Metody analizy sieci społecznościowych

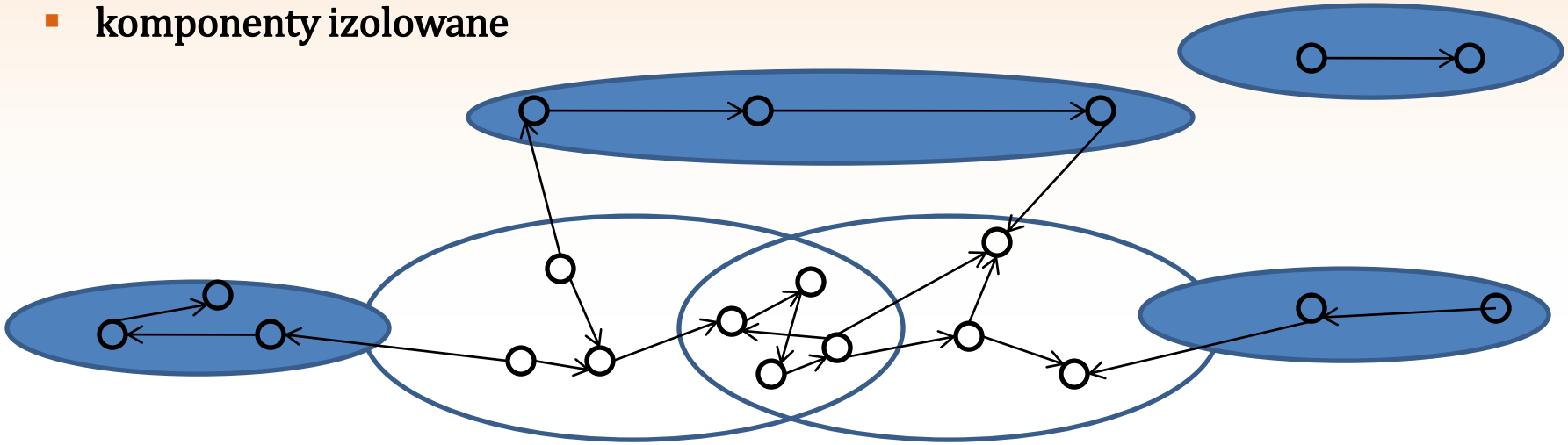
Część II

Mikołaj Morzy

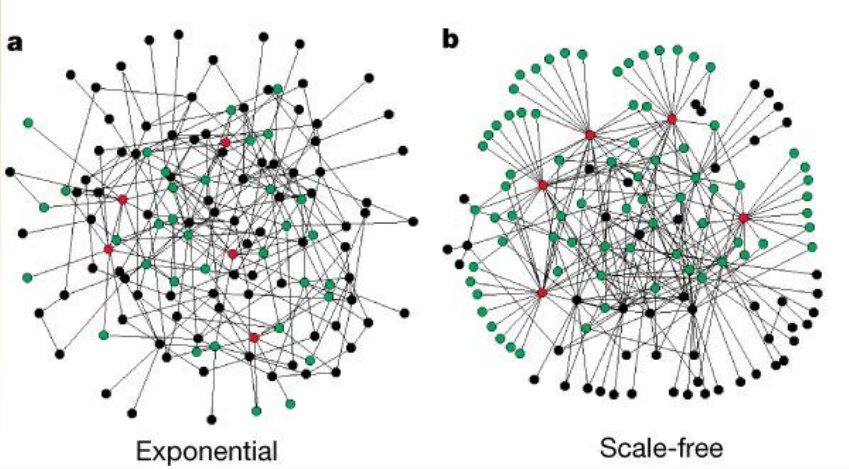
- współczynnik grupowania
- korelacje międzywęzłowe
- modularność sieci
- sieci bezskalowe
- matematyka praw potęgowych

Komponenty sieci

- W sieci wyróżniamy
  - **NSSK**: największy silnie spójny komponent
  - **komponent wejściowy**: wszystkie węzły z których można się dostać do NSSK
  - **komponent wyjściowy**: wszystkie węzły, do których można się dostać z NSSK
  - **wąsy**: węzły, do których ani z których nie można dotrzeć do NSSK
  - **komponenty izolowane**



Rozkład stopni wierzchołków



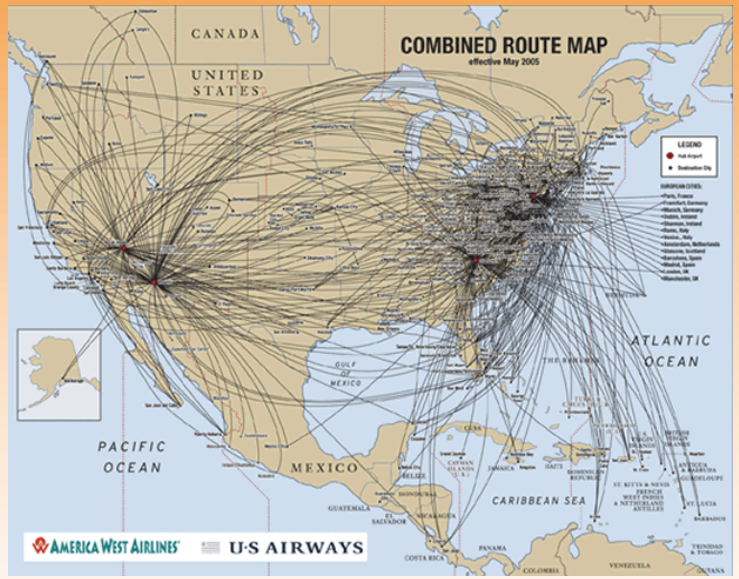
P(k) – prawdopodobieństwo, że wybrany węzeł sieci ma stopień *k*

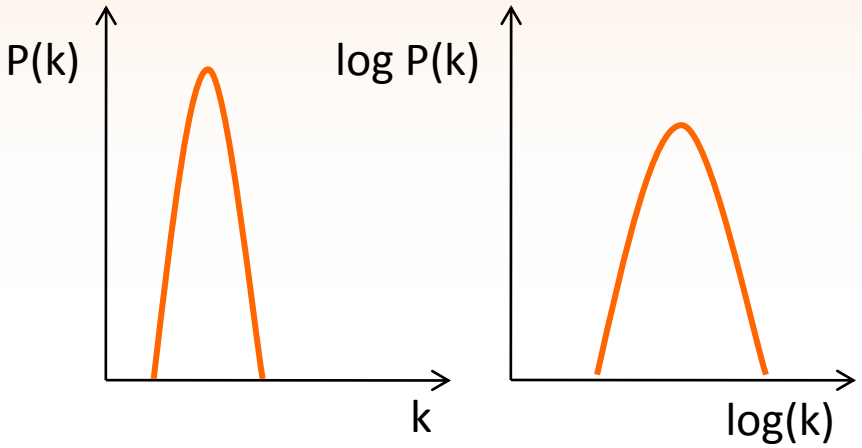
W sieciach społecznych najczęściej zachodzi zależność

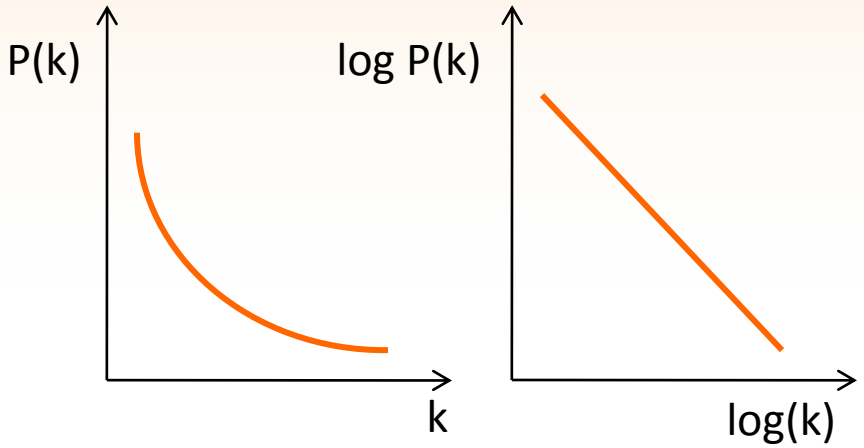
$$P(k) = \frac{C}{k^{\alpha}}$$

$$\ln P(k) = -\alpha \ln k + \ln C$$

Porównanie dwóch rozkładów



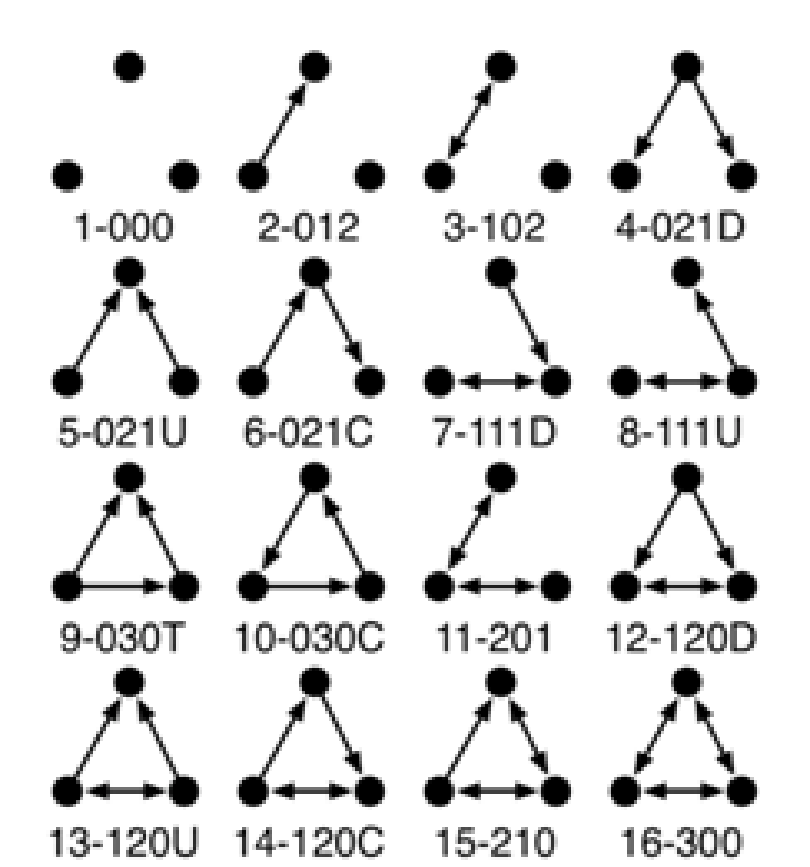




Statystyki rzeczywistych sieci

nazwa	→	N	E	<k>	α	l	r
aktorzy	nie	449 913	25 516 482	113.43	2.3	3.48	0.208
e-mail	tak	59 912	86 300	1.44	1.5	4.95	
www	tak	269 504	1 497 135	5.55	2.1	11.27	-0.06
blog.onet	tak	141 755		0.81	2.7	7.6	
kolej	nie	587	19 603	66.79		2.16	
pokarm	tak	135	598	4.43		2.05	-0.26
pakiety pr.	tak	1 439	1 723	1.20	1.6	2.42	-0.01
nauka	nie	52 909	245 300	9.27		6.19	0.36

