Modelowanie i analiza sieci złożonych

IV. Metryki sieci

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

Przypomnij sobie

Co to jest odległość w grafie

$$d(i,j) = ?$$

Poznane już metryki sieci

- średni stopień wierzchołka

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i,$$

- średnia droga w grafie

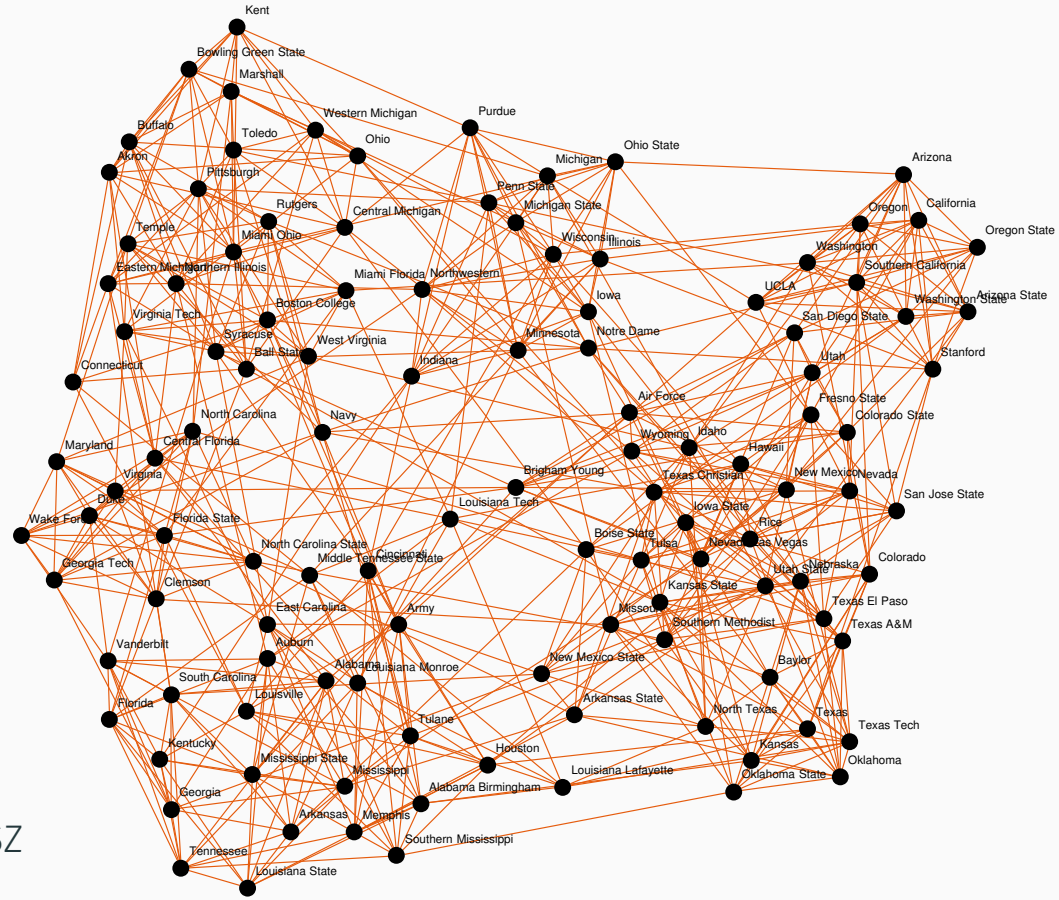
$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j).$$

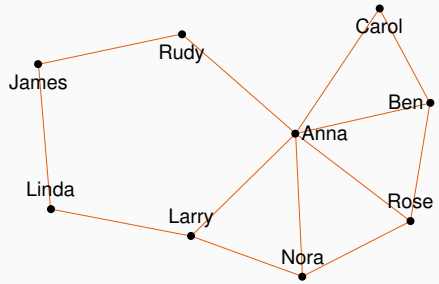
Lemat o uścisku dłoni

Wykład

Jakie cechy sieci możemy zmierzyć?

