

Justyna Signerska

Grafy losowe jako modele sieci



Główne metody konstruowania sieci:

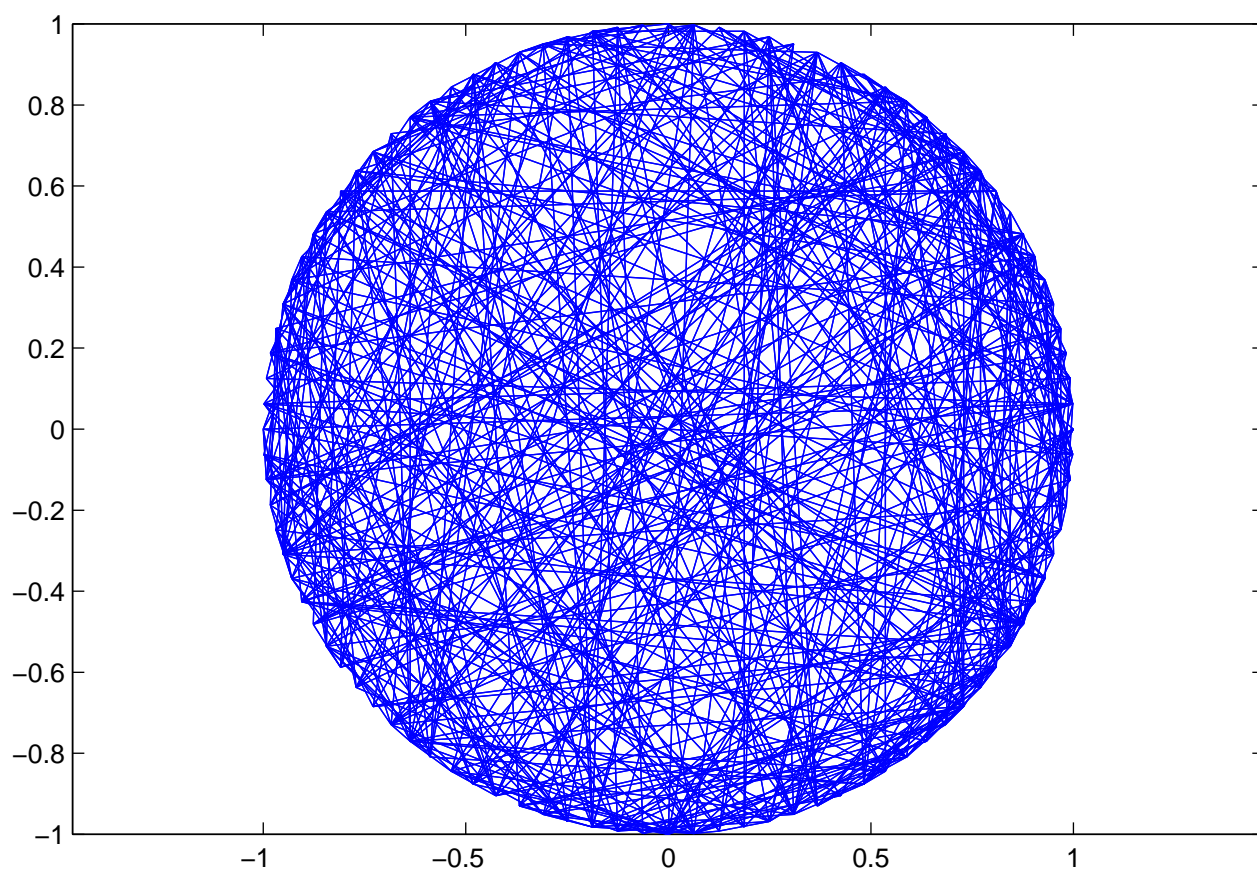
- klasyczny graf losowy $G_{n,p}$ (Erdős, Rényi - 1960)
- graf losowy z ustalonym rozkładem stopni wierzchołków (\sim 1972)-
tzw. model konfiguracyjny
- sieć "małych światów" (*Small-world networks*, Watts i Strogatz - 1998)
- sieć Barabasi-Albert (1999)

$G_{n,p}$:

średni stopień wierzchołka

$$z = \frac{n(n-1)p}{n} = (n-1)p \approx np \quad (1)$$

(ostatnie przybliżenie właściwe dla dostatecznie dużych n)