Triady



Triady to prosty model mikro-społeczności występujących w sieci

Analiza triad pozwala na ilościową ocenę takich zjawisk jak:

- *homofilia*
- *układy sfrustrowane*

Współczynnik grupowania

Współczynnik grupowania (gronowania, klastrowania) wierzchołka (ang. *clustering coefficient*) to stosunek liczby istniejących krawędzi między sąsiadami wierzchołka do liczby wszystkich możliwych krawędzi

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

Współczynnik grupowania sieci jest uśrednioną wartością wszystkich współczynników grupowania wierzchołków

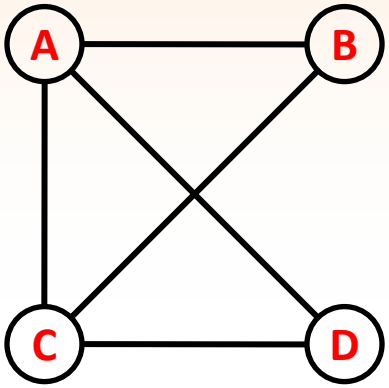
$$C = \langle C_i \rangle$$

Alternatywna definicja (socjologia)

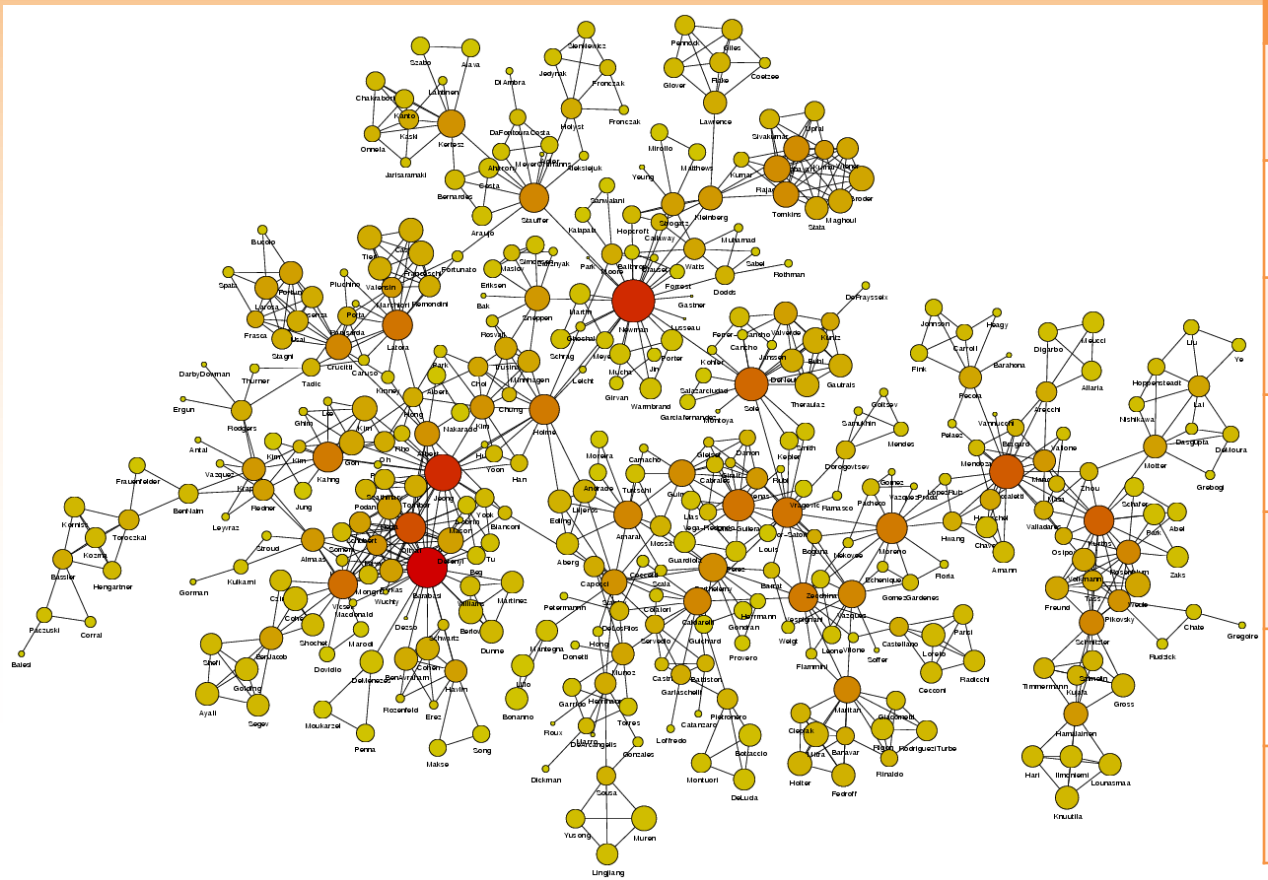
W socjologii czasem wykorzystuje się alternatywną definicję współczynnika grupowania

- N_Δ: liczba trójkątów w sieci
- L_2: liczba dróg o długości 2 w sieci

$$C_\Delta = \frac{3N_\Delta}{|L_2|}$$



Współczynnik grupowania w rzeczywistych sieciach



nazwa	C
domeny	0.18
aktorzy	0.79
nauka	0.49
E.Coli	0.59
synonimy	0.7
energetyka	0.08
WWW	0.1078

Współczynnik grupowania w sieciach ważonych

Dla grafów ważonych istnieje kilka definicji, jedną z najbardziej popularnych jest definicja Alaina Barrata

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,k} \frac{w_{ij} + w_{ik}}{2} a_{ij} a_{jk} a_{ik} = \frac{1}{k_i(k_i - 1)} \sum_{j,k} \frac{1}{\langle w_i \rangle} \frac{w_{ij} + w_{ik}}{2} a_{ij} a_{jk} a_{ik}$$

Wkład każdego trójkąta do C^w_i jest ważony przez stosunek średniej wagi dwóch krawędzi danego trójkąta przylegających do wierzchołka *i* do średniej wagi wszystkich krawędzi przylegających do tego wierzchołka

Średnia odległość w sieci

Średnia najkrótsza droga w sieci $l = \frac{1}{N(N-1)} \sum_{i \neq j} d(i, j)$

- w grafie losowym $l \propto \frac{\ln N}{\ln \langle k \rangle}$
- w sieci kwadratowej $l \propto \sqrt{N}$
- w sieciach bezskalowych

$$l \propto \ln N, \quad \alpha > 3$$

$$l \propto \frac{\ln N}{\ln \ln N}, \quad \alpha = 3$$

zjawisko ultra-małych światów → $l \propto \ln \ln N, \quad \alpha \in (2,3)$

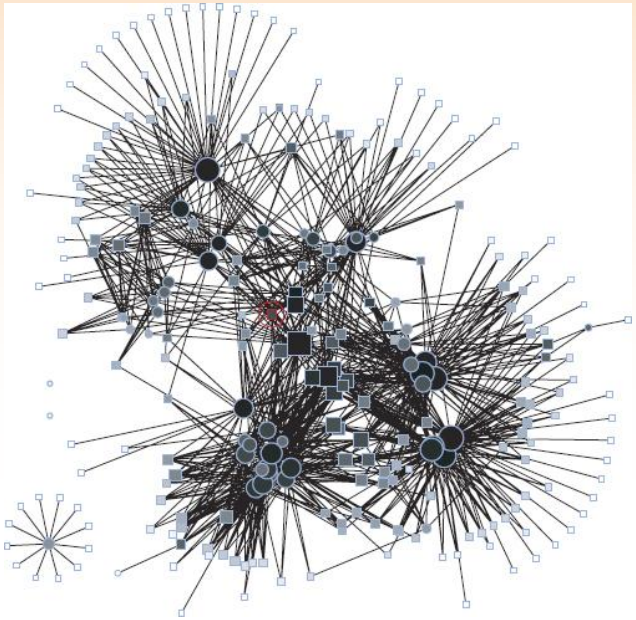
Wydajność sieci

Wykorzystywanie odległości międzywęzłowych jest kłopotliwe dla grafów o wielu składowych spójności. W takim przypadku wygodniej jest wykorzystać średnią harmoniczną odległości, zwaną wydajnością sieci

$$\mathrm{E} = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}^{-1}$$

Pośrednictwo

Pośrednictwo (ang. *betweenness*) wierzchołka to liczba najkrótszych dróg łączących dowolne dwa wierzchołki przechodzących przez dany wierzchołek



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

W sieciach bezskalowych rozkład pośrednictwa też jest potęgowy $P(B) \propto B^{-\eta}$