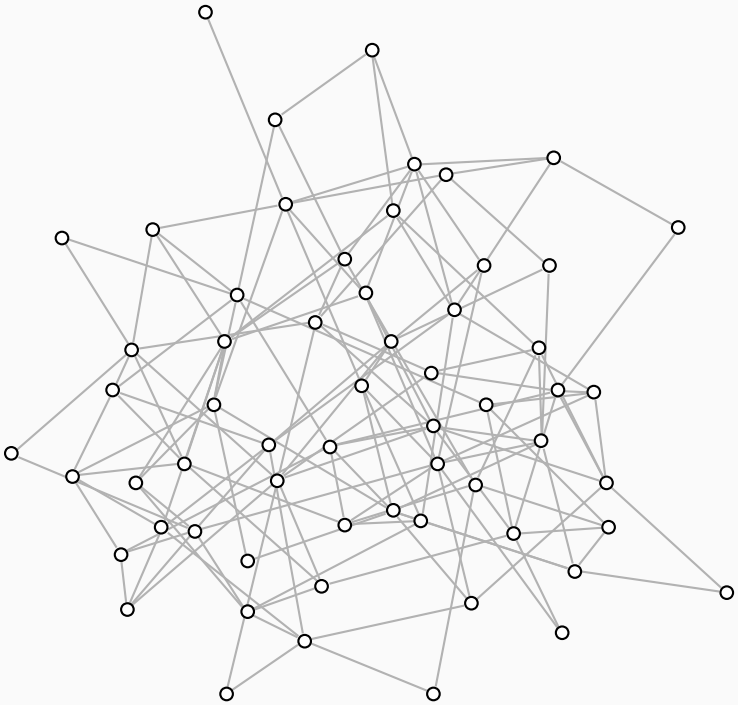
- jak duża jest sieć?

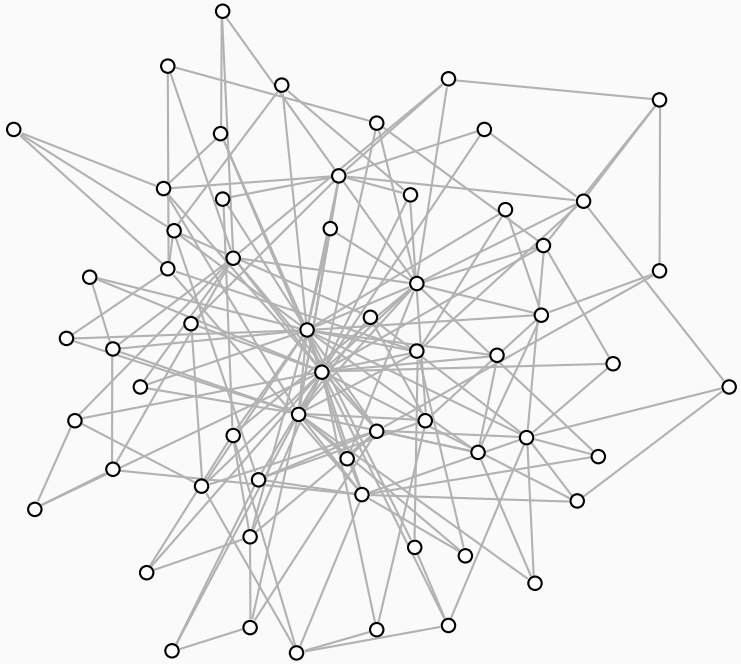




Jakie cechy sieci możemy zmierzyć?

- jak duża jest sieć?
- jak gęsta jest sieć?





Jakie cechy sieci możemy zmierzyć?

- jak duża jest sieć?
- jak gęsta jest sieć?
- jaka jest topologia połączeń?

?

Uwaga!

W sieciologii topologia ma nieco inne znaczenie niż w matematyce!

Jakie znamy potoczne metryki/miary sieci?

Naturalne/naiwne metryki:

- liczba wierzchołków N (rozmiar),

Jakie znamy potoczne metryki/miary sieci?

Naturalne/naiwne metryki:

- liczba wierzchołków N (rozmiar),
- liczba krawędzi E (rozmiar, gęstość),

Jakie znamy potoczne metryki/miary sieci?

Naturalne/naiwne metryki:

- liczba wierzchołków N (rozmiar),
- liczba krawędzi E (rozmiar, gęstość),
- dlaczego $\langle k \rangle = 2E/N$? (gęstość)

Jakie znamy potoczne metryki/miary sieci?

Naturalne/naiwne metryki:

- liczba wierzchołków N (rozmiar),
- liczba krawędzi E (rozmiar, gęstość),
- dlaczego $\langle k \rangle = 2E/N$? (gęstość)
- największy stopień węzła (celebrytki/celebryci?),

Jakie znamy potoczne metryki/miary sieci?

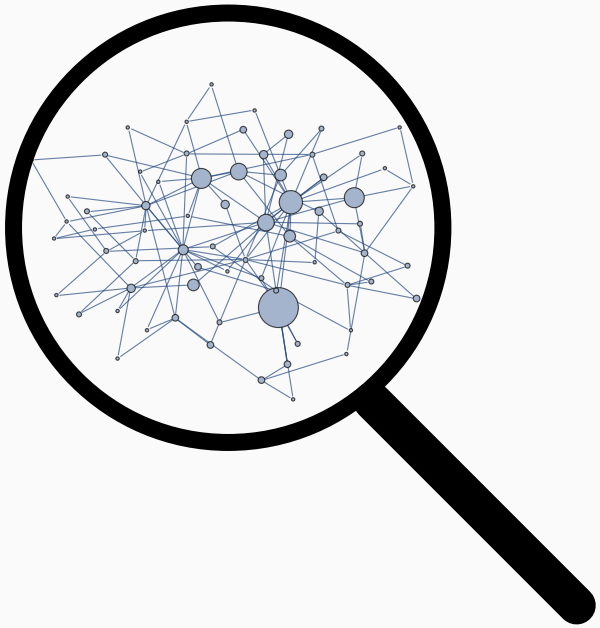
Naturalne/naiwne metryki:

- liczba wierzchołków N (rozmiar),
- liczba krawędzi E (rozmiar, gęstość),
- dlaczego $\langle k \rangle = 2E/N$? (gęstość)
- największy stopień węzła (celebrytki/celebryci?),
- inne?

