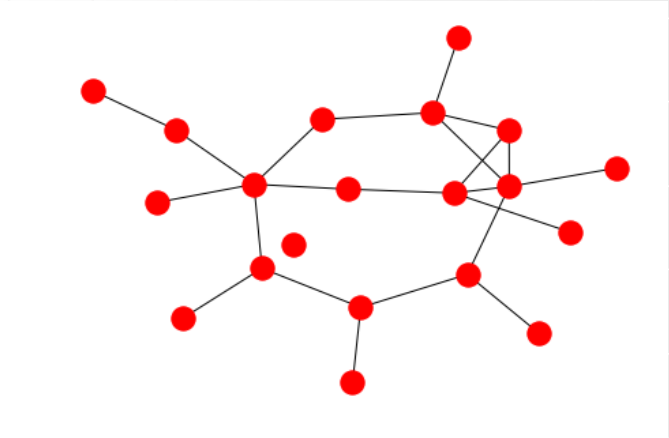
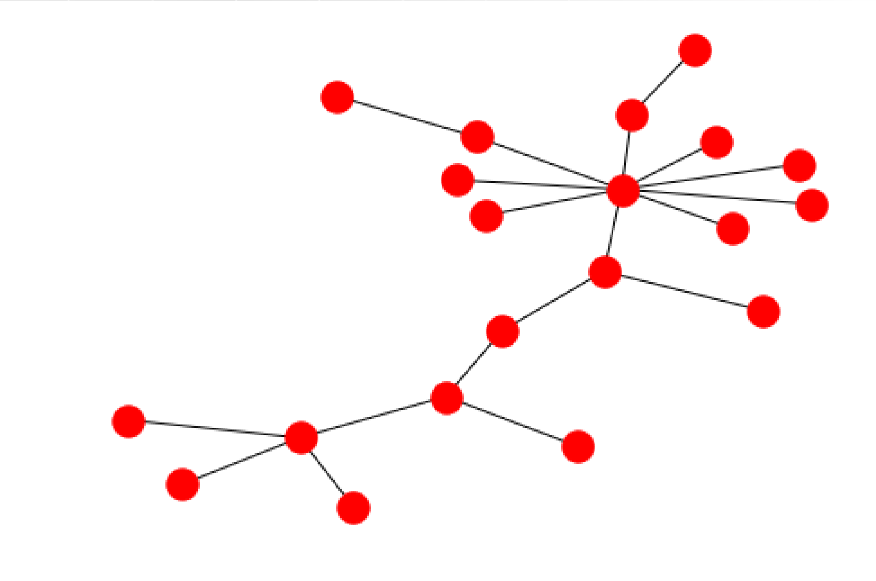
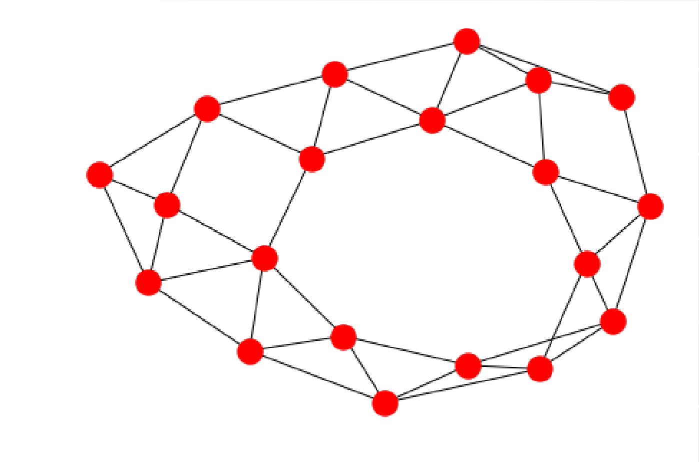
Politechnika Wrocławska