Korelacje dwuwęzłowe

- Czy węzły o dużych stopniach chętniej łączą się z innymi, dobrze usieciowionymi węzłami?
  - **tak**: towarzyskie związki między ludźmi
  - **nie**: sieć powiązań między serwerami

$$R(k_i,k_j)=\frac{P(k_i,k_j)}{P_u(k_i,k_j)}, \quad P(k_i,k_j)=\frac{N(k_i,k_j)}{|E|}$$

prawdopodobieństwo dla sieci nieskorelowanej o takim samym rozkładzie stopni wierzchołków

$$P_u(k_i,k_j)=\frac{k_i k_j P(k_i)P(k_j)}{\langle k\rangle^2}$$

Analiza korelacji dwuwęzłowych

Prawdopodobieństwo warunkowe

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

W zależności od rozkładu korelacji warunkowej można wyznaczyć dwa rodzaje sieci

- **dysasortatywne**: systemy autonomiczne
- **asortatywne**: sieci społeczne

Korelacja względem najbliższego sąsiada

W rzeczywistych sieciach występuje problem zbyt małej liczby wystąpień $N(k_i,k_j)$ do wyliczenia $P(k_i,k_j)$. Rozwiązaniem jest wprowadzenie średniego stopnia najbliższego sąsiada

$$\langle k \rangle_{nn} = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j$$

Można tę wartość wyznaczyć w funkcji stopnia węzła

$$\langle k \rangle_{nn}(k_i) = \sum_{k_j} k_j P(k_j \mid k_i)$$

Korelacja względem najbliższego sąsiada

Dla sieci nieskorelowanych, w których najbliższe sąsiedztwo każdego węzła jest jednakowe, otrzymujemy

$$\langle k \rangle_{nn}(k_i) = \frac{\langle k^2 \rangle}{\langle k \rangle}$$

Dla poszczególnych rodzajów sieci funkcja jest:

- sieci **nieskorelowane**: stała
- sieci **dysasocjacyjne**: malejąca
- sieci **asocjacyjne**: rosnąca

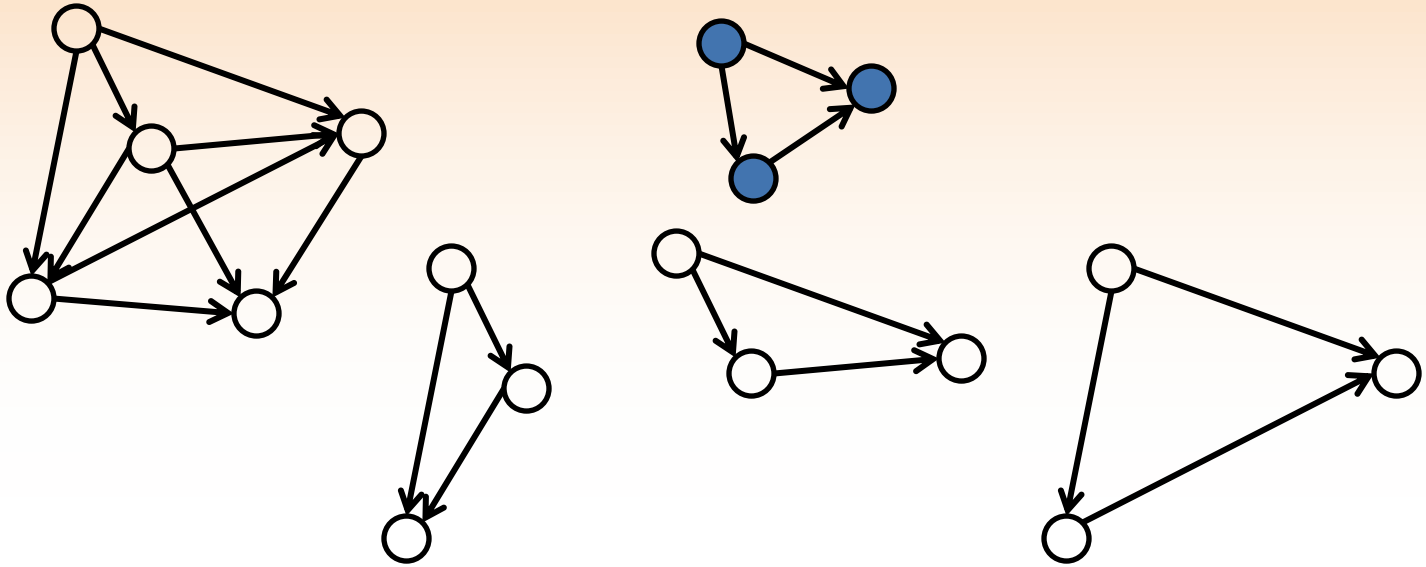
Korelacja Pearsona

Najprostszym sposobem pomiaru korelacji jest współczynnik Pearsona zdefiniowany dla sieci jako

$$r = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\langle k_i^2 \rangle - \langle k_i \rangle^2}$$

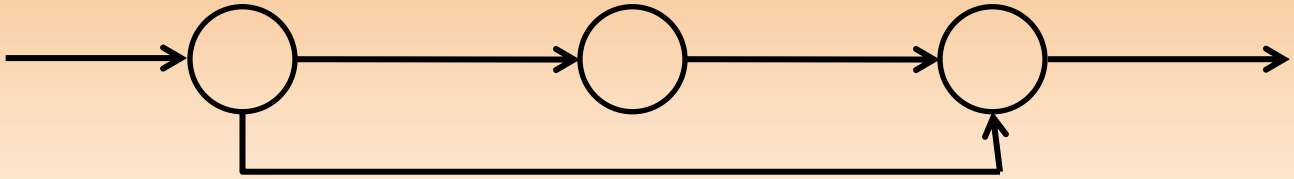
Motywy

Motyw to wzorzec połączeń w sieci, który występuje częściej niż wynikałoby to z założenia o losowości sieci

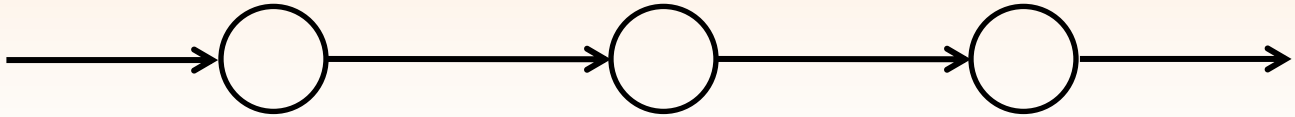


Popularne motywy w sieciach

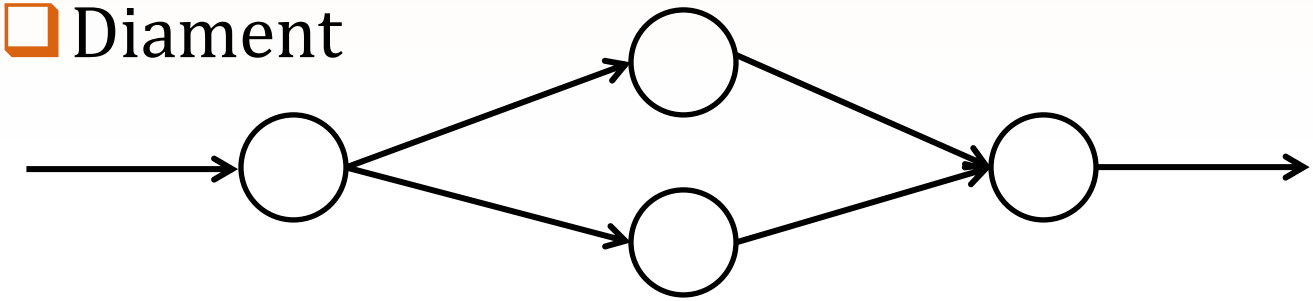
- Pętla ze sprzężeniem do przodu



- Prosty łańcuch



- Diament



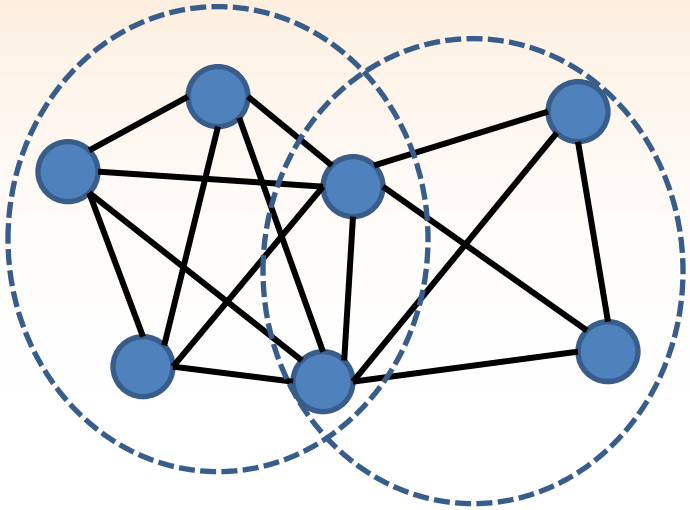
Definicja modułu

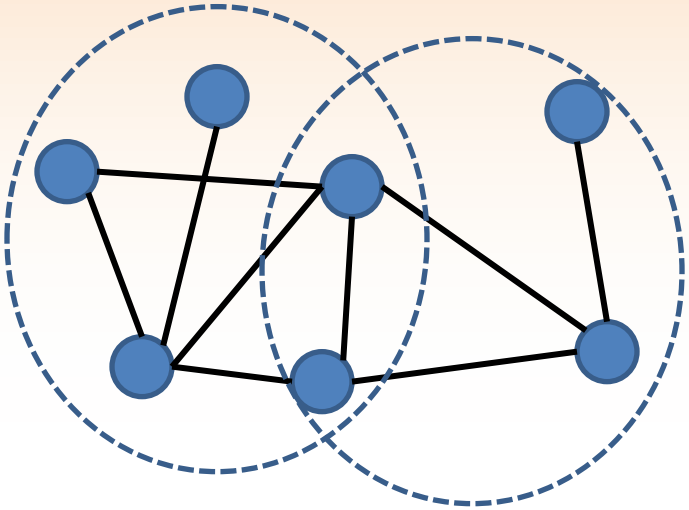
- W sieciach społecznych jednostki łączą się w wyraźne i odrębne społeczności lokalne

