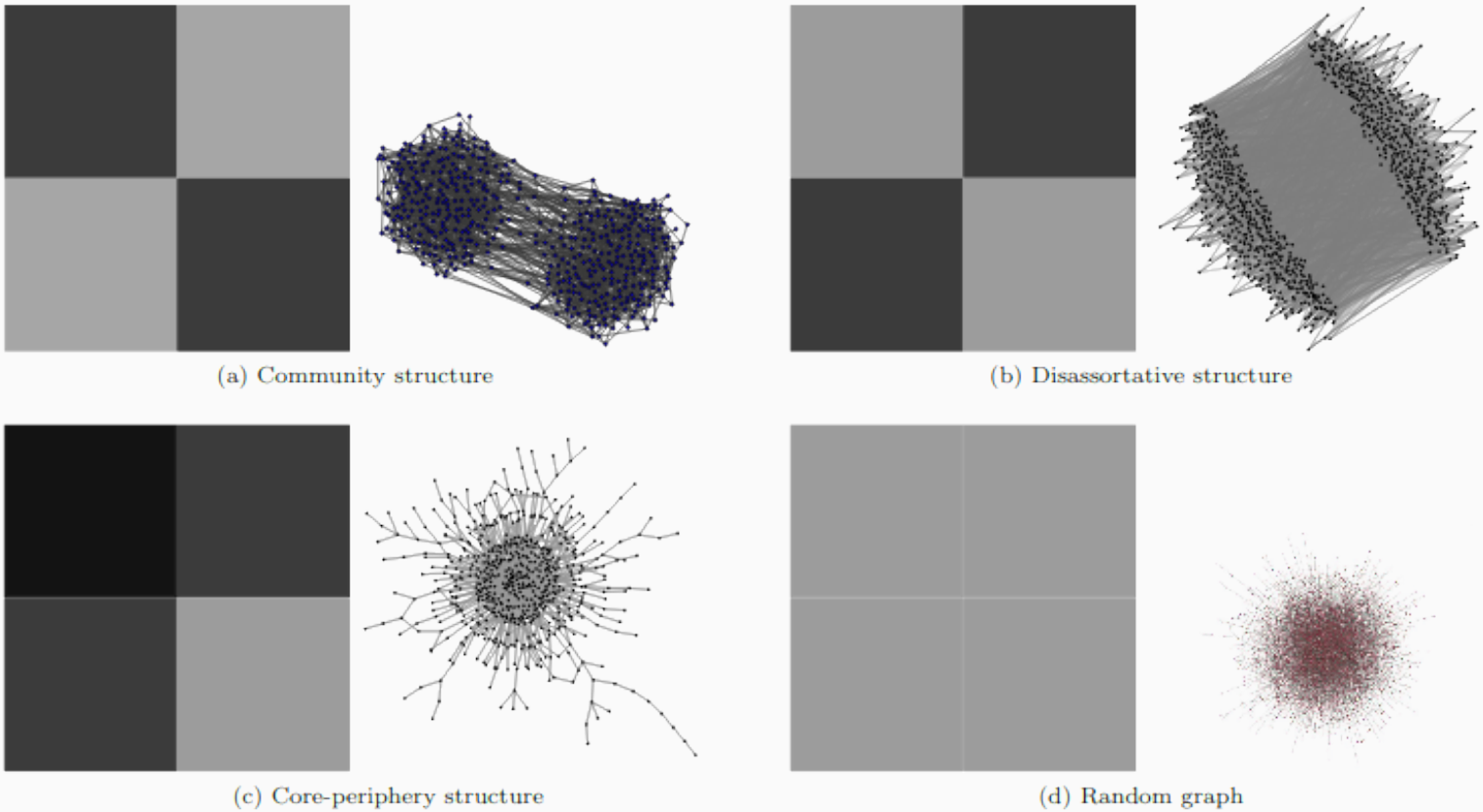
**Przypomnienie – prosty model sieci ze społecznościami**

Uogólniamy grafy Erdösa-Rényi do modelu blokowego (ang. *stochastic block model*).

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

- $K$ - liczba społeczności,
- $N > K$ liczba wierzchołków.

# Stochastic block model (z monografii Fortunato i Hrica)



(a) Community structure

(b) Disassortative structure

(c) Core-periphery structure

(d) Random graph

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). 

# Metody spektralne

## Ogólny opis

- Poszukujemy wartości własnych macierzy sąsiedztwa (lub innych powiązanych).
- Wyszukujemy skupień tych wartości własnych w $\mathbb{R}^2$.
- Wektory własne odpowiadające tym skupieniom *powinny* wyznaczać podział na klastry w grafie.

# Metody spektralne

## Ogólny opis

- Poszukujemy wartości własnych macierzy sąsiedztwa (lub innych powiązanych).
- Wyszukujemy skupień tych wartości własnych w $\mathbb{R}^2$.
- Wektory własne odpowiadające tym skupieniom *powinny* wyznaczać podział na klastry w grafie.

## Wady:

- metoda zawodzi dla rzadkich sieci.

Polecam lekturę: rozdział VII w
`https://arxiv.org/pdf/0906.0612`.

# Metody oparte o wnioskowanie statystyczne

## Ogólny opis

- Zakładamy, że rozważaną sieć można opisać modelem blokowym.
- Poszukujemy estymatora największej wiarogodności dla parametrów modelu.

Polecam lekturę: `https://arxiv.org/abs/1008.3926`.

# Metody oparte o wnioskowanie statystyczne

## Ogólny opis

- Zakładamy, że rozważaną sieć można opisać modelem blokowym.
- Poszukujemy estymatora największej wiarogodności dla parametrów modelu.

## Wady:

- metoda wymaga znajomości liczby społeczności.

Polecam lekturę: `https://arxiv.org/abs/1008.3926`.

# Metody oparte o dyfuzję

## Ogólny opis

- Generujemy ścieżkę błądzenia losowego na zadanej sieci.
- Próbujemy ją optymalnie zakodować, co jest równoważne poszukiwaniu podziału na społeczności.

Polecam lekturę: `https://arxiv.org/pdf/0707.0609.pdf`.

# Metody oparte o dyfuzję

## Ogólny opis

- Generujemy ścieżkę błądzenia losowego na zadanej sieci.
- Próbujemy ją optymalnie zakodować, co jest równoważne poszukiwaniu podziału na społeczności.

## Wady:

- wymaga *zwiedzania* całej sieci.

Polecam lekturę: `https://arxiv.org/pdf/0707.0609.pdf`.

# Inne metody

- metody oparte o dynamikę spinów,

# Inne metody

- metody oparte o dynamikę spinów,
- metody optymalizacyjne (wybór funkcji celu),

# Inne metody

- metody oparte o dynamikę spinów,
- metody optymalizacyjne (wybór funkcji celu),
- każda z przedstawionych metod posiada liczne wariacje!

# Inne metody

- metody oparte o dynamikę spinów,
- metody optymalizacyjne (wybór funkcji celu),
- każda z przedstawionych metod posiada liczne wariacje!

Dziękuję za uwagę!

Fundusze Europejskie
Wiedza Edukacja Rozwój

Politechnika Warszawska

Unia Europejska
Europejski Fundusz Społeczny

Zadanie 10 pn.
„Przygotowanie i uruchomienie nowego kierunku studiów na studiach II stopnia - Inżynieria i Analiza Danych (IAD)”

realizowane jest w ramach projektu
„NERW PW. Nauka – Edukacja – Rozwój – Współpraca”
współfinansowanego ze środków Unii Europejskiej
w ramach Europejskiego Funduszu Społecznego