Sieci złożone





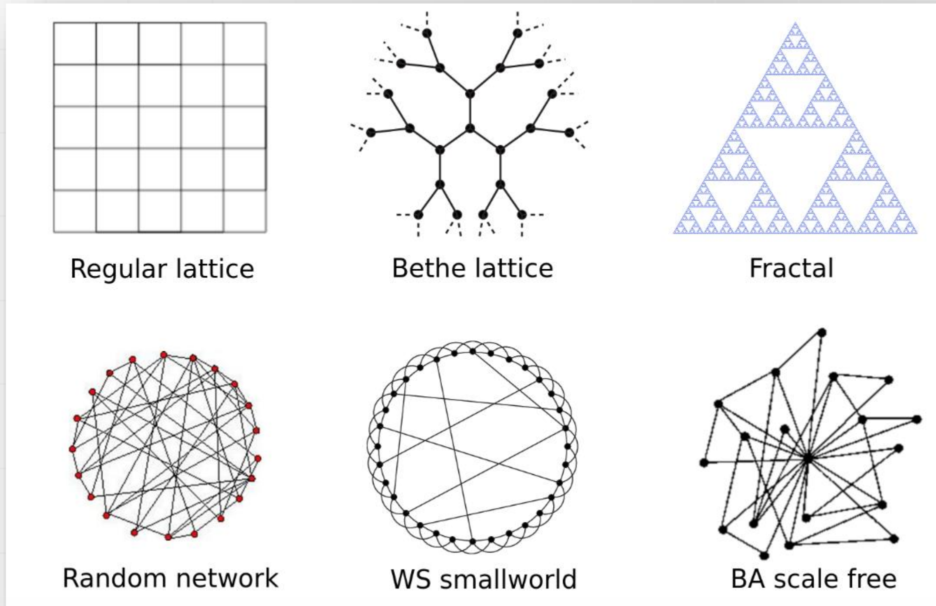


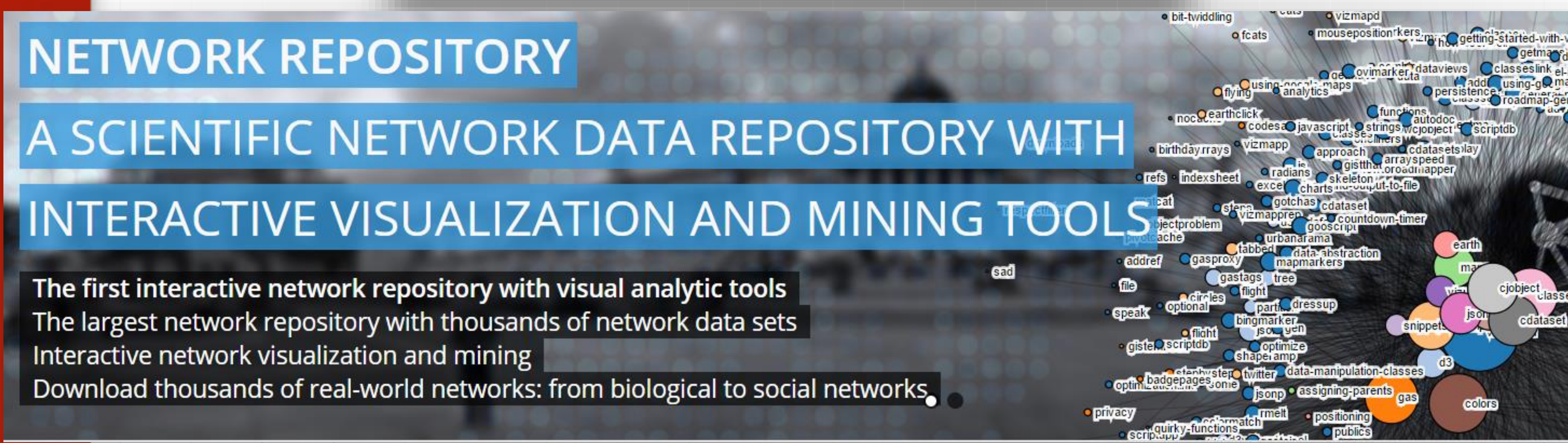
HR EXCELLENCE IN RESEARCH

Dane rzeczywiste

A co jeżeli nie mamy:

Sztuczne i darmowe





Politechnika Wrocławska

Model sieci losowej

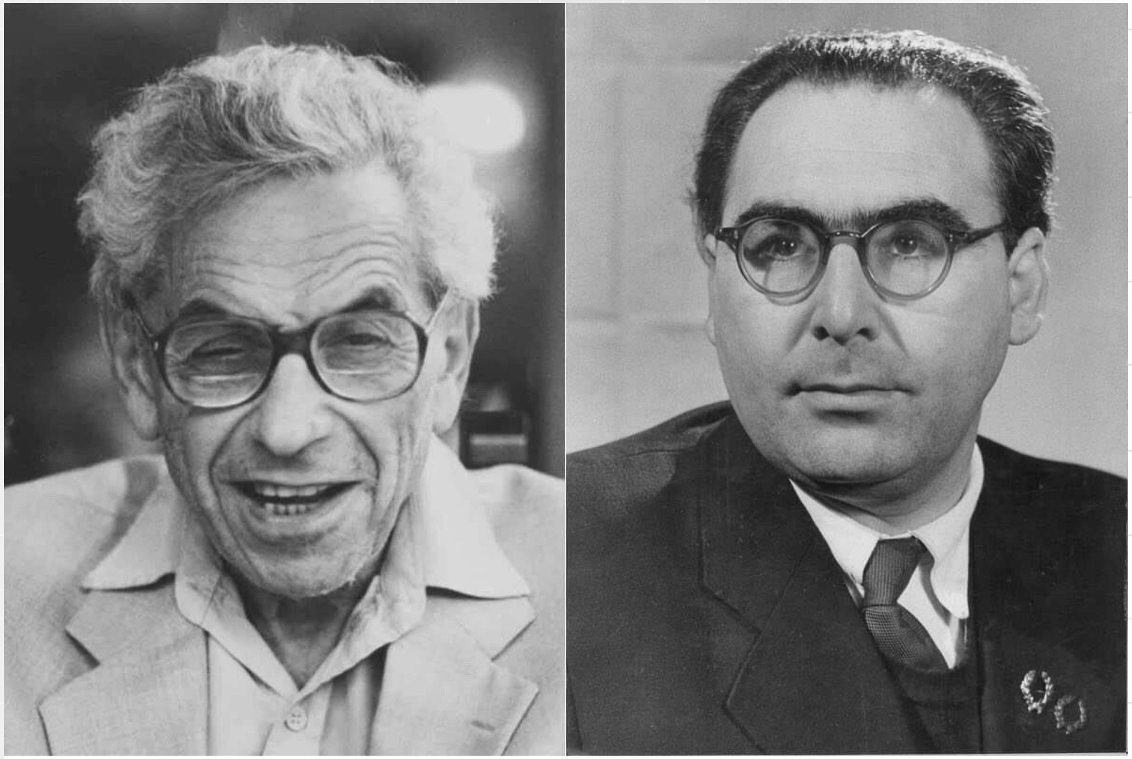
- Erdős, P. Rényi, A (1959) "On Random Graphs I" in Publ. Math. Debrecen 6, p. 290–297
- Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, *30*(4), 1141-1144.

HR EXCELLENCE IN RESEARCH

Politechnika Wrocławska

Model sieci losowej

- Erdös-Rényi model (1959) - Erdős, P. Rényi, A (1959) "On Random Graphs I" in Publ. Math. Debrecen 6, p. 290–297



Paul Erdős (1913-1996) Alfréd Rényi (1921-1970) https://mathscinet.ams.org/mathscinet/collaborationDistance.html?group_source=1025821