Jakie znamy potoczne metryki/miary sieci?

Naturalne/naiwne metryki:

- liczba wierzchołków N (rozmiar),
- liczba krawędzi E (rozmiar, gęstość),
- dlaczego $\langle k \rangle = 2E/N$? (gęstość)
- największy stopień węzła (celebrytki/celebryci?),
- inne?



Bardziej formalne metody opisu sieci – rozkłady

Na zajęciach mówiliśmy już o

- rozkładzie stopni wierzchołków $\mathcal{P}(k)$,

Bardziej formalne metody opisu sieci – rozkłady

Na zajęciach mówiliśmy już o

- rozkładzie stopni wierzchołków $\mathcal{P}(k)$,
- w szczególności o rozkładach potęgowych z parametrem α

$$\mathcal{P}(k) \propto k^{-\alpha},$$

- jak jednak estymować α?

Bardziej formalne metody opisu sieci – rozkłady

Na zajęciach mówiliśmy już o

- rozkładzie stopni wierzchołków $\mathcal{P}(k)$,
- w szczególności o rozkładach potęgowych z parametrem α

$$\mathcal{P}(k) \propto k^{-\alpha},$$

- jak jednak estymować α?

Kto odrobił pracę domową?

- M.E.J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics, **46**, 323-351 (2005).
  i/lub
- rozdziały 3.1-3.3 w A. Fronczak, P. Fronczak, *Świat sieci złożonych* PWN (2009).

Bardziej formalne metody opisu sieci – rozkłady

Na zajęciach mówiliśmy już o

- rozkładzie stopni wierzchołków $\mathcal{P}(k)$,
- w szczególności o rozkładach potęgowych z parametrem α

$$\mathcal{P}(k) \propto k^{-\alpha},$$

- jak jednak estymować α?

Kto odrobił pracę domową?

- M.E.J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics, **46**, 323-351 (2005).
  i/lub
- rozdziały 3.1-3.3 w A. Fronczak, P. Fronczak, *Świat sieci złożonych* PWN (2009).

Więcej na ten temat w części projektowej.

Miary korelacji węzłów

Znamy już:

Korelacje dwuwęzłowe $\mathcal{R}(k_i, k_j)$ czyli prawdopodobieństwo, że losowo wybrana krawędź łączy węzły o stopniach k_i i k_j

$$\mathcal{R}(k_i, k_j) = \frac{P(k_i, k_j)}{P_u(k_i, k_j)},$$

a P_u odpowiada sieci nieskorelowanej o tym samym rozkładzie.
Niestety nie jest to zbyt praktyczne narzędzie.

Miary korelacji węzłów

Znamy już:

Korelacje dwuwęzłowe $\mathcal{R}(k_i, k_j)$ czyli prawdopodobieństwo, że losowo wybrana krawędź łączy węzły o stopniach k_i i k_j

$$\mathcal{R}(k_i, k_j) = \frac{P(k_i, k_j)}{P_u(k_i, k_j)},$$

a P_u odpowiada sieci nieskorelowanej o tym samym rozkładzie.
Niestety nie jest to zbyt praktyczne narzędzie.

Słyszeliśmy także o:

Prawdopodobieństwie warunkowym

$$\mathcal{P}(k_i|k_j) = \frac{\mathcal{P}(k_i, k_j)}{k_j \mathcal{P}(k_j)/\langle k \rangle}$$

Czy to się *dobrze* estymuje? Niestety nie...

Jak obniżyć korelacje?

Losowe przełączanie węzłów:

Zachowuje stopnie wierzchołków!

Czemu to robimy?

- żeby pozbyć się niechcianych korelacji,
- żeby określić ich znaczenie dla danej sieci,
- żeby zyskać model referencyjny o tym samym rozkładzie.

Miary korelacji węzłów cd.

Wprowadźmy:

Średni stopień najbliższego węzła (dla węzła o stopniu k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

Pytamy o zależność średniego stopnia najbliższego sąsiada $\langle k \rangle_{nn}$ od stopnia wierzchołka k_i

Miary korelacji węzłów cd.

