Modelowanie i analiza sieci złożonych

V. Statyczne grafy przypadkowe

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

Do przypomnienia

- rozkład Bernoulliego $\mathbb{P}(X = k) = \binom{N}{k} p^k (1 - p)^{N-k}$,
- rozkład Piossona $\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$,
- zbieżność jednego do drugiego (w jakim sensie?)
- lemat o uściskach dłoni $\sum_i k_i = 2E$ (alternatywnie: $\langle k \rangle = 2\frac{E}{N}$.).

Pytanie

Dlaczego we wzorze pojawia sie czynnik 2?

Wykład

Przypomnienie czym jest sieciologia

$$\text{sieciologia} = \text{dane} + \text{metryki} + \text{modele} + \ldots$$

Przypomnienie czym jest sieciologia

$$\text{sieciologia} = \text{dane} + \text{metryki} + \text{modele} + \ldots$$

Po co nam modele?

- do generowania sieci w kontrolowanych warunkach,

Przypomnienie czym jest sieciologia

$$\text{sieciologia} = \text{dane} + \text{metryki} + \text{modele} + \ldots$$

Po co nam modele?

- do generowania sieci w kontrolowanych warunkach,
- pozwalają ekstrahować istotne z danego punktu widzenia cechy sieci,

Przypomnienie czym jest sieciologia

$$\text{sieciologia} = \text{dane} + \text{metryki} + \text{modele} + \ldots$$

Po co nam modele?

- do generowania sieci w kontrolowanych warunkach,
- pozwalają ekstrahować istotne z danego punktu widzenia cechy sieci,
- dają szansę na budowę dobrych statystyk przy analizie dynamiki na sieciach (zajęcia 11. i 12.),

$$\text{sieciologia} = \text{dane} + \text{metryki} + \text{modele} + \ldots$$

Po co nam modele?

- do generowania sieci w kontrolowanych warunkach,
- pozwalają ekstrahować istotne z danego punktu widzenia cechy sieci,
- dają szansę na budowę dobrych statystyk przy analizie dynamiki na sieciach (zajęcia 11. i 12.),
- dają nam wgląd w mechanizmy stojące za rzeczywistymi procesami (patrz kolejne zajęcia i sieci BA),

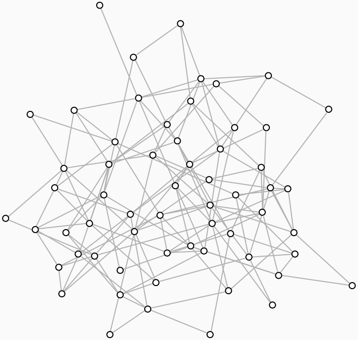
Przypomnienie czym jest sieciologia

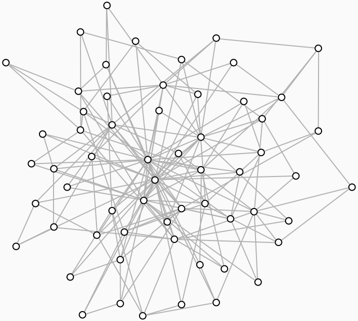
$$\text{sieciologia} = \text{dane} + \text{metryki} + \text{modele} + \ldots$$

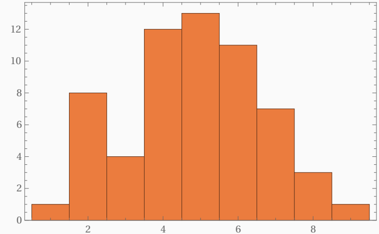
Po co nam modele?

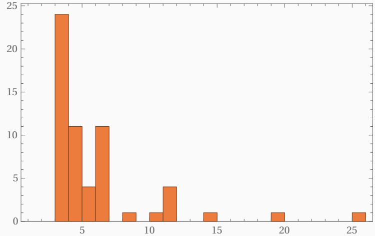
- do generowania sieci w kontrolowanych warunkach,
- pozwalają ekstrahować istotne z danego punktu widzenia cechy sieci,
- dają szansę na budowę dobrych statystyk przy analizie dynamiki na sieciach (zajęcia 11. i 12.),
- dają nam wgląd w mechanizmy stojące za rzeczywistymi procesami (patrz kolejne zajęcia i sieci BA),
- to bardzo wdzięczna dziedzina matematyki stosowanej.