Model sieci losowej - Erdös Number

Erdos Number = 5		
Piotr Bródka	coauthored with	Przemysław Kazienko
Przemysław Kazienko	coauthored with	Edwin Lughofer
Edwin Lughofer	coauthored with	Bernhard Alois Moser
Bernhard Alois Moser	coauthored with	István Joó
István Joó	coauthored with	Paul Erdős[1]

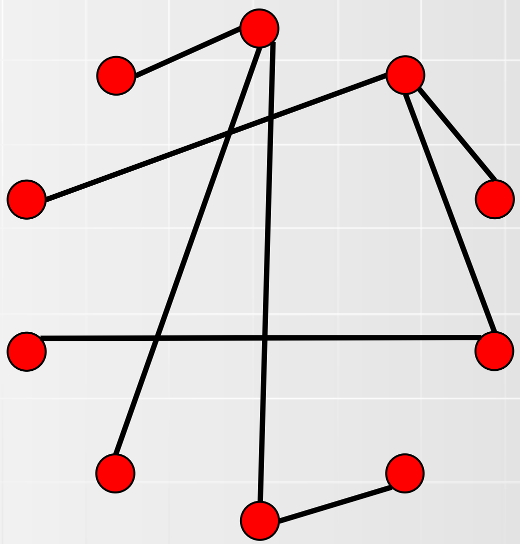
Model sieci losowej

- **Model *G(N,L)* - Erdös-Rényi 1959**
  - Erdös-Rényi model (1959) - Erdős, P. Rényi, A (1959) "On Random Graphs I" in Publ. Math. Debrecen 6, p. 290–297
  - *N* – liczba wierzchołków
  - *L* – liczba krawędzi losowo łącząca wierzchołki

- **Model *G(N,p)* – Gilbert 1959**
  - Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, *30*(4), 1141-1144.
  - *N* – liczba wierzchołków
  - *p* – prawdopodobieństwo istnienia krawędzi pomiędzy parą wierzchołków

Model sieci losowej

- Graf losowy to graf posiadający ***N*** wierzchołków gdzie każda para wierzchołków jest połączona z prawdopodobieństwem ***p***