$G_{n,p}$ a rzeczywiste sieci

Niech p_k oznacza prawdopodobieństwo, że losowo wybrany wierzchołek ma stopień k

1. klasyczny graf losowy posiada dwumianowy rozkład stopni wierzchołków

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}, \quad (2)$$

który w granicy $n \gg k$ przechodzi w rozkład Poissona:

$$p_k = \frac{z^k e^{-z}}{k!} \quad (3)$$

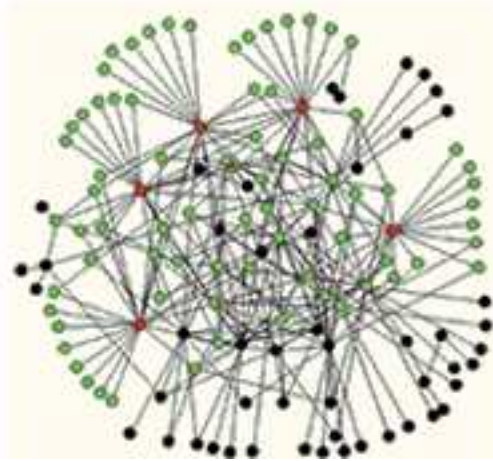
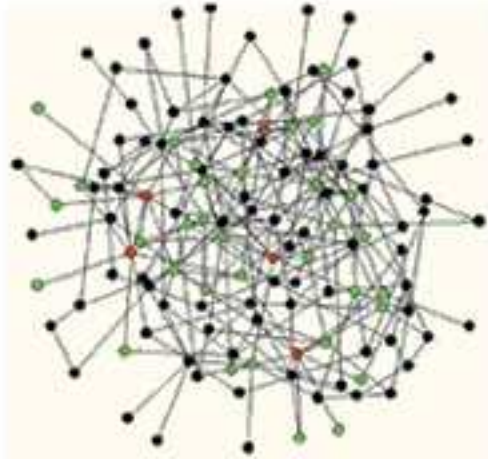
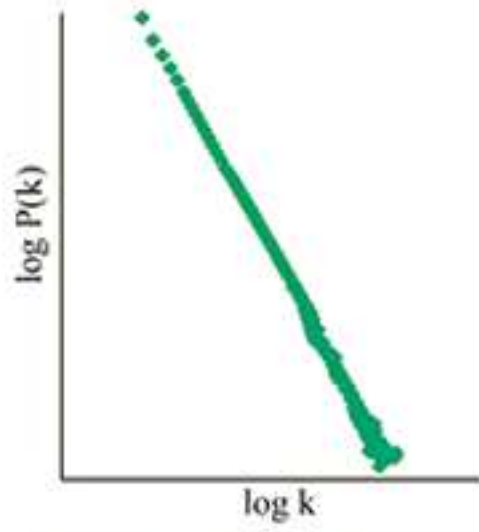
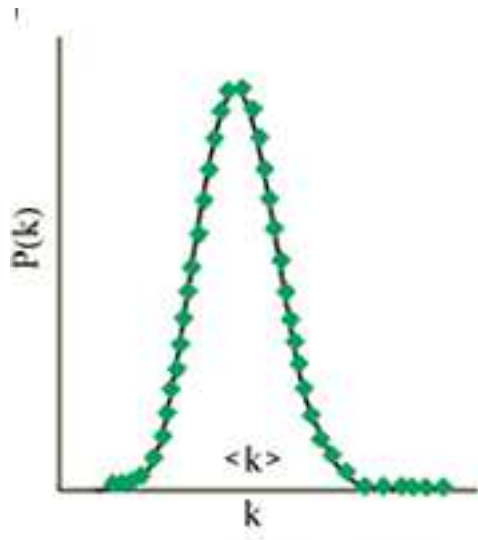
(większość sieci rzeczywistych posiada rozkład potęgowy)

2. niski współczynnik grupowania

$$C \approx \frac{z}{n} \quad (4)$$

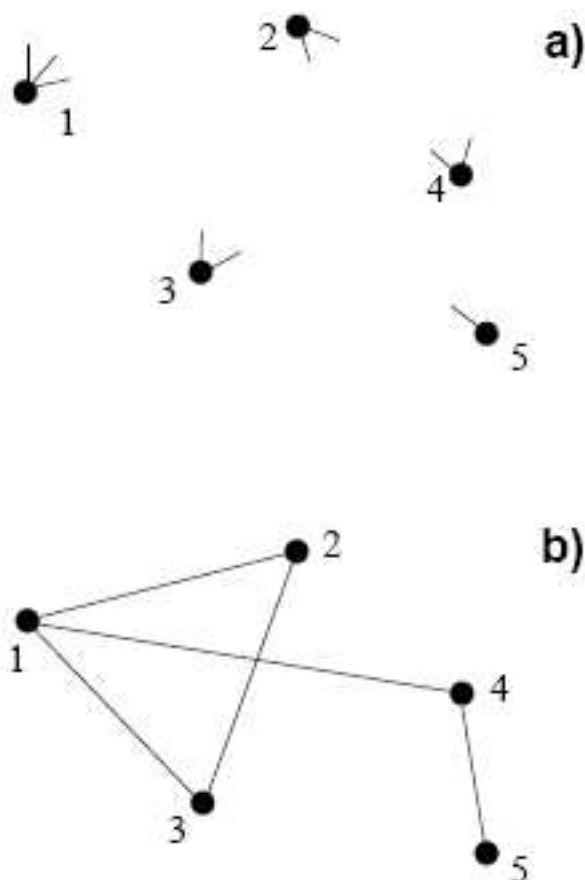
- większość rzeczywistych sieci ma wysoki współczynnik C (zjawisko *the friend of my friend is also my friend*)