Uwaga! Grafy ER są *normalne*



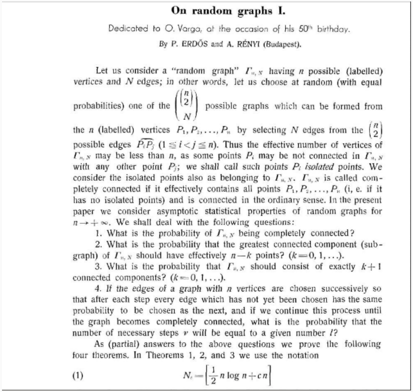






A tak naprawdę poissonowskie, jak przekonamy się za chwilę...

Model Erdősa-Rényiego



On random graphs I.

Dedicated to O. Varga, at the occasion of his 50th birthday.

By P. ERDŐS and A. RÉNYI (Budapest).

Let us consider a "random graph" $\Gamma_{n,N}$ having n possible (labelled) vertices and N edges; in other words, let us choose at random (with equal probabilities) one of the $\binom{\binom{n}{2}}{N}$ possible graphs which can be formed from the n (labelled) vertices $P_1, P_2, \ldots, P_n$ by selecting N edges from the $\binom{n}{2}$ possible edges $\widehat{P_i P_j}$ $(1 \leq i < j \leq n)$. Thus the effective number of vertices of $\Gamma_{n,N}$ may be less than n, as some points P_i may be not connected in $\Gamma_{n,N}$ with any other point P_j; we shall call such points P_i *isolated* points. We consider the isolated points also as belonging to $\Gamma_{n,N}$. $\Gamma_{n,N}$ is called completely connected if it effectively contains all points $P_1, P_2, \ldots, P_n$ (i. e. if it has no isolated points) and is connected in the ordinary sense. In the present paper we consider asymptotic statistical properties of random graphs for $n \to +\infty$. We shall deal with the following questions:

1. What is the probability of $\Gamma_{n,N}$ being completely connected?

2. What is the probability that the greatest connected component (subgraph) of $\Gamma_{n,N}$ should have effectively $n-k$ points? $(k = 0, 1, \ldots)$.

3. What is the probability that $\Gamma_{n,N}$ should consist of exactly $k+1$ connected components? $(k = 0, 1, \ldots)$.

4. If the edges of a graph with n vertices are chosen successively so that after each step every edge which has not yet been chosen has the same probability to be chosen as the next, and if we continue this process until the graph becomes completely connected, what is the probability that the number of necessary steps ν will be equal to a given number l?

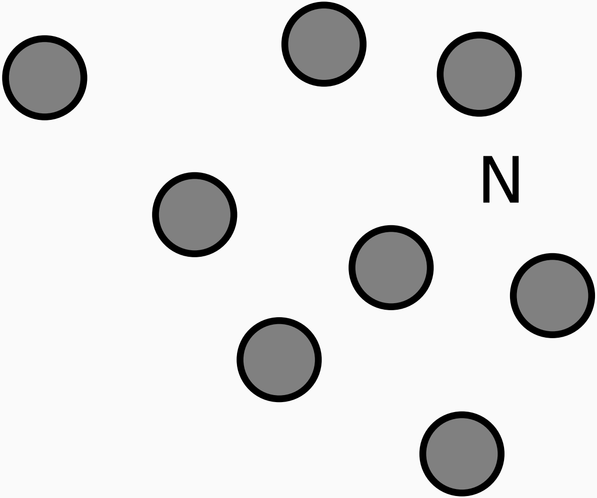
As (partial) answers to the above questions we prove the following four theorems. In Theorems 1, 2, and 3 we use the notation