- Modularność dotyczy także innych sieci
  - fizyczne sieci komputerowe
  - sieci połączeń drogowych
  - sieci neuronowe

Kliki i n-kliki

- klika: zbiór wierzchołków, którego każda para wierzchołków jest połączona krawędzią
- n-klika: zbiór wierzchołków, którego każda para wierzchołków jest od siebie oddalona co najwyżej o n krawędzi





Nie działa, bo $E \infty N$, a powinno być $E \infty N^2$

Moduł

- Moduł to zbiór wierzchołków taki, że gęstość połączeń wewnątrz modułu jest większa niż gęstość połączeń między wierzchołkami należącymi do różnych modułów
  - Znajdowanie modułów jest wykonywane przez algorytmy analizy hierarchii skupień
  - Liczba modułów występujących w sieci najczęściej nie jest znana *a priori*

Szkic metody znajdowania modułów

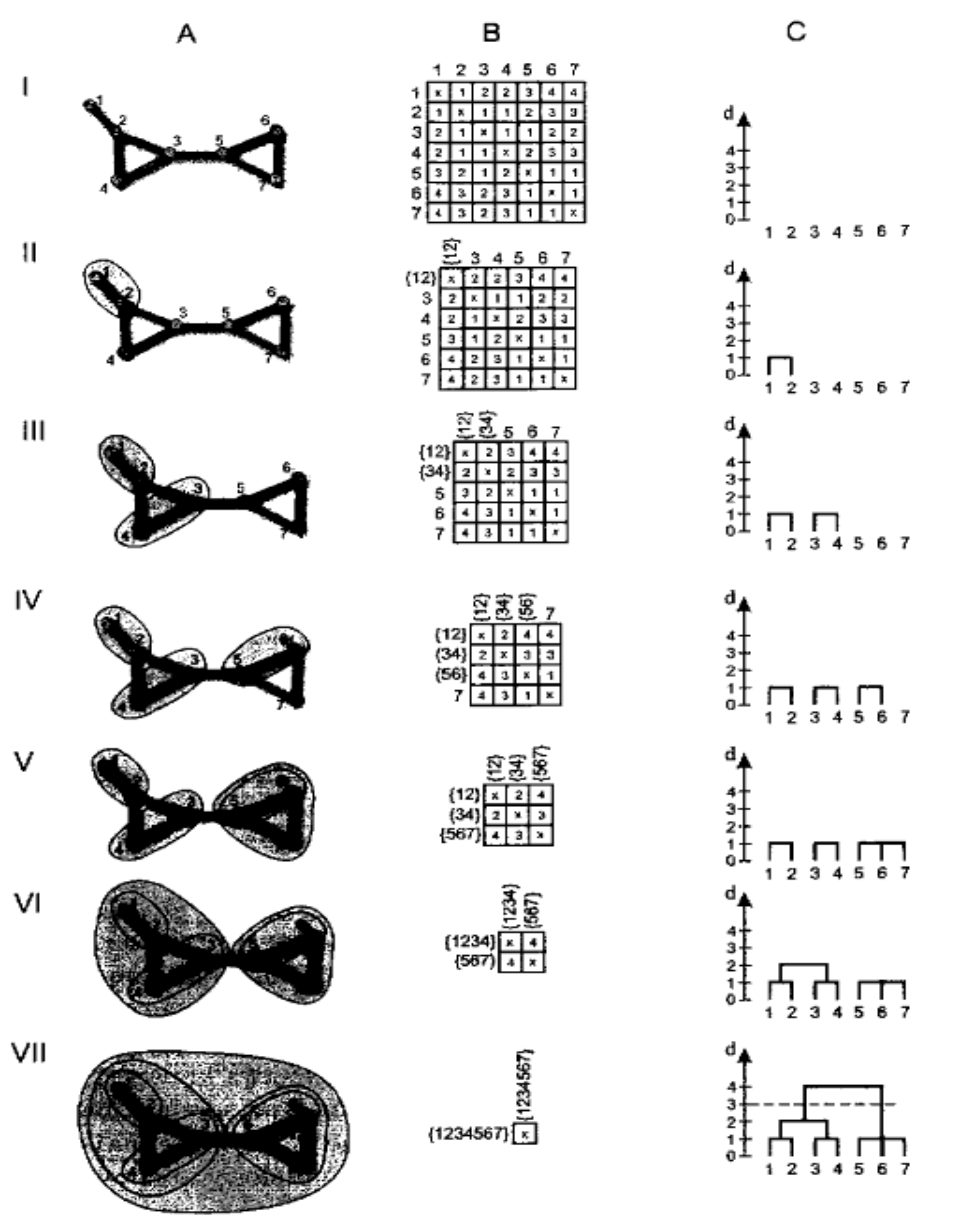
- Macierz odległości D o rozmiarze N×N
- Odległość strukturalna: węzły są strukturalnie równoważne, jeśli mają tych samych sąsiadów

$$D_{ij} = \sqrt{\sum_{k=1}^{N} (a_{ik} - a_{jk})^2} \quad k \neq i, j$$

1. Znajdź parę najbliższych elementów w sieci (wierzchołków lub modułów)
2. Oblicz nową macierz odległości *D*
3. Powtarzaj kroki 1-2 aż uzyskasz jeden moduł

Przykład analizy hierarchii skupień

metoda aglomeracyjna

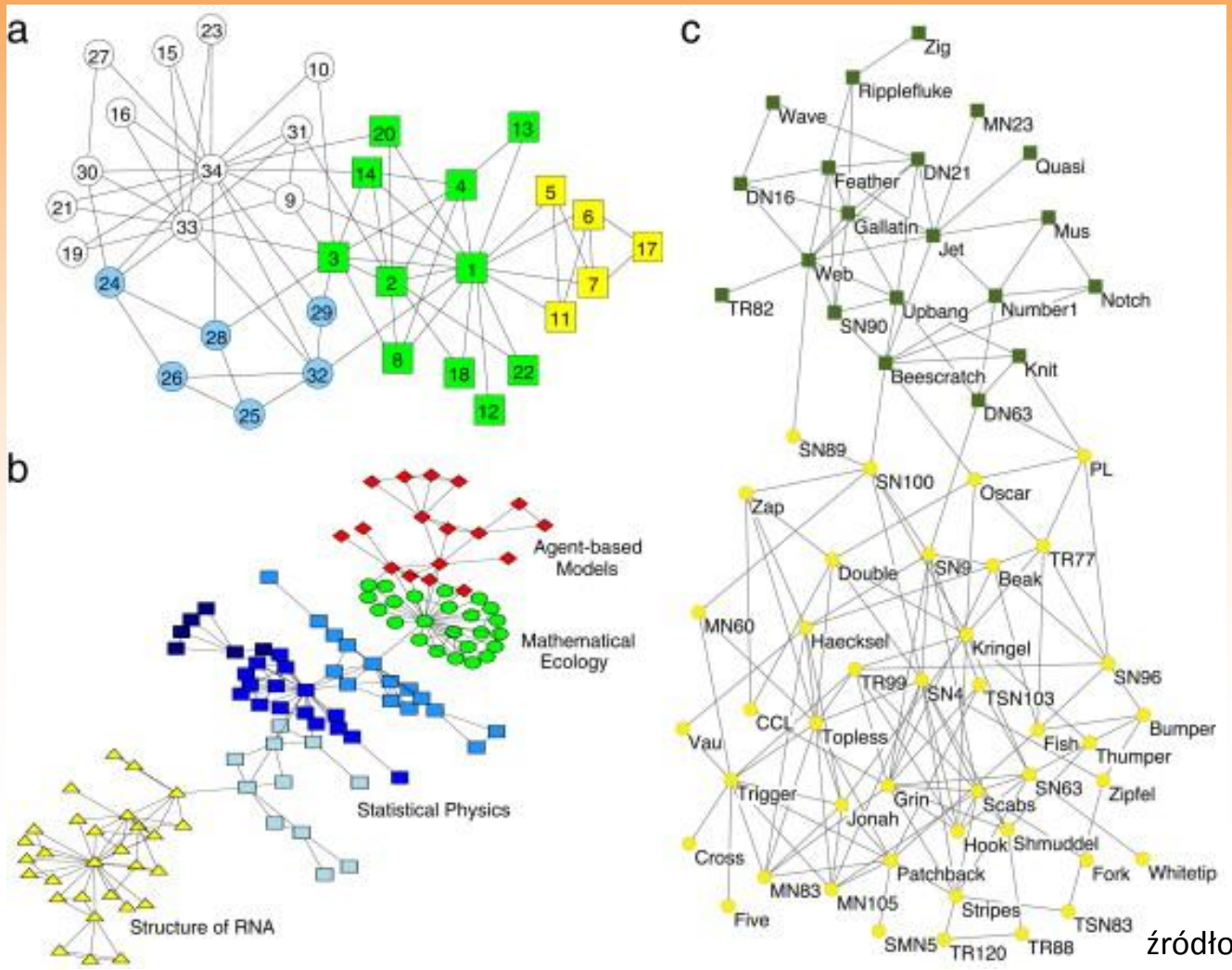


Metody podziałowe analizy hierarchii skupień

- W przeciwieństwie do metod aglomeracyjnych, metody podziałowe rozpoczynają od jednego wielkiego modułu, dzieląc go mniejsze moduły

- Kryterium podziału: wysokie pośrednictwo krawędzi
  - wysokie pośrednictwo charakteryzuje krawędzi, które łączą ze sobą wierzchołki należące do różnych modułów
  - algorytm Girvan-Newmana

Przykład metody podziałowej



źródło: www.sciencedirect.com

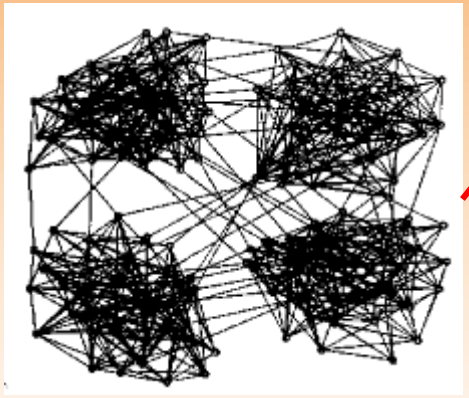
Skąd wiadomo, ile powinno być modułów?