network	n	z	clustering coefficient C	
			measured	random graph
Internet (autonomous systems) ^a	6 374	3.8	0.24	0.00060
World-Wide Web (sites) ^b	153 127	35.2	0.11	0.00023
power grid ^c	4 941	2.7	0.080	0.00054
biology collaborations ^d	1 520 251	15.5	0.081	0.000010
mathematics collaborations ^e	253 339	3.9	0.15	0.000015
film actor collaborations ^f	449 913	113.4	0.20	0.00025
company directors ^f	7 673	14.4	0.59	0.0019
word co-occurrence ^g	460 902	70.1	0.44	0.00015
neural network ^c	282	14.0	0.28	0.049
metabolic network ^h	315	28.3	0.59	0.090
food web ⁱ	134	8.7	0.22	0.065



W jaki sposób uogólnić klasyczny model grafu losowego, aby lepiej modelował sieci rzeczywiste?

- graf posiadający specyficzny rozkład stopni wierzchołków p_k (lub ustalony ciąg stopni wierzchołków $\{k_i\}$, $i = 1, 2, \dots, n$ zbiegający do p_k dla $n \rightarrow \infty$).



Dla tak zdefiniowanego modelu:

$$z = \langle k \rangle = \sum_k k p_k, \quad (5)$$

$$z_2 = \langle k^2 \rangle - \langle k \rangle, \quad (6)$$

gdzie z_2 - średnia liczba sąsiadów drugiego rzędu dla losowo wybranego wierzchołka

Ogólnie:

$$z_m = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} z_{m-1} = \left(\frac{z_2}{z_1} \right)^{m-1} z_1 \quad (7)$$

Przejścia fazowe

Przejście fazowe to taka zmiana układu, której towarzyszy nagła zmiana parametrów układu, np. zmiana stanu skupienia układu lub jego składowych, perkolacja.

Wyróżniamy dwa parametry:

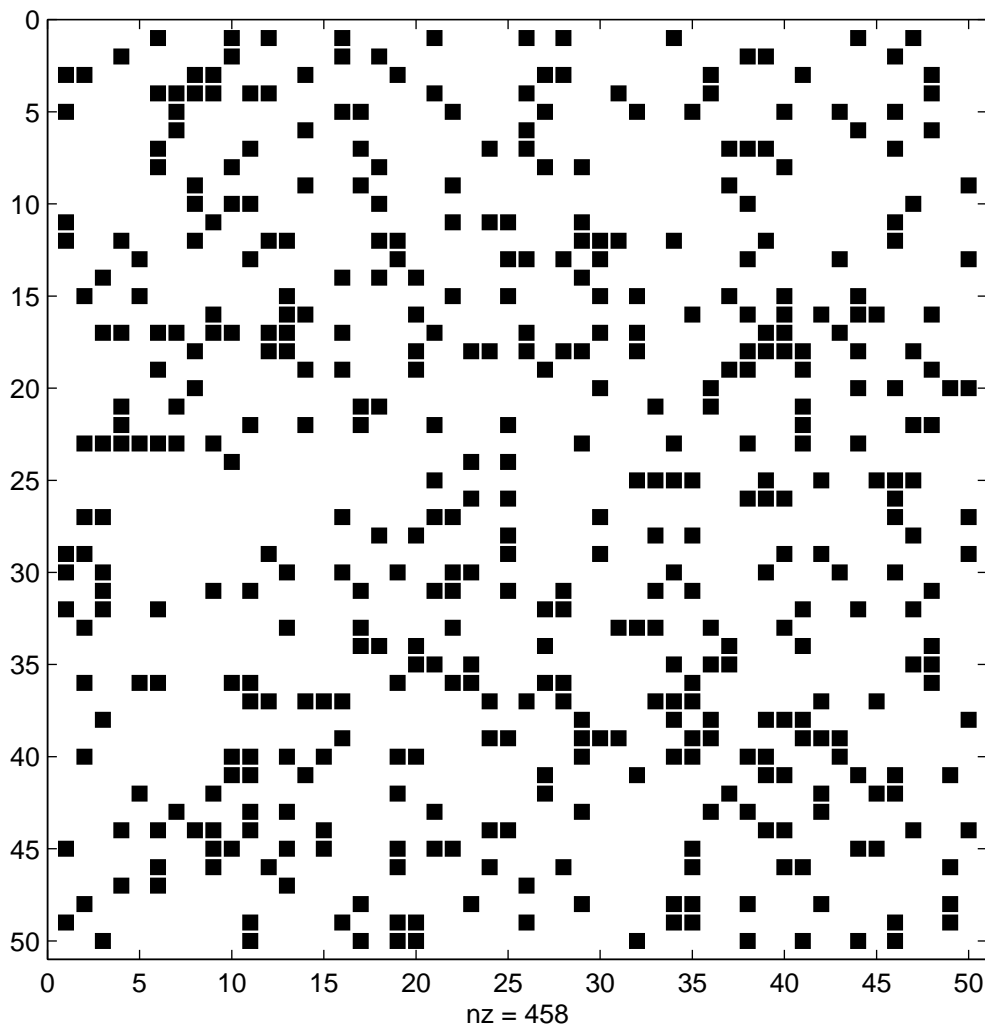
- parametr kontroli -np. p w modelu $G(n, p)$
- parametr porządku -np. S -liczba wierzchołków grafu $G(n, p)$ w tzw. *giant connected component* w stosunku do n