$$N_c = \left[\frac{1}{2} n \log n + c\,n\right] \tag{1}$$

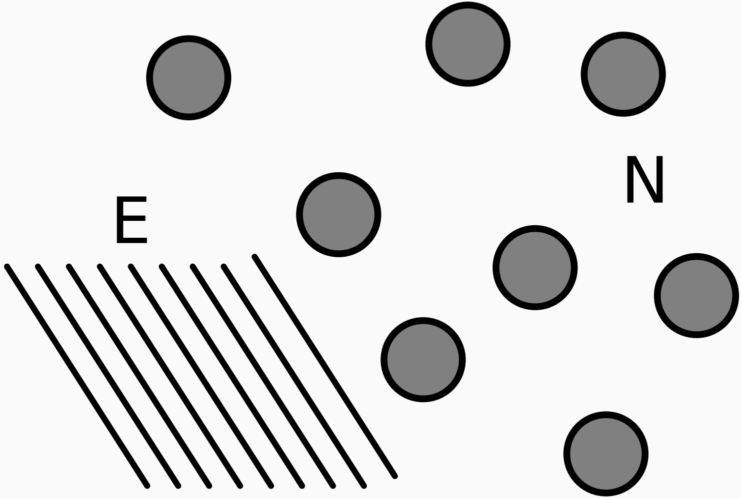
Dwa warianty modelu

- $G_{N,E}$ (Erdős-Rényi),
- $G_{N,p}$ (E. Gilbert)

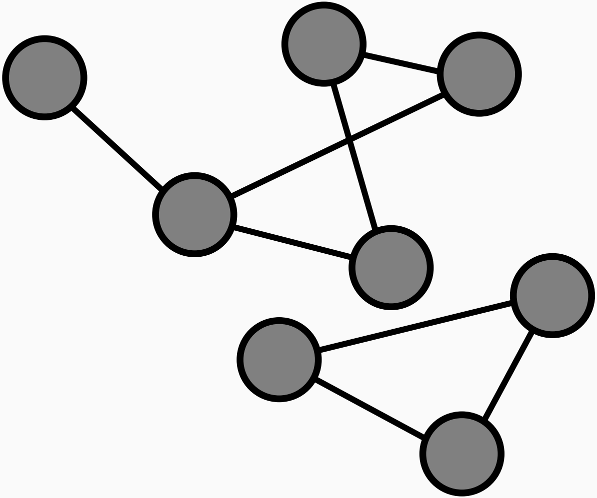
Model Erdősa-Rényiego w wersji $G_{N,E}$



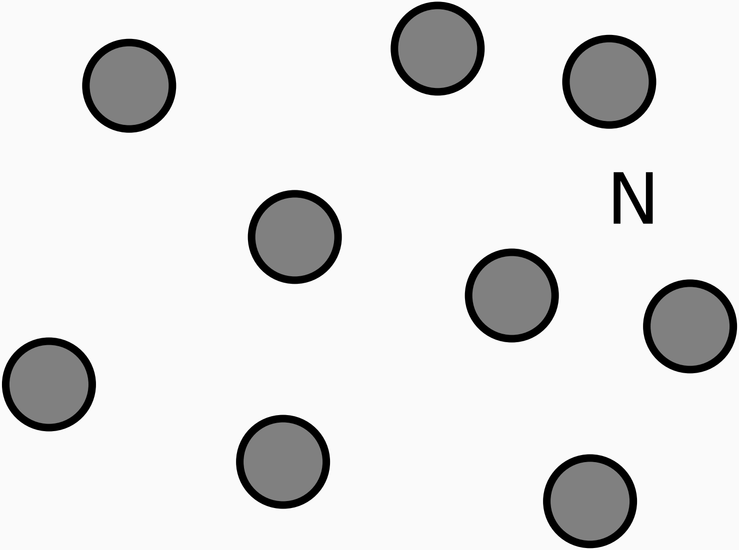
Model Erdősa-Rényiego w wersji $G_{N,E}$



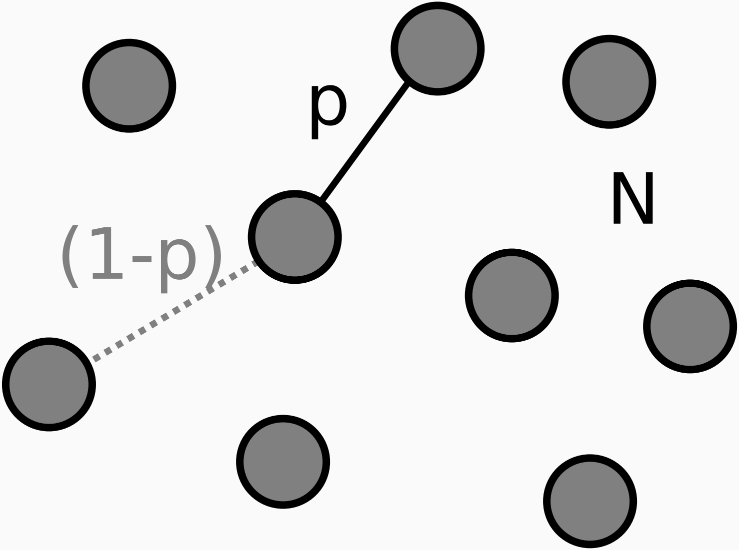
Model Erdősa-Rényiego w wersji $G_{N,E}$