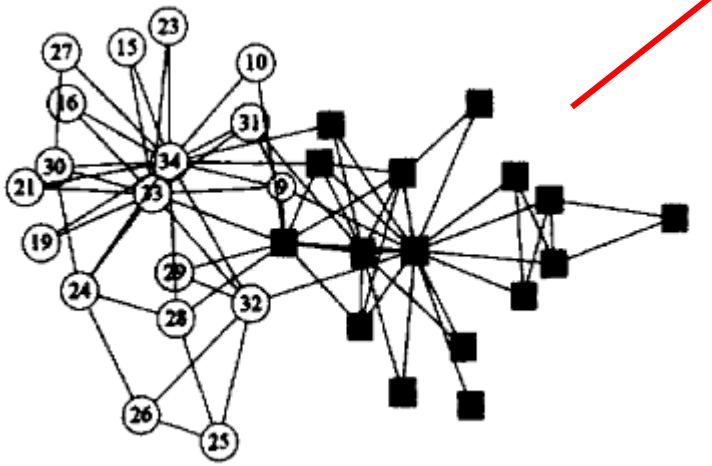
- Modularność: różnica między rzeczywistą liczbą krawędzi wewnątrz modułu a liczbą wynikającą z przypadkowego podziału na moduły

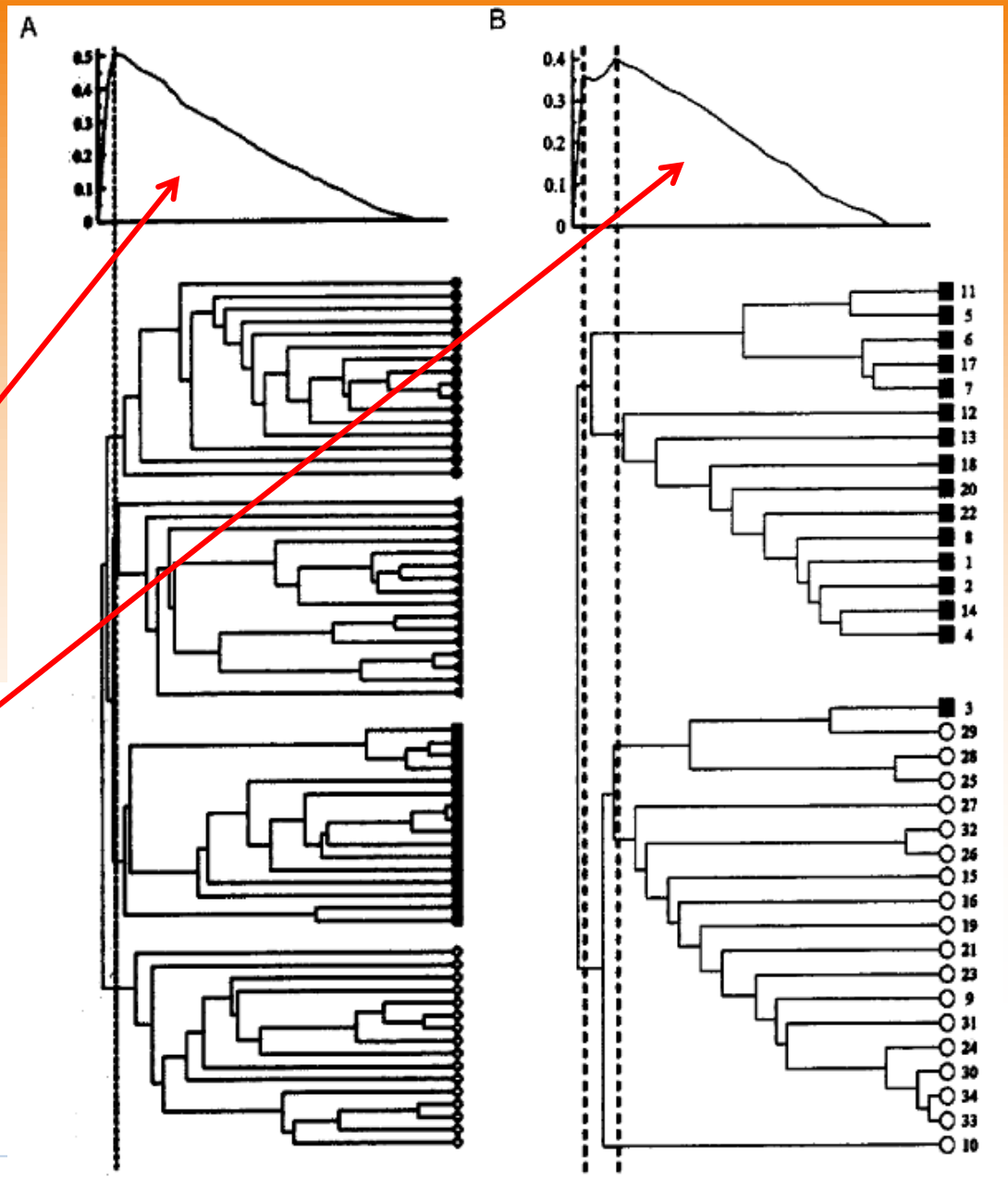
$$Q = \sum_i (e_{ii} - p^2{}_{\to i})$$

- macierz e o rozmiarze $k \times k$ (k oznacza liczbę modułów)
- e_{ij}: ułamek krawędzi z modułu i do modułu j
- ślad macierzy Tr $e = \sum_i e_{ii}$
- suma elementów w wierszu (kolumnie) $p_{\to i} = \sum_j e_{ij}$
- prawdopodobieństwo krawędzi w module i w grafie przypadkowym $p^2{}_{\to i} = p_{\to i} * p_{\to i}$