Podstawową zasadą, która konstytuuje dziedzinę fizyki zajmującą się teorią przejść fazowych jako samodzielny obszar badawczy, jest fakt, że zupełnie różne substancje przejawiają w ramach zjawisk towarzyszących przejściom fazowym takie samo zachowanie, co jest treścią hipotezy uniwersalności opisu przejść fazowych.

Przejście fazowe w grafie losowym polega na pojawieniu się tzw. *giant connected component*.

Perkolacja

W matematyce **teoria perkolacji** opisuje zachowanie się połączonych grup wierzchołków w grafie losowym. Znajduje ona także szersze zastosowanie, np. w chemii czy inżynierii materiałowej.





W klasycznym grafie losowym $G_{n,p}$ obserwujemy przejście fazowe, gdy $z = 1$.

W grafie z danym rozkładem stopni wierzchołków, gdy $z_1 = z_2$ lub równoważnie, gdy:

$$\langle k^2 \rangle - 2\langle k \rangle = 0 \iff \sum_{k=0}^{\infty} k(k-2)p_k = 0 \quad (8)$$

(Molloy, Reed 1995)

Powyżej przejścia fazowego możemy rozważać średnią odległość między dwoma wierzchołkami grafu - l :

$$l = \frac{\log(n/z_1)}{\log(z_2/z_1)} + 1 \quad (9)$$

Zauważmy, że nawet w bardzo dużych sieciach l jest dosyć małe-zjawisko to znane jest jako **small-world effect**

Współczynnik grupowania dla uogólnionego grafu losowego:

$$C = \frac{\langle k_i k_j \rangle}{nz} = \frac{z}{n} \left[\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle^2} \right]^2 = \frac{z}{n} \left[c_v^2 + \frac{z-1}{z} \right]^2 \quad (10)$$

Funkcje generujące rozkłady prawdopodobieństwa

- a) p_k -prawdopodobieństwo, że losowo wybrany wierzchołek ma stopień k :

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k, \quad (11)$$

$$p_k = \frac{1}{k!} \left[\frac{d^k G_0}{dx^k} \right]_{x=0} \quad (12)$$

- b) q_k -prawdopodobieństwo, że losowo wybrana krawędź kończy się wierzchołkiem stopnia $k + 1$:

$$\begin{aligned} G_1(x) &= \sum_{k=0}^{\infty} q_k x^k = \frac{\sum_{k=0}^{\infty} (k+1)p_{k+1} x^k}{\sum_j j p_j} = \\ &= \frac{\sum_{k=0}^{\infty} (k)p_k x^{k-1}}{\sum_j j p_j} = \frac{G_0'(x)}{z} \quad (13) \end{aligned}$$

Własności funkcji generujących:

- 1) jeżeli rozkład, który generuje funkcja jest poprawnie znormalizowany, to:

$$G_0(1) = \sum_k p_k = 1 \quad (14)$$

- 2) wartość oczekiwaną możemy obliczyć jako:

$$G'_0(1) = \sum_k k p_k = \langle k \rangle \quad (15)$$

- 3) ogólnie, n -ty moment rozkładu obliczamy jako:

$$\langle k^n \rangle = \sum_k k^n p_k = \left[\left(x \frac{d}{dx} \right)^n G_0(x) \right]_{x=1} \quad (16)$$

- 4) jeśli funkcja generuje rozkład prawdopodobieństwa dla pewnej własności k danego obiektu (np. stopień wierzchołka w grafie), to rozkład tej własności dla n niezależnych obiektów jest generowany przez $[G_0(x)]^n$

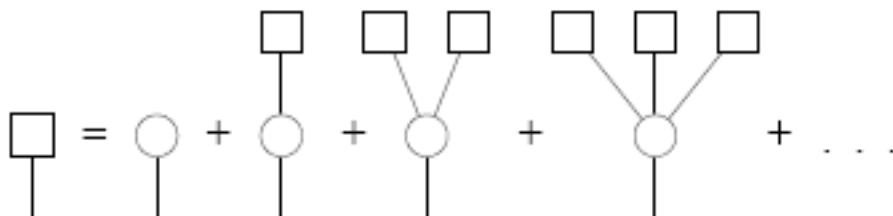
Rozmiary składowych spójności grafu

A. Poniżej przejścia fazowego

Każda skończona składowa grafu nie ma cykli - ma strukturę drzewa ($C \rightarrow 0$ dla $n \rightarrow \infty$). Wybierzmy losową krawędź. Rozważmy zbiór wierzchołków, które są "osiągalne" z jednego końca tej krawędzi - *klaster*.