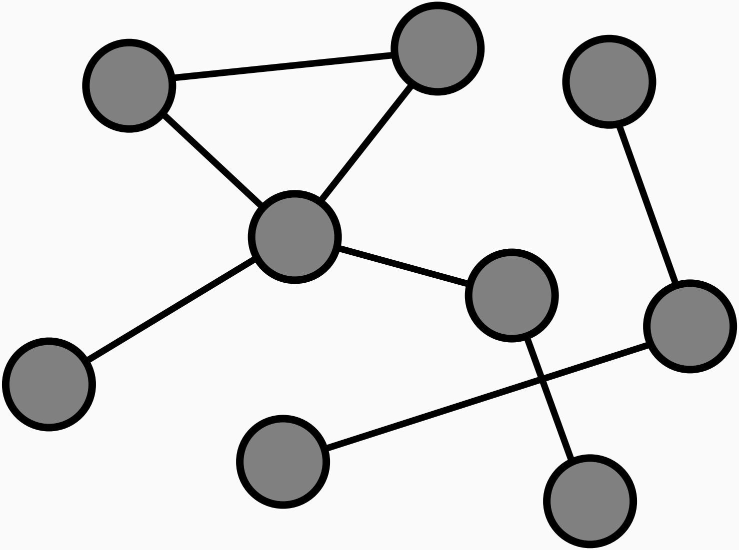
Model Erdősa-Rényiego w wersji $G_{N,p}$



Model Erdősa-Rényiego w wersji $G_{N,p}$



Model Erdősa-Rényiego w wersji $G_{N,p}$

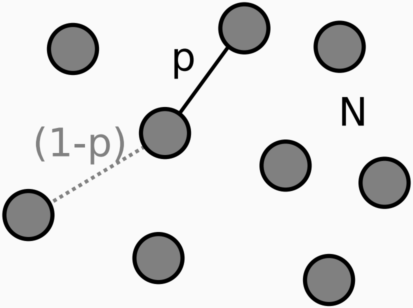


Skupmy się na modelu $G_{N,p}$

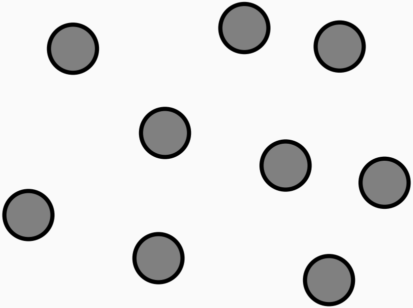
Założenia

Analizujemy ansambl opisany dwoma parametrami:

- N liczba wierzchołków,
- p prawdopodobieństwo, że dwa wierzchołki będą połączone.



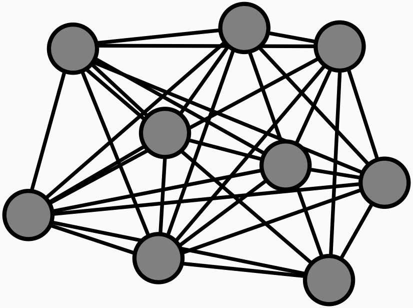
Przypadek graniczny: $p = 0$



Wniosek:

$p = 0$ to przypadek trywialny.

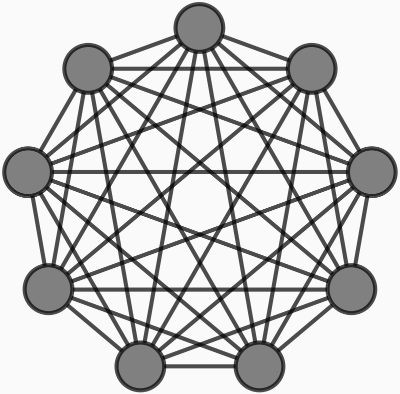
Przypadek graniczny – $p = 1$



Wniosek 1:

To bardzo nieudana wizualizacja...