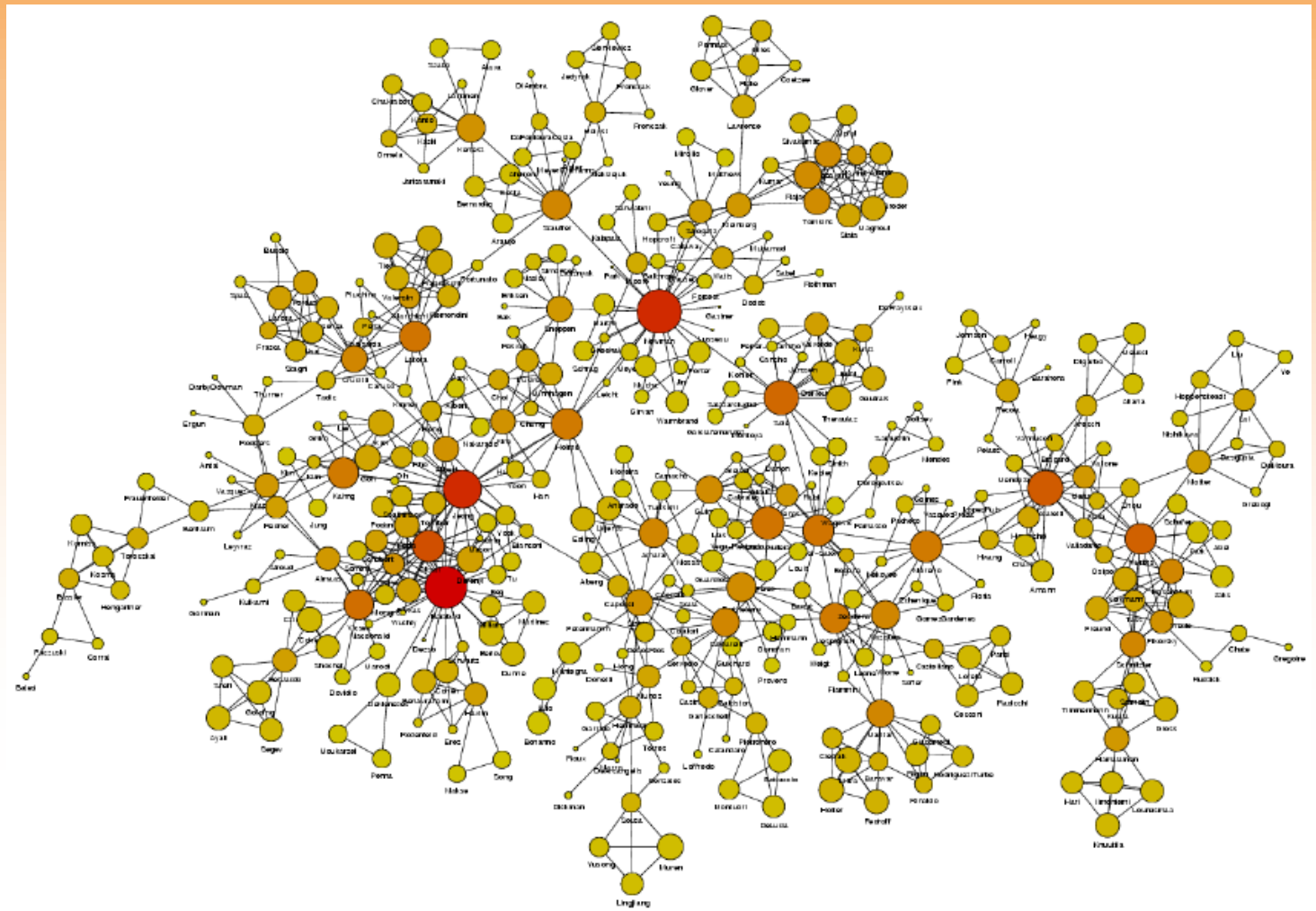
Przykład wykorzystania modularności



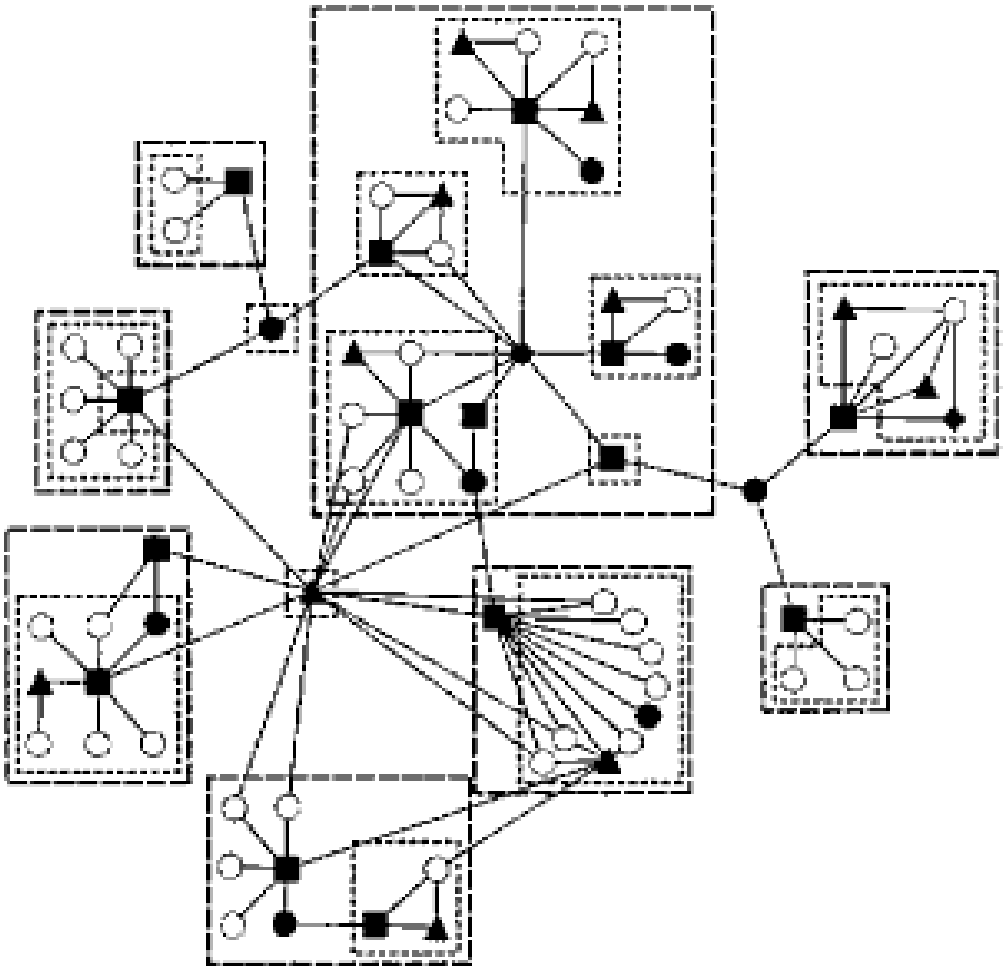




Hierarchiczność w sieciach (1/2)



Hierarchiczność w sieciach (2/2)



Lokalny współczynnik grupowania

$$C_i(k) \propto k^{-1}$$

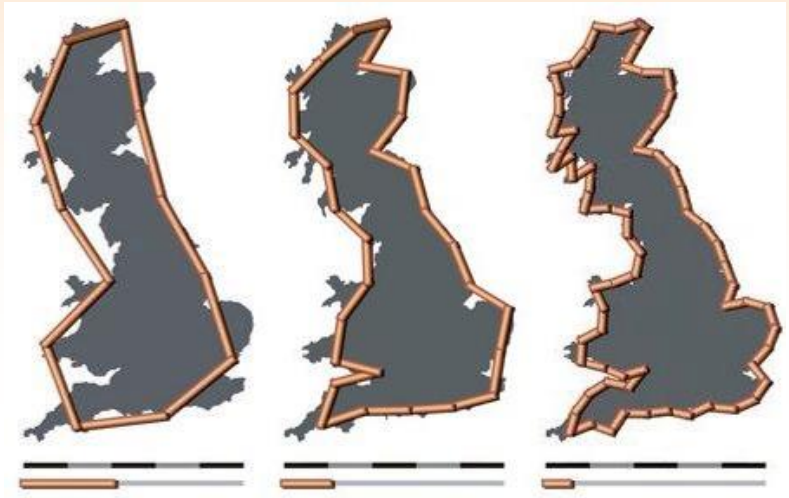
- trawy
- pasożyty
- roślinożercy
- hiperpasożyty
- pasożyty III rzędu

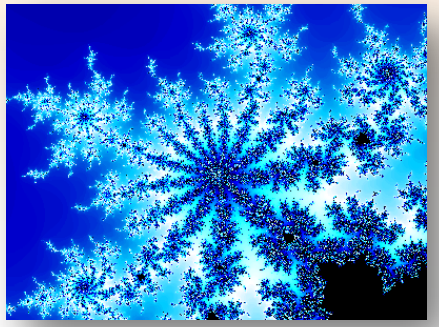
Układy złożone

- Układ złożony: dowolny układ, który
  - jest bardzo wrażliwy na warunki początkowe
  - jest bardzo wrażliwy na małe zakłócenia
  - zawiera wiele interakcji między komponentami, a właściwości układu nie da się przewidzieć w prosty sposób badając indywidualne komponenty
  - nieustannie ewoluuje i rozwija się

Co to jest bezskalowość?

- Istnieją obiekty i zjawiska których nie można zmierzyć
- Niemierzalność jest najczęściej związana z samopodobieństwem struktur
- Samopodobieństwo jest ściśle związane z prawami potęgowymi





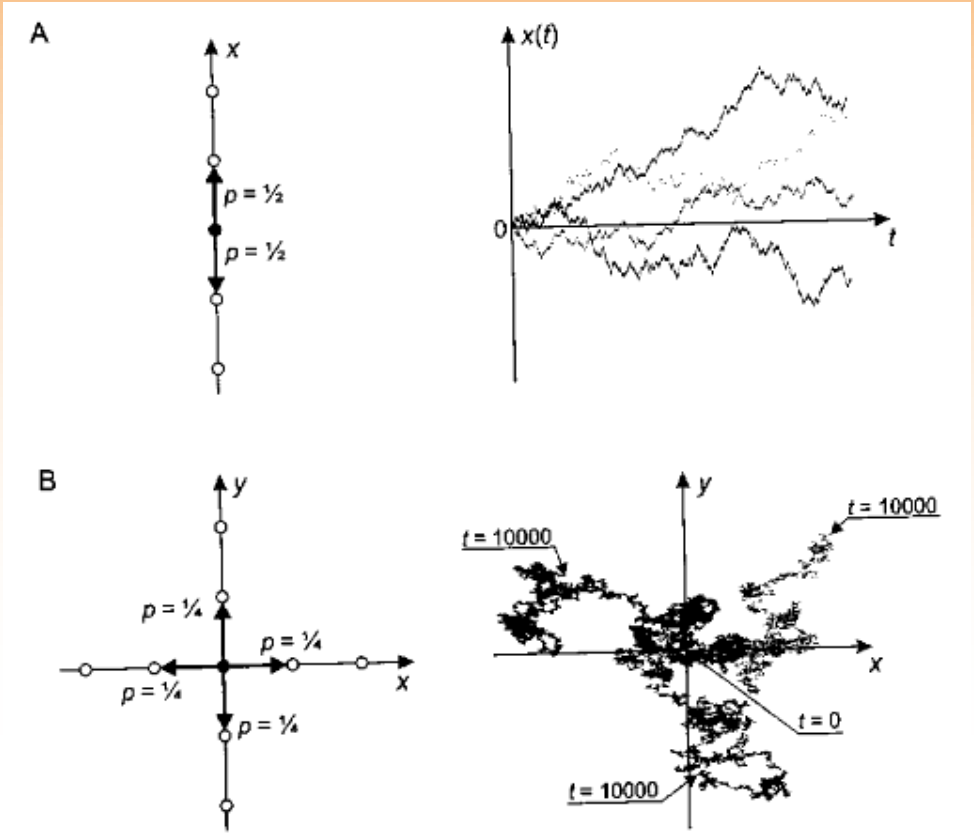
Ogólna postać praw potęgowych

- Istnieją układy i zjawiska, które nie posiadają cechy samopodobieństwa, a mimo to nie posiadają naturalnej skali
- Zjawiska takie można najczęściej opisać za pomocą potęgowych rozkładów prawdopodobieństwa

  - liczba hiperlinków na stronie WWW
  - liczba e-maili wysłanych przez użytkownika
  - liczba cytować publikacji naukowych
  - liczba wystąpień nazwisk
  - liczebność gmin w Polsce

$$P(x) \propto x^{-\alpha}$$

Problem błądzącej mrówki



prawdopodobieństwo powrotu w chwili τ

$$F(\tau) \propto \frac{1}{\tau^{3/2}}$$

średni czas oczekiwania na powrót

$$\bar{\tau} = \int_0^\infty \tau F(\tau) d\tau \propto \int_0^\infty \frac{1}{\sqrt{\tau}} d\tau = \infty$$