N=10

p=1/6

Politechnika Wrocławska

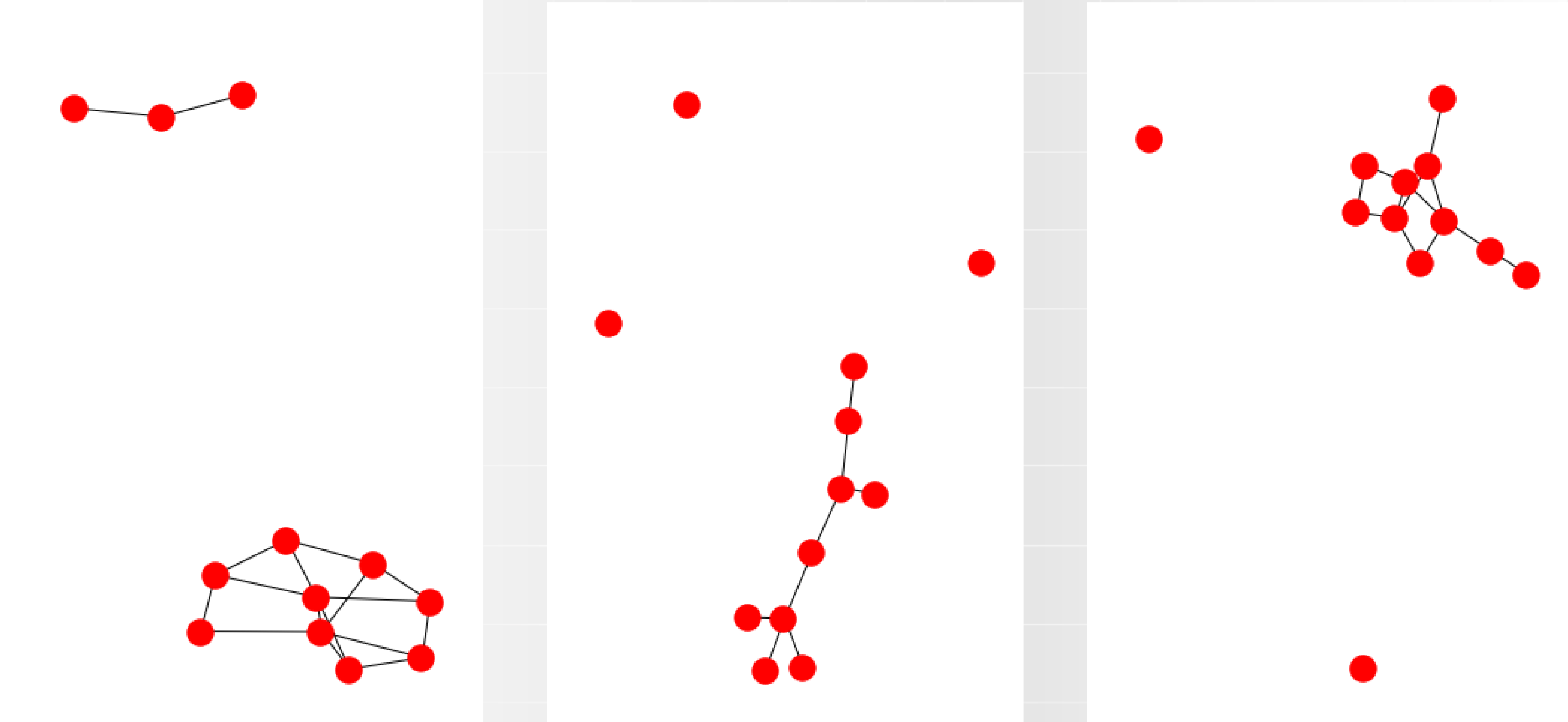
Model sieci losowej

Aby zbudować losową sieć, postępuj zgodnie z następującymi krokami:

1. Zacznij od N izolowanych węzłów.
2. Wybierz parę węzłów i wygeneruj losową liczbę z zakresu od 0 do 1. Jeśli liczba jest poniżej p, połącz wybraną parę węzłów krawędzią, w przeciwnym razie pozostaw je nie połączone.
3. Powtórz krok (2) dla każdej z $N(N\text{-}1)/2$ par węzłów.