Przypadek graniczny – $p = 1$



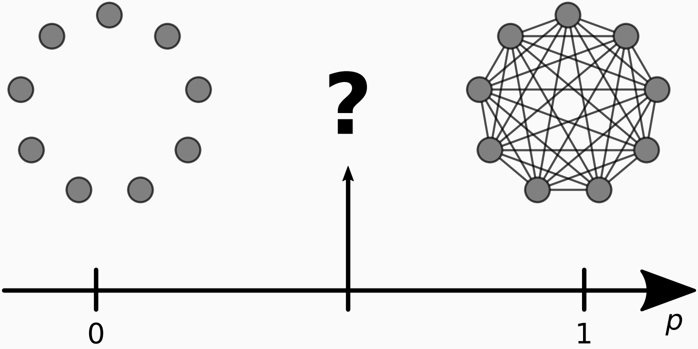
Wniosek 1:

~~To bardzo nieudana wizualizacja...~~

Wniosek 2:

$p = 1$ to graf pełny.

Grafy ER – zależność od parametru p



Rozwiązanie modelu $G_{N,p}$

Pytanie

Ile *przeciętnie* pojawi się krawędzi w jednym układzie?

Rozwiązanie modelu $G_{N,p}$

Pytanie

Ile *przeciętnie* pojawi się krawędzi w jednym układzie?

Policzmy

$$\langle E \rangle = p \times \frac{N(N-1)}{2},$$

co z lematu o uścisku dłoni daje

$$\langle k \rangle = p \frac{N(N-1)}{2} \frac{2}{N} = p(N-1) \approx pN.$$

Rozwiązanie modelu $G_{N,p}$

Pytanie

Ile *przeciętnie* pojawi się krawędzi w jednym układzie?

Policzmy

$$\langle E \rangle = p \times \frac{N(N-1)}{2},$$

co z lematu o uścisku dłoni daje

$$\langle k \rangle = p \frac{N(N-1)}{2} \frac{2}{N} = p(N-1) \approx pN.$$

Znamy zatem średni stopień

A co z rozkładem stopni wierzchołków?

Rozwiązanie modelu $G_{N,p}$

Zmienna losowa opisująca stopień wierzchołka K

$$K = \sum_{i=1}^{N-1} X_k,$$

gdzie X_k są zmiennymi iid o $\mathbb{P}(X_k = 1) = p$ $\mathbb{P}(X_k = 0) = 1 - p$.

Rozwiązanie modelu $G_{N,p}$

Zmienna losowa opisująca stopień wierzchołka K

$$K = \sum_{i=1}^{N-1} X_k,$$

gdzie X_k są zmiennymi iid o $\mathbb{P}(X_k = 1) = p$ $\mathbb{P}(X_k = 0) = 1 - p$.

K ma rozkład Bernoulliego

$$\mathcal{P}(k) = \mathbb{P}(K = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

Rozwiązanie modelu $G_{N,p}$

Zmienna losowa opisująca stopień wierzchołka K

$$K = \sum_{i=1}^{N-1} X_k,$$

gdzie X_k są zmiennymi iid o $\mathbb{P}(X_k = 1) = p$ $\mathbb{P}(X_k = 0) = 1 - p$.

K ma rozkład Bernoulliego

$$\mathcal{P}(k) = \mathbb{P}(K = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

Przybliżenie rozkładem Poissona

$$\mathcal{P}(k) \approx \frac{e^{-\langle k \rangle} \langle k \rangle^k}{k!},$$

bo $\langle k \rangle \approx Np$.

Niefizyczność grafów ER

Przyjmując przybliżenie poissonowskie wyznaczmy wariancję

$$\mathbb{E}(K) = \sum_{k=0}^{\infty} \frac{k e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle,$$

$$\mathbb{E}(K^2) = \sum_{k=0}^{\infty} \frac{k^2 e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle + \langle k \rangle^2.$$

$$\mathrm{Var}(K) = \mathbb{E}(K^2) - [\mathbb{E}(K)]^2 = \langle k \rangle^2$$

Niefizyczność grafów ER

Przyjmując przybliżenie poissonowskie wyznaczmy wariancję

$$\mathbb{E}(K) = \sum_{k=0}^{\infty} \frac{k e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle,$$

$$\mathbb{E}(K^2) = \sum_{k=0}^{\infty} \frac{k^2 e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle + \langle k \rangle^2.$$

$$\mathrm{Var}(K) = \mathbb{E}(K^2) - [\mathbb{E}(K)]^2 = \langle k \rangle^2$$

A szczegóły rachunków?