średnie położenie po czasie t

$$\mu_t = \left\langle \sum_{i=1}^{t} x_i \right\rangle = t\langle x_i \rangle = 0$$

błąd oszacowania położenia

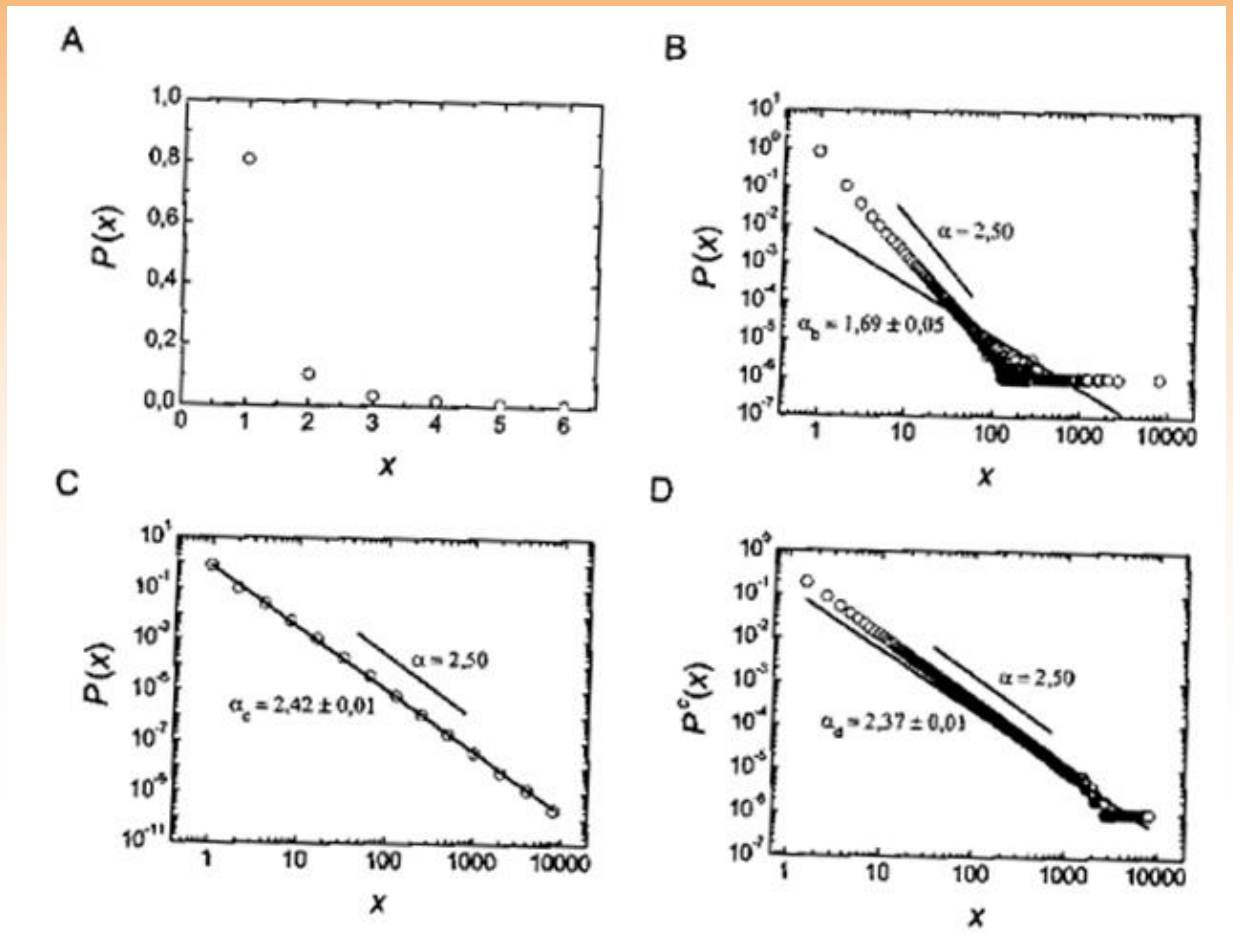
$$\sigma_t = \sqrt{\left\langle \sum_{i=1}^{t} (x_i - \langle x_i \rangle)^2 \right\rangle} = \sqrt{t\langle x_i^{\,2} \rangle} = \sqrt{t}$$

Jak sprawdzić, czy rozkład jest potęgowy?

- Czy wystarczy proste zliczenie N(x) i wyznaczenie zależności funkcyjnej?
  - w granicy N>>1 jest to rozkład prawdopodobieństwa $P(x) = \frac{N(x)}{\sum_x N(x)}$
  - wiele układów jest opisanych prawami potęgowymi w granicach dużych i rzadkich zdarzeń
  - konieczny jest reprezentatywny zbiór pomiarów
  - wielkość N wpływa na rozmiary zdarzeń

$$\langle x_{\max} \rangle \propto N^{\frac{1}{\alpha-1}}$$

Przykład analizy



A: oryginalne dane
B: skala log-log
C: uśrednianie w przedziałach
D: uśrednianie logarytmiczne

Dyskretyzacja rozkładów potęgowych

Uśrednianie w przedziałach

- podział zmiennych na przedziały
- zliczenie w przedziałach

$$\Delta x_n = \langle zn, z(n+1) \rangle, \quad |\Delta x_n| = z$$

$$\langle N_n \rangle = \frac{\sum_{\Delta x_n} N(x)}{|\Delta x_n|}$$

Dyskretyzacja logarytmiczna

- podział zmiennych na przedziały
- zliczenie w przedziałach

$$\Delta x_n = \langle z^n, z^{n+1} \rangle, \quad |\Delta x_n| = z^n(z-1)$$

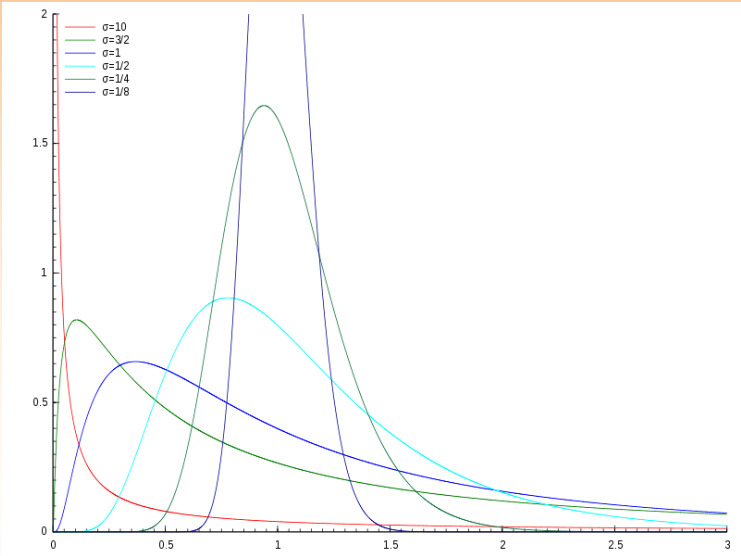
$$\langle N_n \rangle = \frac{\sum_{\Delta x_n} N(x)}{|\Delta x_n|}$$

Rozkłady log-normalne

- Procesy i zjawiska opisane rozkładem log-normalnym

$$P(x) \propto \frac{1}{x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

$$\ln P(x) = \frac{1}{2\sigma^2} \ln(x)^2 + \left(\frac{\mu}{\sigma^2} - 1\right) \ln x + C$$



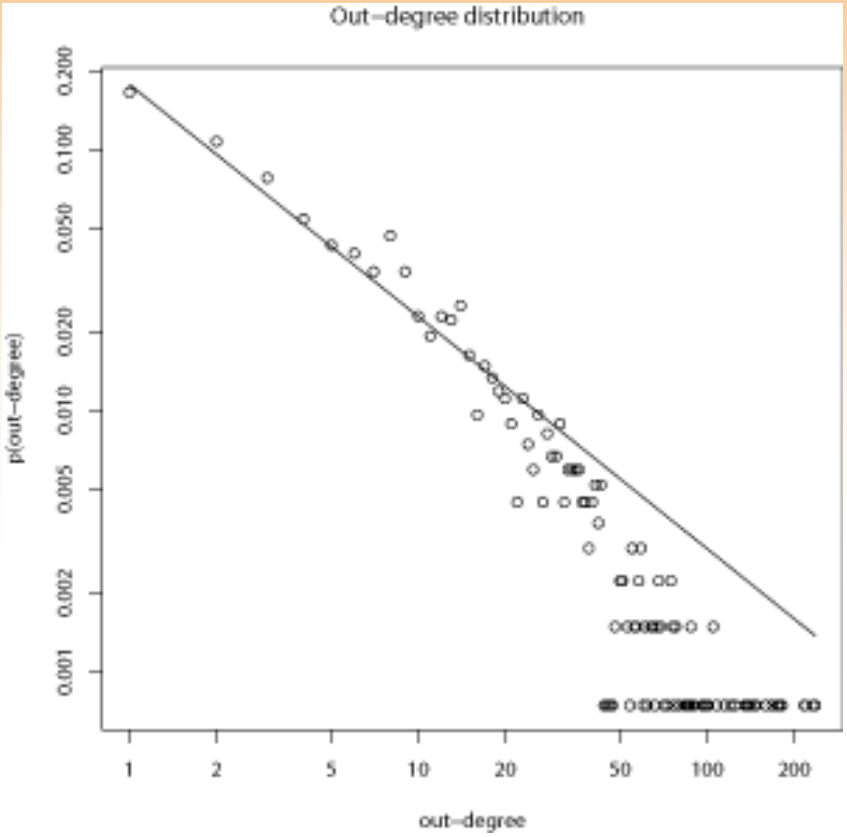
Rozkłady log-normalne najczęściej pojawiają się w wyniku multiplikatywnych procesów stochastycznych, gdzie $x_n = x_0 \Pi y_i$. Przykłady takich procesów to np. zmiana liczności populacji lub zachowania inwestorów giełdowych