Niech $H_1(x)$ generuje rozkład liczby wierzchołków w takim klasterze (jego rozmiar):

$$H_1(x) = x \sum_{k=0}^{\infty} q_k [H_1(x)]^k = xG_1(H_1(x)) \quad (17)$$



Funkcja generująca rozkład liczby wierzchołków w składowej spójności, do której należy losowo wybrany wierzchołek:

$$H_0(x) = x \sum_{k=0}^{\infty} p_k [H_1(x)]^k = xG_0(H_1(x)). \quad (18)$$

Średni rozmiar składowej spójności:

$$\begin{aligned} \langle s \rangle = H_0'(1) &= \left[G_0(H_1(x)) + xG_0'(H_1(x))H_1'(x) \right]_{x=1} \\ &= 1 + G_0'(1)H_1'(1) \end{aligned} \quad (19)$$

lub równoważnie:

$$\langle s \rangle = 1 + \frac{z_1^2}{z_1 - z_2} \quad (20)$$

Przejście fazowe obserwujemy, gdy $z_1 = z_2$ lub $G_1'(1) = 1$.

B. Powyżej przejścia fazowego

-większość sieci badanych doświadczalnie znajduje się w tym stanie, ma tzw. *giant connected component* (GCC). GCC nie ma struktury drzewa dla $n \rightarrow \infty$.

$H_0(x)$, $H_1(x)$ - funkcje generujące dla rozkładu wielkości składowych spójności z wyłączeniem GCC

P_s -rozkład prawdopodobieństwa dla rozmiarów składowych spójności (poza GCC):

$$H_0(1) = \sum_s P_s,$$

$H_0(1)$ = liczba wierzchołków grafu poza GCC w stosunku do n .

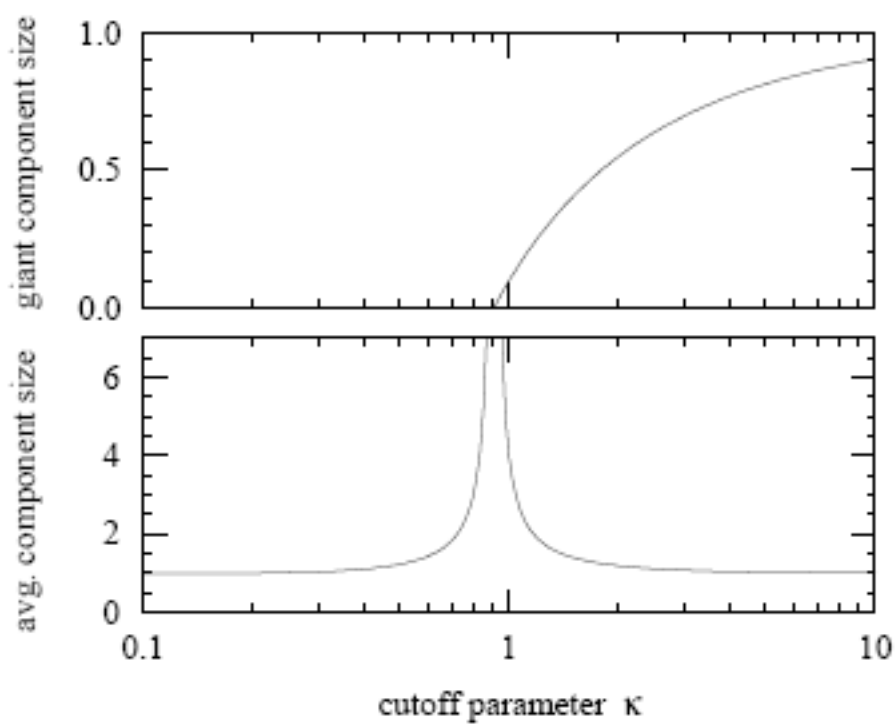
Rozmiar GCC, S , musi być rozwiązaniem układu:

$$S = 1 - G_0(v), \quad v = G_1(v), \quad (21)$$

gdzie $v \equiv H_1(1)$.