W części projektowej.

Niefizyczność grafów ER

Przyjmując przybliżenie poissonowskie wyznaczmy wariancję

$$\mathbb{E}(K) = \sum_{k=0}^{\infty} \frac{k e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle,$$

$$\mathbb{E}(K^2) = \sum_{k=0}^{\infty} \frac{k^2 e^{-\langle k \rangle} \langle k \rangle^k}{k!} = \cdots = \langle k \rangle + \langle k \rangle^2.$$

$$\mathrm{Var}(K) = \mathbb{E}(K^2) - [\mathbb{E}(K)]^2 = \langle k \rangle^2$$

A szczegóły rachunków?

W części projektowej.

Tak mała wariancja nie pojawia się w danych rzeczywistych...

Niefizyczność grafów ER

Przypomnijmy współczynnik gronowania

Wyznaczymy wartość współczynnika gronowania dla grafów ER

$$C_i = \frac{2E_i}{k_i(k_i - 1)}.$$

Niefizyczność grafów ER

Przypomnijmy współczynnik gronowania

Wyznaczymy wartość współczynnika gronowania dla grafów ER

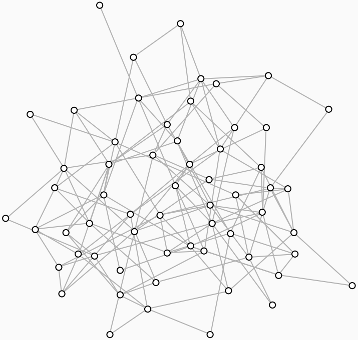
$$C_i = \frac{2E_i}{k_i(k_i - 1)}.$$

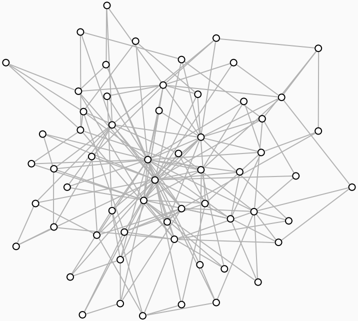
Odpowiedź:

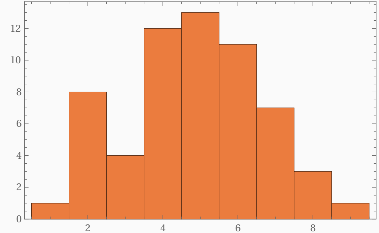
$$\langle C \rangle = \frac{p\langle k \rangle(\langle k \rangle - 1)}{\langle k \rangle(\langle k \rangle - 1)} = p.$$

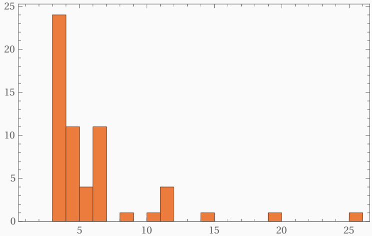
Tak mały współczynnik gronowania raczej nie pojawia się w danych rzeczywistych...

Niefizyczność grafów ER







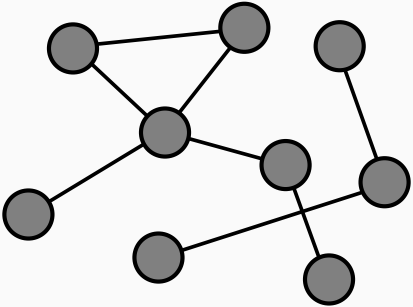


Na kolejnych zajęciach bardziej rzeczywisty model BA.↗

Model konfiguracyjny

Pomysł:

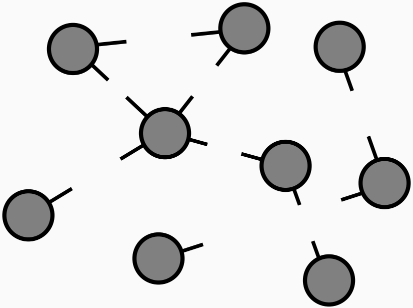
Zbudujmy graf z klocków, jakie mamy.