Rozkłady potęgowe z obcięciem wykładniczym

- Procesy i zjawiska opisane rozkładem potęgowym z obcięciem wykładniczym

$$P(x) \propto \frac{1}{x^{\alpha}} e^{-\frac{x}{x_0}}$$

  - rozkład jest potęgowy dla $x << x_0$, ponieważ wykładnik jest bliski 1



Rozkłady potęgowe z obcięciem wykładniczym

- W jaki sposób parametr obcięcia x_0 jest związany z zagadnieniem niemierzalności układu?
  - jeśli x_0 jest stałe (nie zależy od wielkości układu) to wszystkie momenty (średnia, wariancja, odchylenie standardowe) są skończone
  - jeśli x_0 zależy od rozmiaru układu, np. $x_0(N) \propto N^{\mu}$ to pojawia się bezskalowość
    - każdy układ z osobna jest mierzalny
    - parametr skali każdego układu jest inny
    - nie istnieje wspólna skala dla całego układu

Wartość oczekiwana, odchylenie standardowe i wariancja rozkładu potęgowego

- Wartość oczekiwana
  - rozbieżna dla $\alpha \leq 2$
  - nieskończona wartość oczekiwana

$$\bar{x} = \int_{x_{\min}}^{\infty} xP(x)dx = C\int_{x_{\min}}^{\infty} x^{-\alpha+1}dx$$

- Drugi moment rozkładu
  - nieskończony dla $\alpha < 3$
  - wariancja jest nieskończona nawet dla $\alpha > 2$ i

$$\overline{x^2} = \int_{x_{\min}}^{\infty} x^2 P(x)dx = C\int_{x_{\min}}^{\infty} x^{-\alpha+2}dx$$

$$\sigma^2 = \left\langle x^2 \right\rangle - \left\langle x \right\rangle^2$$

$$\left\langle x \right\rangle = \frac{C}{2-\alpha} x_{\min}{}^{-\alpha+2} = \frac{\alpha-1}{\alpha-2} x_{\min}$$

- Ogólnie, momenty $\langle x^m \rangle$ są rozbieżne dla $m \geq \alpha - 1$

Zdarzenia ekstremalne

- Prawdopodobieństwo zajścia największego zdarzenia

$$P(x_{\max}) = \frac{1}{N}$$

- Prawdopodobieństwo, że w badanym zbiorze jakiekolwiek zdarzenie będzie większe niż $x_{\max}$

$$\int_{x_{\max}}^{\infty} P(x)dx = P^C(x_{\max}) = \frac{1}{N}$$

- Po podstawieniu skumulowanego rozkładu prawdopodobieństwa

$$P^C(x) = \int_{x}^{\infty} P(x)dx = \left(\frac{x}{x_{\min}}\right)^{-\alpha+1}$$

$$x_{\max} = x_{\min} N^{\frac{1}{\alpha-1}} \qquad \langle x_{\max} \rangle \propto N^{\frac{1}{\alpha-1}}$$

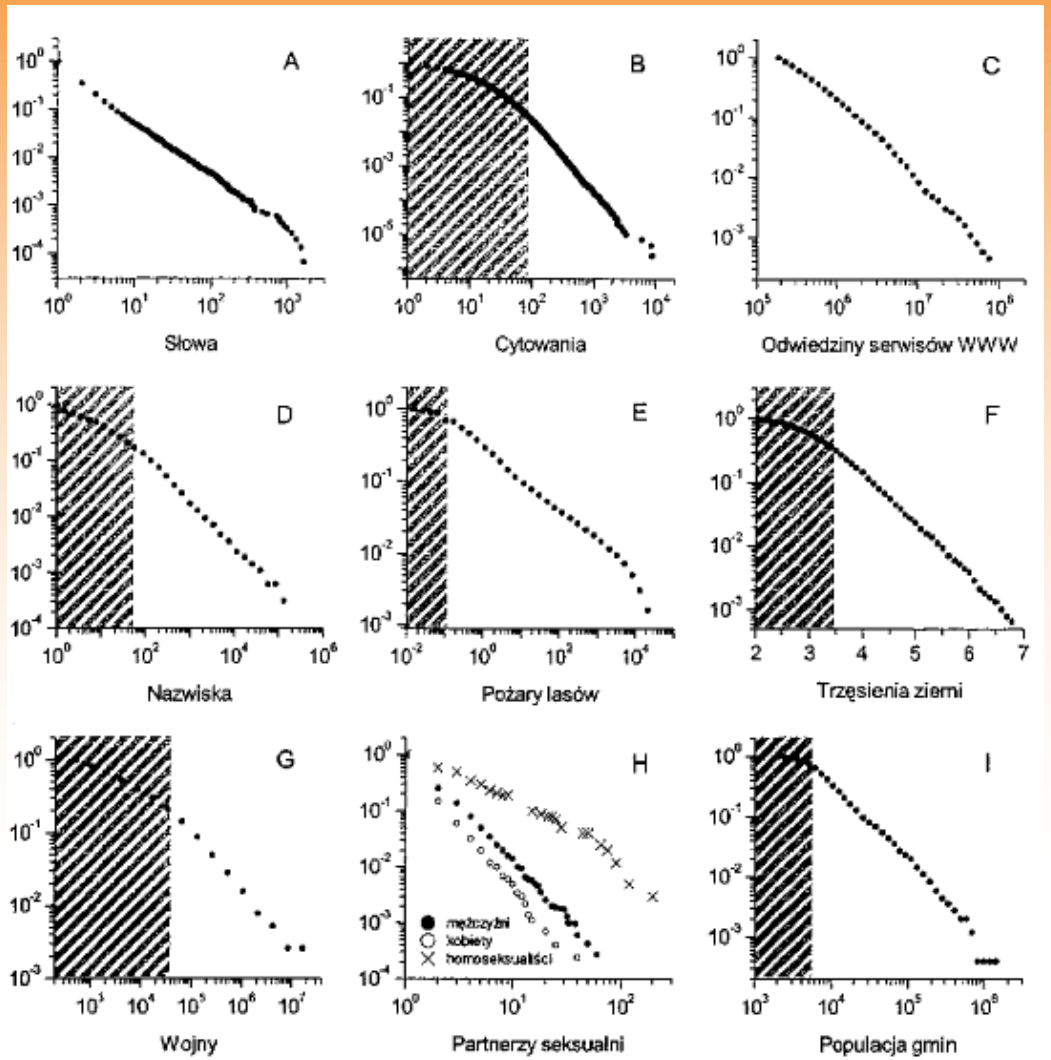
Przykład wyznaczania wartości ekstremalnych

- Warszawa liczy sobie $N=1.7*10^6$ mieszkańców
- Znajomości między ludźmi tworzą sieć społeczną rządzoną prawem potęgowym o wykładniku charakterystycznym $\alpha=2.5$
- Stopień najlepiej usieciowionego człowieka w wynosi

$$x_{\max} = x_{\min} N^{\frac{1}{\alpha-1}} = 1 \cdot \left(1.7 \cdot 10^6\right)^{\frac{1}{2.5-1}} = 1700000^{0.66} \approx 13000$$

  - w sieci losowej wynosiłby $x_{\max} \approx 3$

Rzeczywiste zjawiska i prawa potęgowe



A: słowa
B: cytowania
C: odwiedziny stron
D: nazwiska
E: pożary lasów
F: trzęsienia ziemi
G: wojny
H: partnerzy seksualni
I: populacja gmin

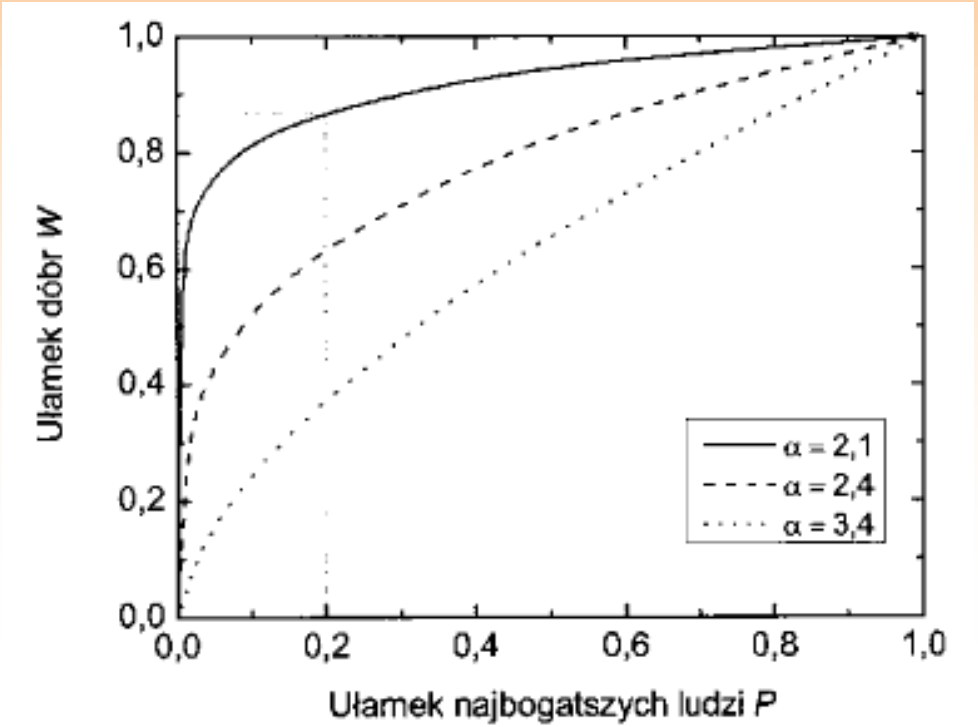
Rozkłady z tłustym ogonem

- Zjawiska i układy spełniające zależność

$$\lim_{x\to\infty} e^{\lambda x} P^{C}(x) = \infty$$

skumulowany rozkład wykładniczy

badany rozkład



- reguła Pareto
- reguła 80/20