Wprowadźmy:

Średni stopień najbliższego węzła (dla węzła o stopniu k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

Pytamy o zależność średniego stopnia najbliższego sąsiada $\langle k \rangle_{nn}$ od stopnia wierzchołka k_i

Co mierzy ta miara?

- jeśli $\langle k \rangle_{nn}(k_i)$ jest funkcją rosnącą wówczas sieć jest asortatywna.

Miary korelacji węzłów cd.

Wprowadźmy:

Średni stopień najbliższego węzła (dla węzła o stopniu k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

Pytamy o zależność średniego stopnia najbliższego sąsiada $\langle k \rangle_{nn}$ od stopnia wierzchołka k_i

Co mierzy ta miara?

- jeśli $\langle k \rangle_{nn}(k_i)$ jest funkcją rosnącą wówczas sieć jest asortatywna.
- jeśli $\langle k \rangle_{nn}(k_i)$ jest funkcją malejącą wówczas sieć jest dysasortatywna.

Miary korelacji węzłów cd.

Wprowadźmy:

Średni stopień najbliższego węzła (dla węzła o stopniu k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

Pytamy o zależność średniego stopnia najbliższego sąsiada $\langle k \rangle_{nn}$ od stopnia wierzchołka k_i

Co mierzy ta miara?

- jeśli $\langle k \rangle_{nn}(k_i)$ jest funkcją rosnącą wówczas sieć jest asortatywna.
- jeśli $\langle k \rangle_{nn}(k_i)$ jest funkcją malejącą wówczas sieć jest dysasortatywna.
- jeśli jest stała, sieć jest nieskorelowana.

Miary korelacji węzłów cd.

Wprowadźmy:

Średni stopień najbliższego węzła (dla węzła o stopniu k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

Pytamy o zależność średniego stopnia najbliższego sąsiada $\langle k \rangle_{nn}$ od stopnia wierzchołka k_i

Co mierzy ta miara?

- jeśli $\langle k \rangle_{nn}(k_i)$ jest funkcją rosnącą wówczas sieć jest asortatywna.
- jeśli $\langle k \rangle_{nn}(k_i)$ jest funkcją malejącą wówczas sieć jest dysasortatywna.
- jeśli jest stała, sieć jest nieskorelowana.
- a co jeśli zależność jest niemonotoniczna?

Miary korelacji węzłów cd.

W praktyce wszystkie poznane miary były zbyt złożony...

A zatem pozostaje nam liczyć współczynnik korelacji

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$$

co przy notacji

- e_{jk} – łączny rozkład pozostałych stopni.
- rozkład pozostałych stopni $q_k = \sum_j e_{jk}$, ale z drugiej strony $q_k = \frac{(k+1)p_{k+1}}{\sum_{j\geq 1} jp_j}$

prowadzi do

$$r = \frac{\frac{1}{M}\sum_i k_i j_i - \left[\frac{1}{2M}\sum_i (j_i + k_i)\right]^2}{\frac{1}{2M}\sum_i (j_i^2 + k_i^2) - \left[\frac{1}{2M}\sum_i (j_i + k_i)\right]^2},$$

gdzie $i = 1, 2, \ldots, M$ numeruje krawędzie, a j_i i k_i to stopnie wierzchołków przyłączonych do i-tej krawędzi.

Współczynnik gronowania

Zjawisko homofilii

Źródło: gazeta.pl

Równowaga Heidera

Model P-O-X Heidera w skrócie:

Równowaga Heidera

Model P-O-X Heidera w skrócie:

- Przyjaciel mojego przyjaciela jest moim przyjacielem.

Równowaga Heidera

Model P-O-X Heidera w skrócie:

- Przyjaciel mojego przyjaciela jest moim przyjacielem.
- Przyjaciel mojego wroga jest moim wrogiem.

Równowaga Heidera

Model P-O-X Heidera w skrócie:

- Przyjaciel mojego przyjaciela jest moim przyjacielem.
- Przyjaciel mojego wroga jest moim wrogiem.
- Wrog mojego przyjaciela jest moim wrogiem.

Równowaga Heidera

Model P-O-X Heidera w skrócie:

- Przyjaciel mojego przyjaciela jest moim przyjacielem.
- Przyjaciel mojego wroga jest moim wrogiem.
- Wrog mojego przyjaciela jest moim wrogiem.
- Wróg mojego wroga jest moim przyjacielem.

Równowaga Heidera

Model P-O-X Heidera w skrócie:

- Przyjaciel mojego przyjaciela jest moim przyjacielem.
- Przyjaciel mojego wroga jest moim wrogiem.
- Wrog mojego przyjaciela jest moim wrogiem.
- Wróg mojego wroga jest moim przyjacielem.

Dotyczy to jednak skierowanych sieci społecznych...

Uprośćmy nasze rozwazania do sieci nieskierowanych.