Średni rozmiar składowej spójności grafu:

$$\langle s \rangle = 1 + \frac{zv^2}{[1 - S][1 - G_1'(v)]} \quad (22)$$



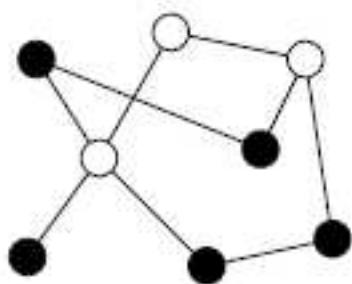
Teoria perkolacji a odporność sieci na "uszkodzenia"

Proces perkolacji w sieci polega ogólnie na losowym podziale wierzchołków lub krawędzi na dwa zbiory: "czynne" i "nieczynne" (*working and not working*).

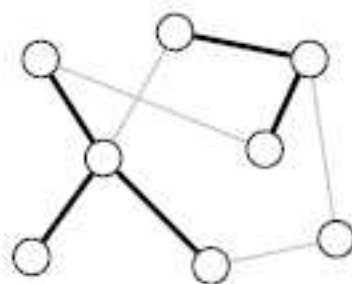
Model perkolacji został po raz pierwszy zaproponowany w latach 50-tych; motywacją była chęć lepszego zrozumienia zjawiska rozprzestrzeniania się chorób zakaźnych.

Wyróżniamy dwa rodzaje perkolacji:

site percolation i **bond percolation**:



site percolation



bond percolation

Miarą odporności sieci na losowe usuwanie wierzchołków może być zmiana (lub brak zmiany) liczby wierzchołków, które znajdują się w największej składowej grafu (GCC).

Proces losowego wyłączenia wierzchołków można rozpatrywać jako **site percolation**.

Wierzchołki, które pozostają w sieci czynne (mogą komunikować się ze sobą) tworzą **giant connected component** w odpowiadającym modelu perkolacji.

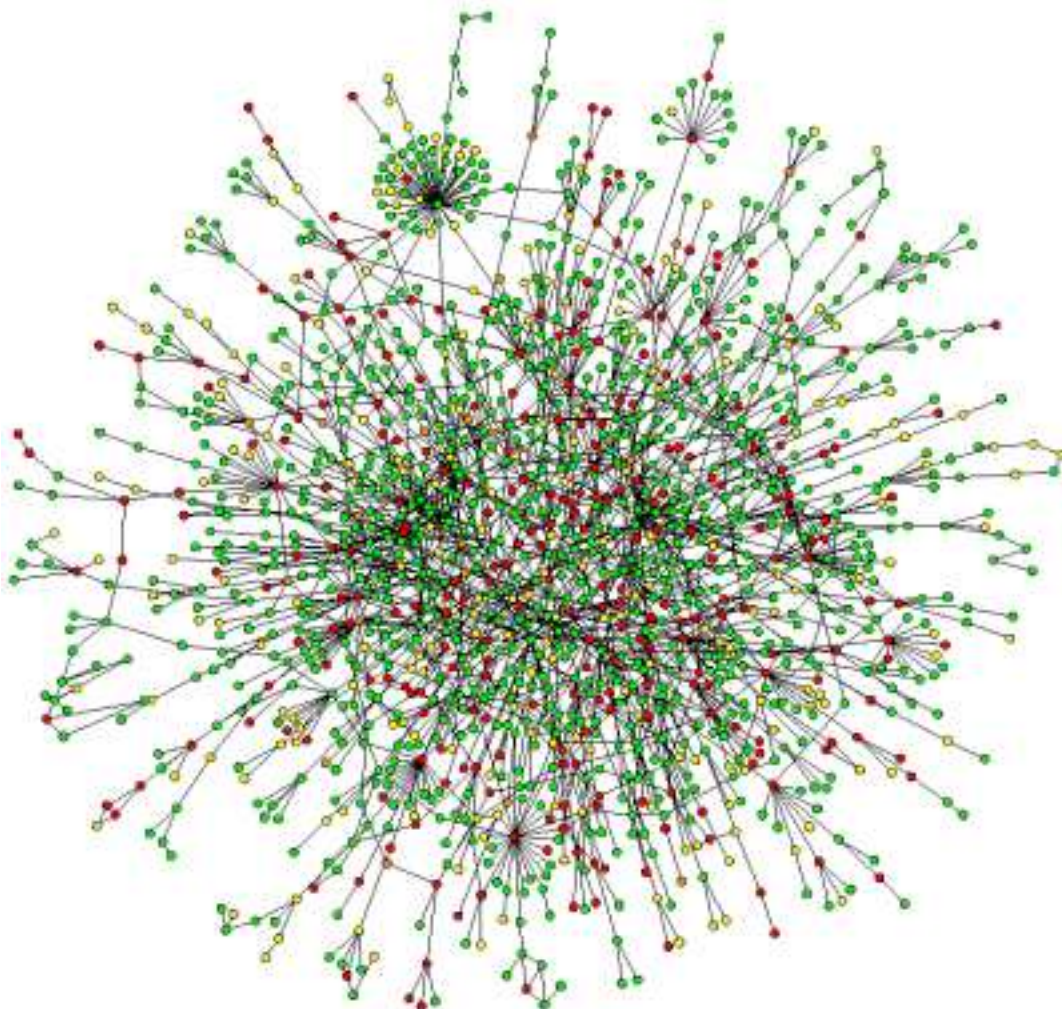
Rozważmy model grafu losowego, gdzie p_k to rozkład prawdopodobieństwa stopni wierzchołków. Załóżmy, że q to część wierzchołków grafu, które są czynne (wierzchołki te losujemy jednostajnie z całego grafu). Wtedy