Model konfiguracyjny

Pomysł:

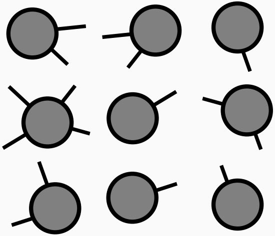
Zbudujmy graf z klocków, jakie mamy.



Model konfiguracyjny

Pomysł:

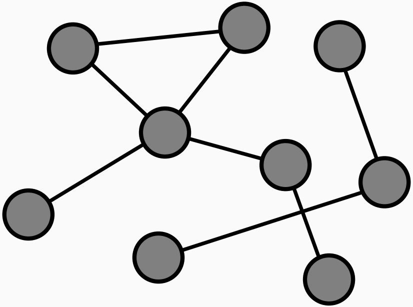
Zbudujmy graf z klocków, jakie mamy.



Model konfiguracyjny

Pomysł:

Zbudujmy graf z klocków, jakie mamy.



A gdyby zastąpić rzeczywiste klocki dobranymi *a priori*?

Model konfiguracyjny

Wybierzmy konfigurację stopni zgodnie z oczekiwanym rozkładem:

Na przykład:

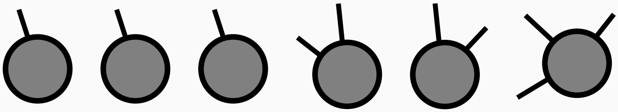
$$\{1, 1, 1, 2, 2, 3\}$$

Model konfiguracyjny

Wybierzmy konfigurację stopni zgodnie z oczekiwanym rozkładem:

Na przykład:

$$\{1, 1, 1, 2, 2, 3\}$$



Pozostaje tylko poskładać te klocki w sieć...

Tylko czy to zawsze jest wykonalne?

Model konfiguracyjny

Wybierzmy konfigurację stopni zgodnie z oczekiwanym rozkładem:

Na przykład:

$$\{1, 1, 1, 2, 2, 3\}$$

Pozostaje tylko poskładać te klocki w sieć...

Tylko czy to zawsze jest wykonalne?

Pytanie:

Znajdź (mały) kontrprzykład na to, że proponowanej procedury nie zawsze uda się ukończyć.

Model blokowy

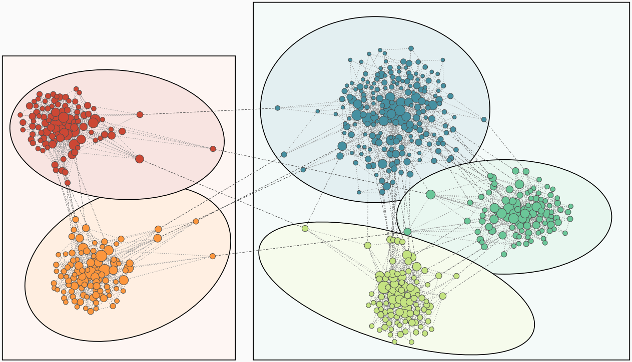
Uogólnienie grafów ER

$$\begin{bmatrix} [p_{11}] & [p_{12}] & \dots & [p_{1N}] \\ [p_{21}] & [p_{22}] & \dots & [p_{2N}] \\ \dots & \dots & \ddots & \dots \\ [p_{N1}] & [p_{N2}] & \dots & [p_{NN}] \end{bmatrix}$$