Współczynnik gronowania

Definicja

Współczynnik gronowania wierzchołka to stosunek liczby E_i istniejących krawędzi pomiędzy sąsiadami wierzchołka do wszystkich możliwych krawędzi pomiędzy tymi sasiadami

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

Współczynnik grafu to średnia po wszystkich wierzchołkach

$$C = \langle C_i \rangle.$$

Współczynnik gronowania

Definicja

Współczynnik gronowania wierzchołka to stosunek liczby E_i istniejących krawędzi pomiędzy sąsiadami wierzchołka do wszystkich możliwych krawędzi pomiędzy tymi sasiadami

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

Współczynnik grafu to średnia po wszystkich wierzchołkach

$$C = \langle C_i \rangle.$$

Alternatywna definicja współczynnika gronowania:

$$C_\triangle = \frac{3 \times \text{liczba trójkątów w sieci}}{\text{liczba dróg o długości 2 w sieci}}.$$

Motywy

Zliczamy *motywy* w sieci

przeważnie porównuje się *Z-score* z ansamblem sieci odkorelowanych

$$Z = \frac{p - \langle p \rangle}{\sigma}.$$

Jak zmierzyć jak mały jest świat sieci?

Średnia odległość

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j)$$

Wydajność

$$\mathcal{E} = \frac{1}{N(N-1)} \sum_{i \neq j} [d(i,j)]^{-1}.$$

Jak zmierzyć jak mały jest świat sieci?

Średnia odległość

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j)$$

Wydajność

$$\mathcal{E} = \frac{1}{N(N-1)} \sum_{i \neq j} [d(i,j)]^{-1}.$$

Pytanie.

Czym różnią się te dwie metryki? Która jest *lepsza*?

Pośrednictwo węzłowe

Który węzeł jest najważniejszy w sieci?

Poszukujemy najważniejszych *stacji przesiadkowych*.

Oznaczenia:

- δ_{jk} oznacza liczbę najkrótszych dróg łączących węzły j oraz k,
- $\delta_{jk}^{(i)}$ oznacza liczbę najkrótszych dróg łączących węzły j oraz k przechodzących przez węzeł i.

Definicja

$$B_i = \frac{2}{(N-1)(N-2)} \sum_k \sum_{j>k} \frac{\delta_{jk}^{(i)}}{\delta_{jk}}.$$

Pośrednictwo krawędziowe

A jeśli zapytamy o najważniejszą krawędź?

Poszukujemy najważniejszych *linii*.

Oznaczenia:

- $\delta_{jk}^{(e)}$ oznacza liczbę najkrótszych dróg łączących węzły j oraz k przechodzących przez krawędź e.

Definicja

$$B_i = \frac{2}{N(N-1)} \sum_k \sum_{j>k} \frac{\delta_{jk}^{(e)}}{\delta_{jk}}.$$

Czy potrzebujemy więcej metryk?

To jeden z głównych celów *sieciologii*

- cała sieć jest zbyt bogatą strukturą, żeby mówić o niej bez uproszczeń,

Czy potrzebujemy więcej metryk?

To jeden z głównych celów *sieciologii*

- cała sieć jest zbyt bogatą strukturą, żeby mówić o niej bez uproszczeń,
- często różne osoby interesują zupełnie inne cechy sieci,

Czy potrzebujemy więcej metryk?

To jeden z głównych celów *sieciologii*

- cała sieć jest zbyt bogatą strukturą, żeby mówić o niej bez uproszczeń,
- często różne osoby interesują zupełnie inne cechy sieci,
- często pewne szczególne miary potrzebne są do opisu pewnych szczególnych typów sieci...

Specyficzne metryki

- Indeks Hirscha – w sieci cytowań,

Specyficzne metryki

- Indeks Hirscha – w sieci cytowań,
- Liczba Erdősa – w sieci cytowań,

Specyficzne metryki

- Indeks Hirscha – w sieci cytowań,
- Liczba Erdősa – w sieci cytowań,
- Liczba Bacona – w sieci aktorów,

Specyficzne metryki

- Indeks Hirscha – w sieci cytowań,
- Liczba Erdősa – w sieci cytowań,
- Liczba Bacona – w sieci aktorów,
- PageRank – w sieci www,

Specyficzne metryki

- Indeks Hirscha – w sieci cytowań,
- Liczba Erdősa – w sieci cytowań,
- Liczba Bacona – w sieci aktorów,
- PageRank – w sieci www,
- próg epidemii – w epidemiologii,

Specyficzne metryki

- Indeks Hirscha – w sieci cytowań,
- Liczba Erdősa – w sieci cytowań,
- Liczba Bacona – w sieci aktorów,
- PageRank – w sieci www,
- próg epidemii – w epidemiologii,
- podatność na ataki/awarie – w inżynierii,