$$p'_k = \sum_{k=k'}^{\infty} p_k \binom{k}{k'} q^{k'} (1 - q)^{k-k'} \quad (23)$$

jest prawdopodobieństwem, że losowo wybrany wierzchołek czynny jest połączony z k' innymi czynnymi wierzchołkami.

Ponieważ wyłączanie wierzchołków jest losowe i niezależne, to podzbiór wierzchołków czynnych tworzy inny *model konfiguracyjny*, gdzie $p_{k'}$ jest rozkładem stopni wierzchołków.

Ciekawe wyniki zostały pokazane dla sieci z rozkładem potęgowym $p_k \sim k^{-\alpha}$ (α ustalone). Dla $\alpha \leq 3$ wartość krytyczna q_c , kiedy dochodzi do przejścia fazowego i tworzy się GCC, jest niedodatnia, co oznacza, że sieć zawsze *perkoluje*. Pokazano ogólnie, że $g_c \leq 0$ dla sieci o rozkładzie p_k , gdzie $\langle k^2 \rangle \rightarrow \infty$ dla $n \rightarrow \infty$.



Prawdopodobieństwo, że dany wierzchołek jest czynny może zależeć od jego stopnia k . Wtedy zamiast stałej q mamy q_k - prawdopodobieństwo, że wierzchołek stopnia k jest czynny. Funkcje generujące:

$$F_0(x) = \sum_{k=0}^{\infty} p_k q_k x^k, \quad F_1(x) = \frac{\sum_k k p_k q_k x^{k-1}}{\sum_k k p_k} \quad (24)$$

Rozkład prawdopodobieństwa rozmiarów składowych grafu, tworzonych przez wierzchołki czynne, do których należy losowo wybrany wierzchołek, jest generowany poprzez funkcję $H_0(x)$:

$$H_0(x) = 1 - F_0(1) + x F_0(H_1(x)), \quad (25)$$

gdzie:

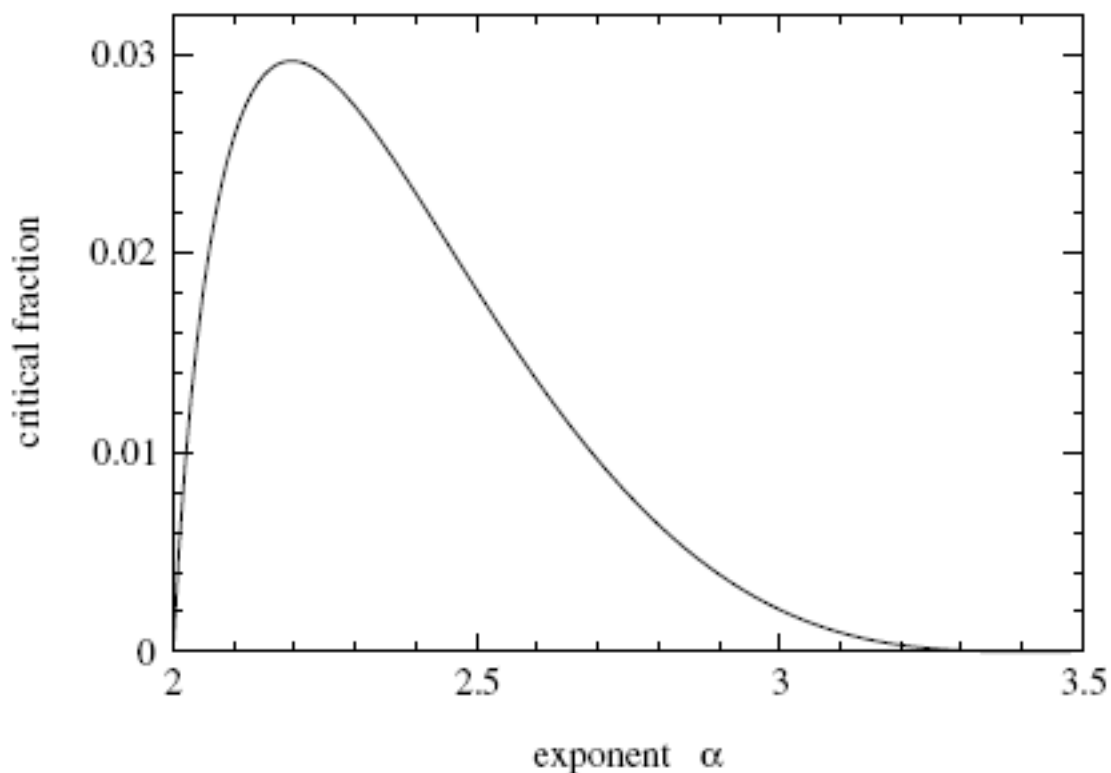
$$H_1(x) = 1 - F_1(1) + x F_1(H_1(x)) \quad (26)$$