Model sieci losowej

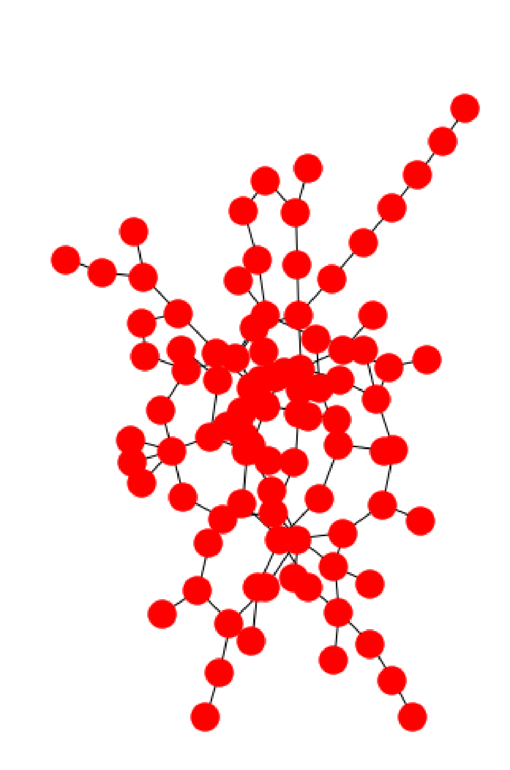
N=12, p=0.2



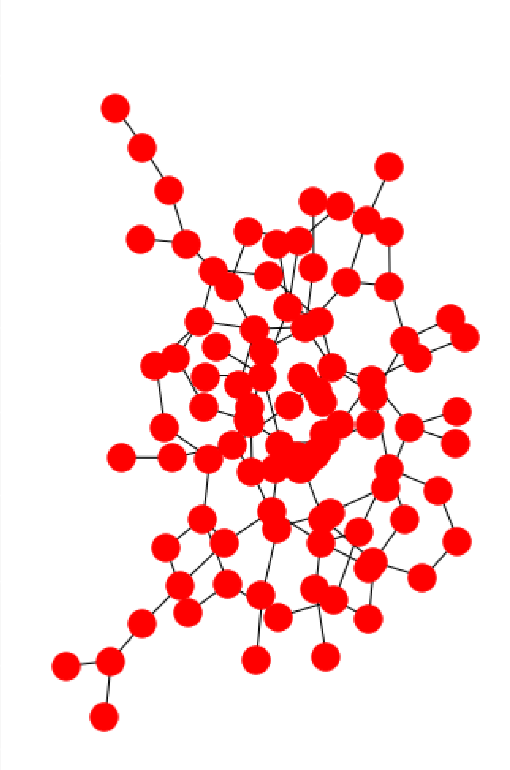
L=17 L=8 L=13

Model sieci losowej

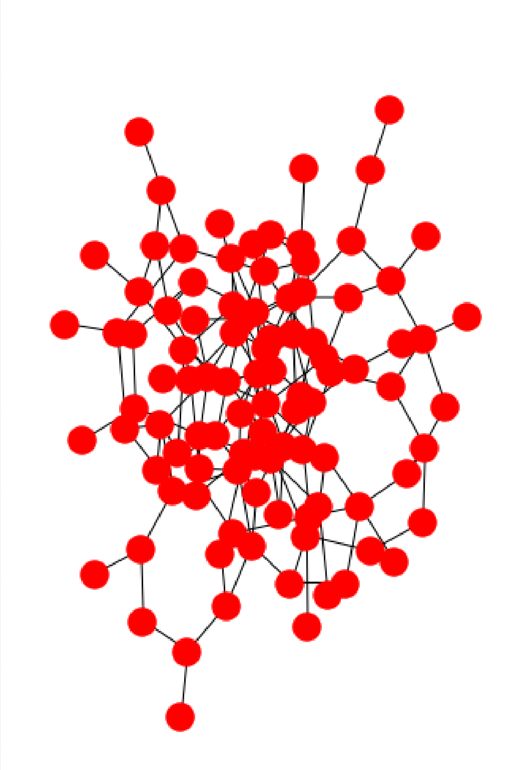
N=100, p=0.02



L=101



L=106



L=152

Politechnika Wrocławska

Rozkład stopni węzła w sieciach losowych

HR EXCELLENCE IN RESEARCH

Politechnika Wrocławska

Rozkład dwumianowy

- Rozklad dwumianowy to dyskretny rozkład prawdopodobieństwa opisujący liczbę sukcesów **x** w ciągu **N** niezależnych prób, z których każda ma stałe prawdopodobieństwo sukcesu równe **p**.
- Funkcja rozkładu prawdopodobieństwa

$$p_x = \binom{N}{x} p^x (1-p)^{N-x}$$

- Wartość oczekiwana (średnia)

$$\langle x \rangle = \sum_{x=0}^{N} x p_x = Np$$

Rozkład stopni węzła w sieciach losowych

W sieci losowej prawdopodobieństwo że wierzchołek **i** ma dokładnie **k** krawędzi składa się z trzech elementów:

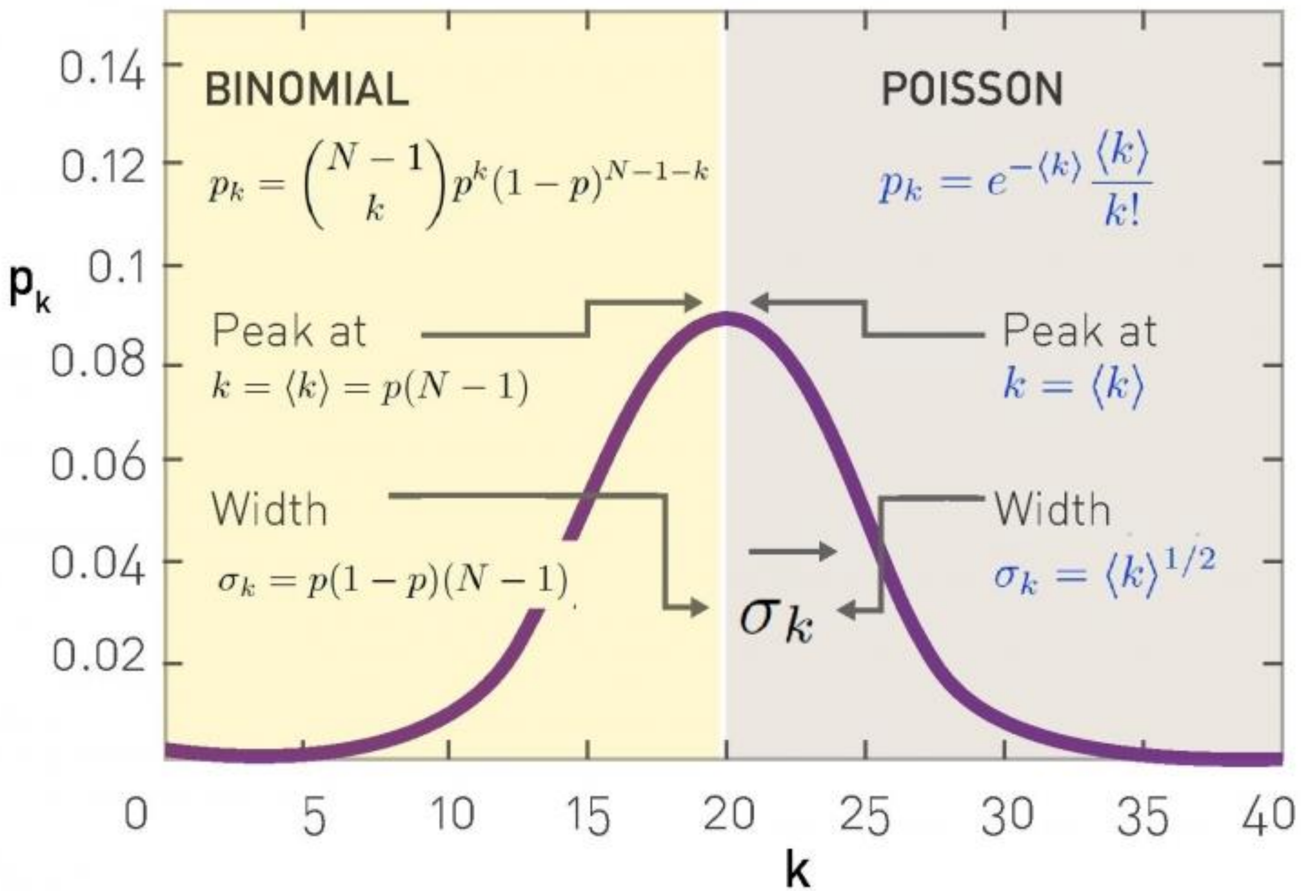
1. Prawdopodobieństwa, że ***k*** jego krawędzi istnieje p^k
2. Prawdopodobieństwa, że pozostałe **(*N-1-k*)** krawędzi nie istnieje $(1-p)^{N-1-k}$
3. Liczby kombinacji bez powtórzeń wyboru ***k*** krawędzi ze zbioru ***N - 1*** potencjalnych krawędzi $\binom{N-1}{k}$

W konsekwencji rozkład stopni węzła w sieci losowej jest rozkładem dwumianowym.