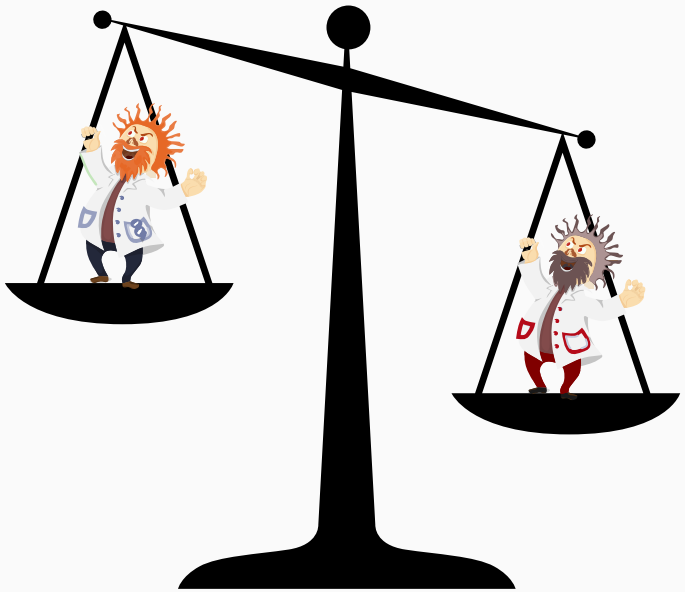
Specyficzne metryki

- Indeks Hirscha – w sieci cytowań,
- Liczba Erdősa – w sieci cytowań,
- Liczba Bacona – w sieci aktorów,
- PageRank – w sieci www,
- próg epidemii – w epidemiologii,
- podatność na ataki/awarie – w inżynierii,
- detekcja społeczności (klastrów),

Specyficzne metryki

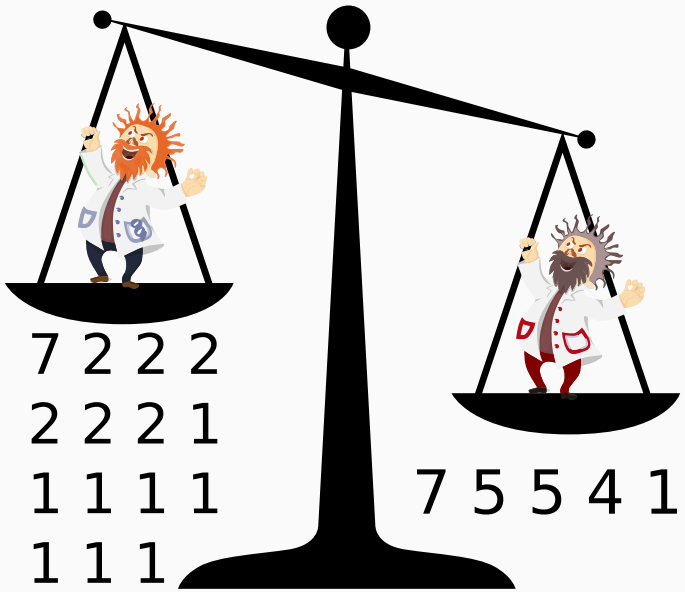
- Indeks Hirscha – w sieci cytowań,
- Liczba Erdősa – w sieci cytowań,
- Liczba Bacona – w sieci aktorów,
- PageRank – w sieci www,
- próg epidemii – w epidemiologii,
- podatność na ataki/awarie – w inżynierii,
- detekcja społeczności (klastrów),
- wiele, wiele innych...

Indeks Hirscha



Jak zmierzyć sukces naukowy?

Indeks Hirscha



Policzmy cytowania (stopnie wierzchołków).