Średni rozmiar składowej grafu (tworzonej przez wierzchołki czynne):

$$\langle s \rangle = F_0(1) + \frac{F_0'(1) F_1(1)}{1 - F_1'(1)} \quad (27)$$

GCC formuje się, gdy $F_1'(1) = 1$.

W niektórych sieciach, np. sieciach przesyłających energię elektryczną, zjawisko, gdy nagle jedna krawędź lub wierzchołek przestaje działać, może w rezultacie zaburzyć funkcjonowanie całej sieci. Sieci opisane rozkładem potęgowym są szczególnie wrażliwe na usuwanie wierzchołów o wysokim stopniu.



Epidemiologia

Standardowe matematyczne podejście do problemu rozprzestrzeniania się choroby zakaźnej w populacji opiera się na założeniu, że każda para osób ma równe szanse kontaktu ze sobą (tzw. *fully mixed approximation*). Założenie to jest nierealistyczne. Realistyczne modele używają struktury sieci.

Najprostszym z nich jest **model SIR** (Reed, Frost \approx 1920):

S - *susceptible*

I - *infective*

R - *recovered*

Proces rozprzestrzeniania się choroby w sieci reprezentującej populację może być utożsamiany z **bond percolation** dla tej samej sieci.

Niech β oznacza prawdopodobieństwo, że osoba zainfekowana zarazi swojego sąsiada (z grupy S) w ustalonej jednostce czasu. Wielkości β są wylosowane z rozkładu $P_i(\beta)$. Niech γ oznacza prawdopodobieństwo, że losowo wybrana osoba zainfekowana wyzdrowieje (w jednostce czasu), gdzie $P_r(\gamma)$ jest rozkładem odpowiadającym tej wielkości. Uzyskany model okazuje się być równoważny z jednostajnym **bond percolation**:

$$T = 1 - \int_0^\infty P_i(\beta) P_r(\gamma) e^{-\beta/\gamma} d\beta d\gamma, \quad (28)$$

gdzie T to prawdopodobieństwo, że losowo wybrana krawędź jest czynna (nastąpi zarażenie).

Model perkolacji dostarcza nam wielu informacji na temat rozprzestrzeniania się choroby, głównie jej rozmiarów:

rozkład rozmiarów składowych grafu utworzonych przez "funkcjonujące" krawędzie to rozkład rozmiarów ognisk choroby zapoczątkowanych przez pojedynczą zainfekowaną osobę, a przejście fazowe to tzw. "próg epidemiologiczny" -stan, powyżej którego możliwy jest wybuch epidemii-jej zasięg to liczba wierzchołków w GCC.

Niestety model ten nie pozwala wnioskować o ewolucji ognisk choroby w czasie.

Interesujące wnioski zostały sformułowane dla sieci z rozkładem potęgowym $p_k \sim k^{-\alpha}$:

jeśli tylko $\alpha \leq 3$, to

"próg epidemiologiczny" ≤ 0 .

Większość realistycznych sieci realizuje to założenie, tak więc "choroby" będą się zawsze na nich rozprzestrzeniały, nie zależnie od rozkładu β (po raz pierwszy pokazano to dla wirusów komputerowych - Pastor-Satorras i Vespignani)

Połączenie zjawiska odporności sieci na losowe usuwanie wierzchołków z problemem rozprzestrzeniania się na niej choroby uzyskujemy, gdy rozważamy **szczepienie losowo wybranych osób** z populacji przeciwko danej chorobie - to z kolei możemy modelować jako **site percolation**. Jeżeli proces site percolation jest dodatnio skorelowany ze stopniem wierzchołka, to otrzymujemy efektywną strategię przeciwdziałania rozprzestrzenianiu się choroby.

Bibliografia:

1. M. Newman: *Random graphs as model of networks.*
2. S.Dorogovtsev, J.Mendes, A.Samukhin:
Modern architecture of random graphs: Constructions and correlations.
3. M. Newman: *The structure and function of complex networks.*
4. S.Janson, D.Knuth, T.Łuczak, B.Pittel
The Birth of the Giant Component.
5. M. Newman, S.Strogatz, D. Watts:
Random graphs with arbitrary degree distributions and their applications.