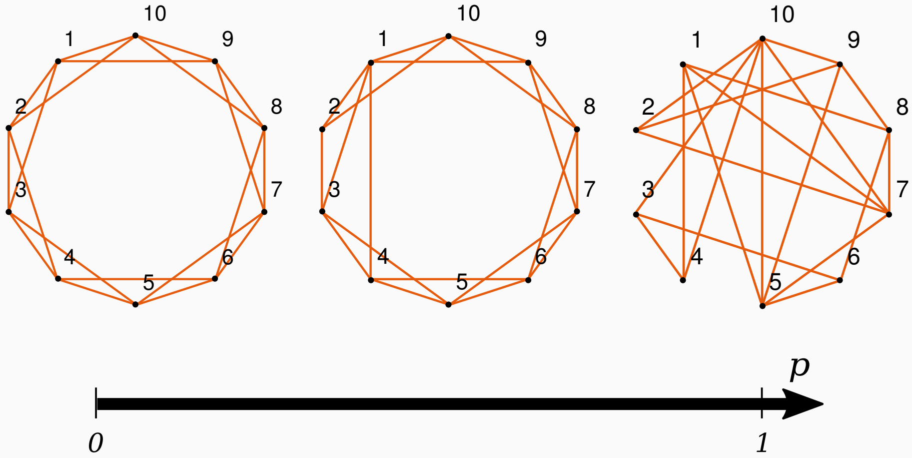
Model blokowy

Uogólnienie grafów ER

$$
\begin{bmatrix}
[p_{11}] & [p_{12}] & \dots & [p_{1N}] \\
[p_{21}] & [p_{22}] & \dots & [p_{2N}] \\
\dots & \dots & \ddots & \dots \\
[p_{N1}] & [p_{N2}] & \dots & [p_{NN}]
\end{bmatrix}
$$



Model Wattsa-Strogatza



Sieci o zadanym hamiltonianie

Rozważmy przestrzeń wszystkich grafów o N wierzchołkach

czyli zbiór $M_N = \mathbb{M}^{N \times N}(\{0, 1\})$. Chcemy zdefiniować na nim rozkład prawdopodobieństwa.

Sieci o zadanym hamiltonianie

Rozważmy przestrzeń wszystkich grafów o N wierzchołkach

czyli zbiór $M_N = \mathbb{M}^{N\times N}(\{0,1\})$. Chcemy zdefiniować na nim rozkład prawdopodobieństwa.

Maksymalizujemy entropię

$$-\sum_{G\in M_N} \mathcal{P}(G)\ln\mathcal{P}(G),$$

Sieci o zadanym hamiltonianie

Rozważmy przestrzeń wszystkich grafów o N wierzchołkach

czyli zbiór $M_N = \mathbb{M}^{N\times N}(\{0,1\})$. Chcemy zdefiniować na nim rozkład prawdopodobieństwa.

Maksymalizujemy entropię

$$-\sum_{G\in M_N} \mathcal{P}(G)\ln\mathcal{P}(G),$$

Pod pewnym warunkiem $f(\mathcal{P}(G)) = 0$,

Sieci o zadanym hamiltonianie

Rozważmy przestrzeń wszystkich grafów o N wierzchołkach czyli zbiór $M_N = \mathbb{M}^{N\times N}(\{0,1\})$. Chcemy zdefiniować na nim rozkład prawdopodobieństwa.

Maksymalizujemy entropię

$$-\sum_{G\in M_N} \mathcal{P}(G) \ln \mathcal{P}(G),$$

Pod pewnym warunkiem $f(\mathcal{P}(G)) = 0$,

Co prowadzi do metody mnożników Lagrange'a

$$\mathcal{L}[\mathcal{P}(G)] = -\sum_{G\in M_N} \mathcal{P}(G) \ln \mathcal{P}(G) + \lambda f(\mathcal{P}(G))$$

Sieci o zadanym hamiltonianie

Rozważmy przestrzeń wszystkich grafów o N wierzchołkach czyli zbiór $M_N = \mathbb{M}^{N\times N}(\{0,1\})$. Chcemy zdefiniować na nim rozkład prawdopodobieństwa.

Maksymalizujemy entropię

$$-\sum_{G\in M_N} \mathcal{P}(G)\ln\mathcal{P}(G),$$

Pod pewnym warunkiem $f(\mathcal{P}(G)) = 0$,

Co prowadzi do metody mnożników Lagrange'a

$$\mathcal{L}[\mathcal{P}(G)] = -\sum_{G\in M_N} \mathcal{P}(G)\ln\mathcal{P}(G) + \lambda f(\mathcal{P}(G))$$

Pozostaje *tylko* rozwiązać

$$\frac{\partial\mathcal{L}}{\partial\mathcal{P}(G)} = 0.$$

Podsumowanie

Na następne zajęcia proponuję:

Poczytać o

- Zasadzie Dulbecco,
- Regule św. Mateusza,
- zasadzie *rich get richer*,
- procesach Yule'a.

Albo przeczytać:

- M. Perc, *Journal of The Royal Society Interface* **11** (2014)

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

Dziękuję za uwagę!