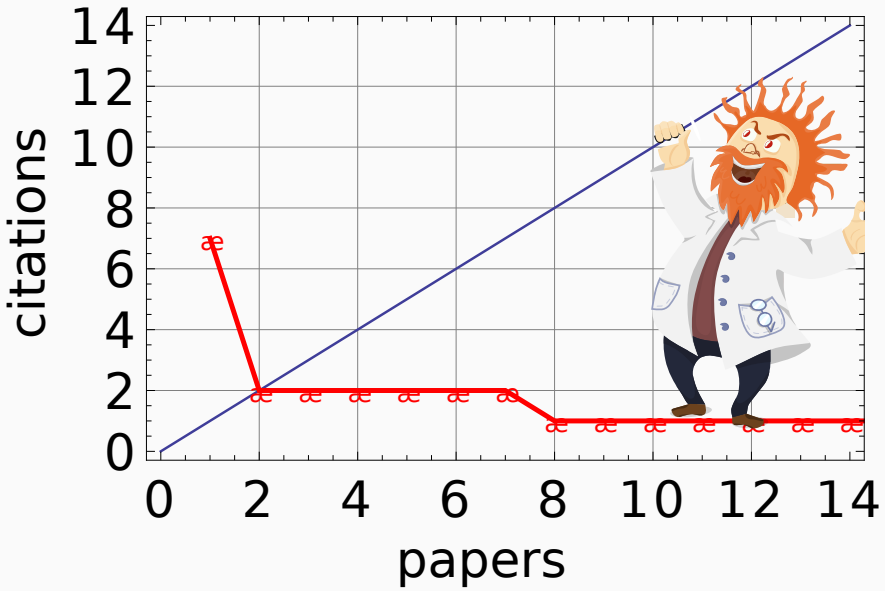
Indeks Hirscha

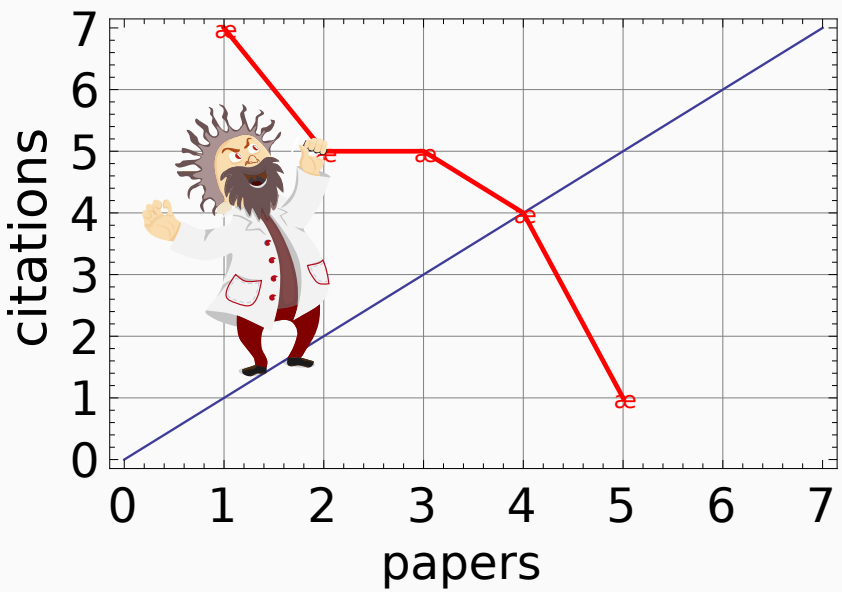
J.E. Hirsch, PNAS **102**, (2005).

$$h\text{-index} = \max\left\{h = 1, \ldots, n : \ X_{(n-h+1)} \geq h\right\},$$

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}.$$

Indeks Hirscha





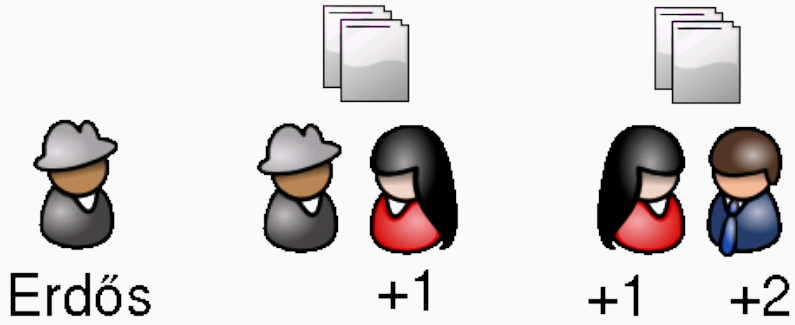
Liczba Erdősa

wikipedia

Paul Erdős 1913-1996

- węgierski matematyk
- na następnych zajęciach poznamy grafy Erdősa-Rényiego.

Liczba Erdősa

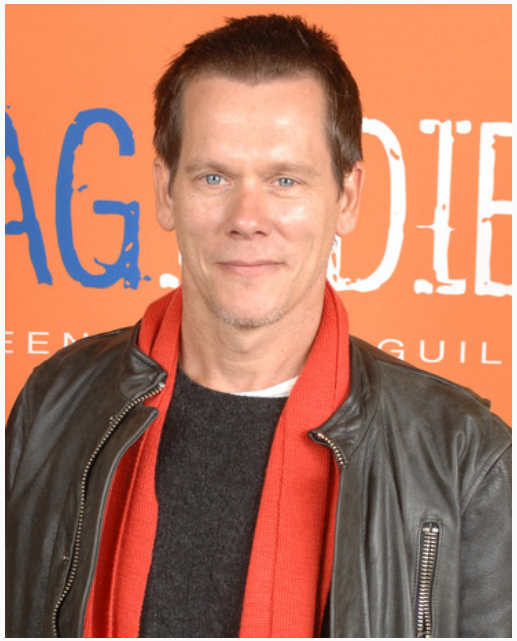


wikipedia

Definicja

- Paul Erdős ma liczbę Erdősa równą 0.
- Liczbę Erdősa każdego innego naukowca określa się jako minimum z $e_i + 1$ gdzie e_i to liczba Erdősa jego/jej współautorów.

Liczba Bacona



wikipedia

Kevin Bacon ur. 1958

- amerykański aktor, reżyser i producent filmowy,

Liczba Bacona

Odpowiednik liczby Erősa w sieci aktorów

Przykładowe wartości:

- Elvis Preasley: 2,
- Ronald Reagan: 2,
- Andrzej Grabowski: 3,
- Andrzej Lepper: 3,
- Zdzisław Maklakiewicz: 3,
- Jan Himilsbach: 3,

Liczba Erdősa-Bacona

Suma liczby Erdősa i liczby Bacona:

- Steven Strogatz $E = 3$ $B = 1 \Rightarrow EB = 4$,
- Richard Feynman $E = 3$ $B = 3 \Rightarrow EB = 6$,
- Stephen Hawking $E = 4$ $B = 2 \Rightarrow EB = 6$,
- Natalie Portman $E = 5$ $B = 2 \Rightarrow EB = 7$,
- Colin Firth $E = 6$ $B = 1 \Rightarrow EB = 7$,
- Kristen Stewart $E = 5$ $B = 2 \Rightarrow EB = 7$,
- Mayim Bialik $E = 5$ $B = 2 \Rightarrow EB = 7$.

Liczba Erdősa-Bacona

Suma liczby Erdősa i liczby Bacona:

- Steven Strogatz $E = 3$ $B = 1 \Rightarrow EB = 4$,
- Richard Feynman $E = 3$ $B = 3 \Rightarrow EB = 6$,
- Stephen Hawking $E = 4$ $B = 2 \Rightarrow EB = 6$,
- Natalie Portman $E = 5$ $B = 2 \Rightarrow EB = 7$,
- Colin Firth $E = 6$ $B = 1 \Rightarrow EB = 7$,
- Kristen Stewart $E = 5$ $B = 2 \Rightarrow EB = 7$,
- Mayim Bialik $E = 5$ $B = 2 \Rightarrow EB = 7$.

Ciekawostka

Proszę poczytać o liczbie Erdősa-Bacona-Black Sabath...

Pozostałe metryki

PageRank

- metoda wyboru najistotniejszych stron www,
- zajmiemy się nią na 11. zajęciach.

Pozostałe metryki

PageRank

- metoda wyboru najistotniejszych stron www,
- zajmiemy się nią na 11. zajęciach.

Próg epidemii

- minimalna liczba zarażonych osób w sieci społecznej, która skutkuje wybuchem epidemii,
- zajmiemy się nim na 12. zajęciach.

Pozostałe metryki

PageRank

- metoda wyboru najistotniejszych stron www,
- zajmiemy się nią na 11. zajęciach.

Próg epidemii

- minimalna liczba zarażonych osób w sieci społecznej, która skutkuje wybuchem epidemii,
- zajmiemy się nim na 12. zajęciach.

Odporność na przypadkowe uszkodzenia i celowe ataki

- zajmiemy się nimi na 7. zajęciach.

Pozostałe metryki

PageRank

- metoda wyboru najistotniejszych stron www,
- zajmiemy się nią na 11. zajęciach.

Próg epidemii

- minimalna liczba zarażonych osób w sieci społecznej, która skutkuje wybuchem epidemii,
- zajmiemy się nim na 12. zajęciach.

Odporność na przypadkowe uszkodzenia i celowe ataki

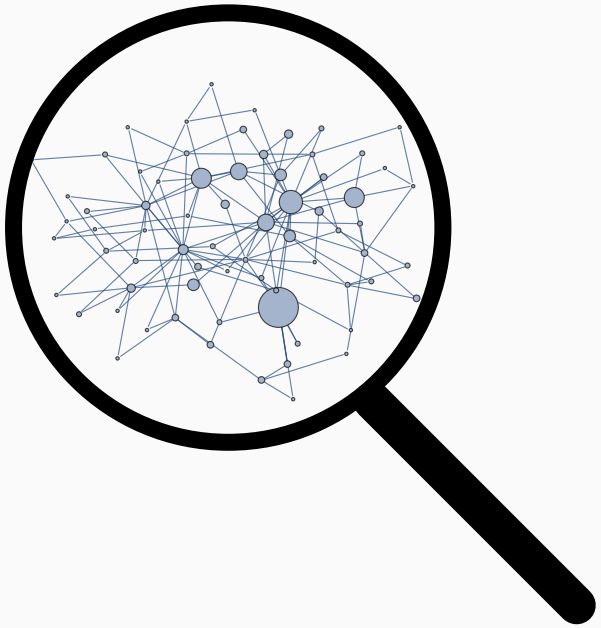
- zajmiemy się nimi na 7. zajęciach.

Detekcja społeczności (klastrów)

- zajmiemy się nią na 8. zajęciach.

Podsumowanie

Podsumowanie



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

Dziękuję za uwagę!