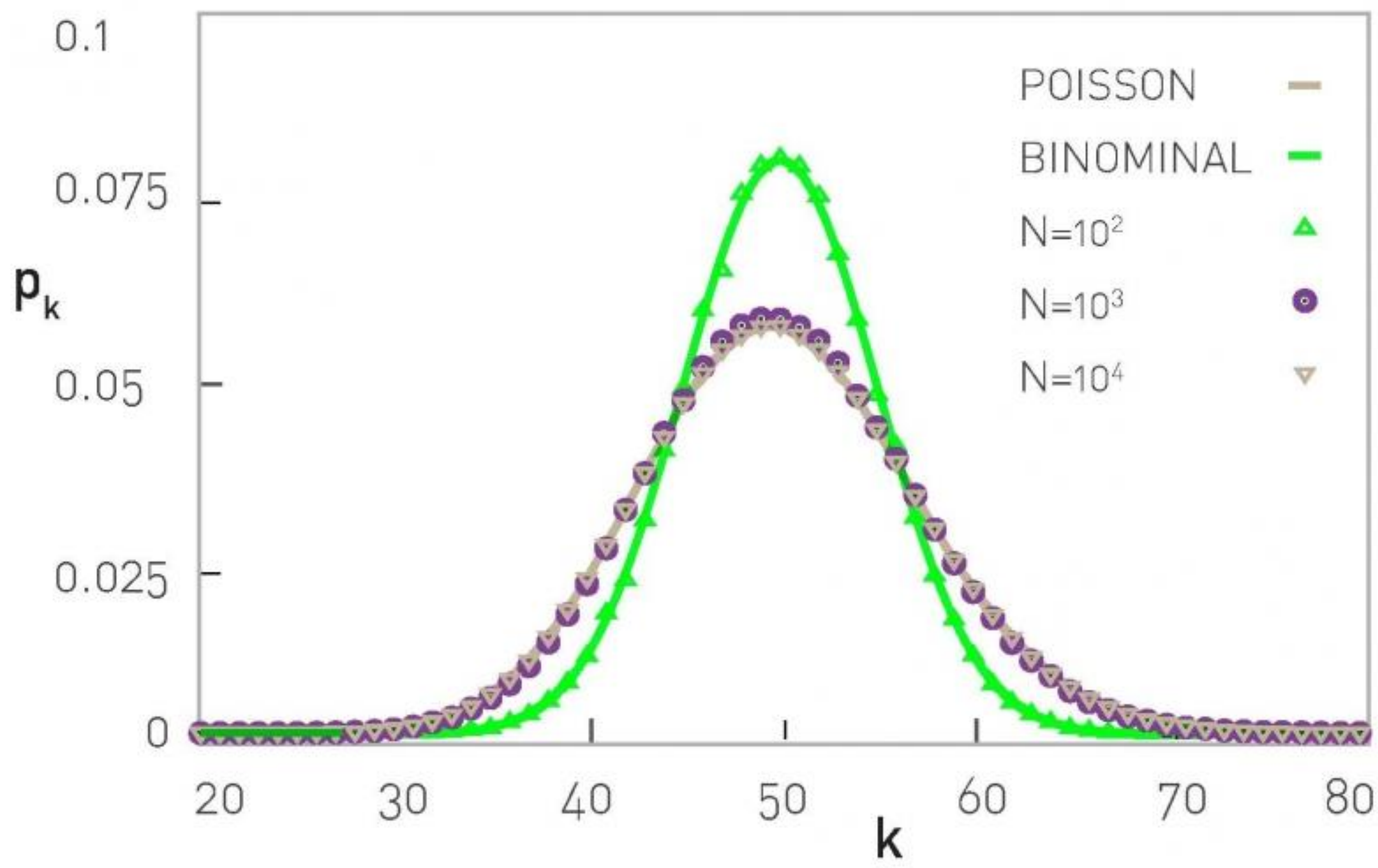
$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \qquad \langle k \rangle = \sum_{k=0}^{N-1} x p_k = (N-1)p$$

https://www.youtube.com/watch?v=eexQyHj6hEA

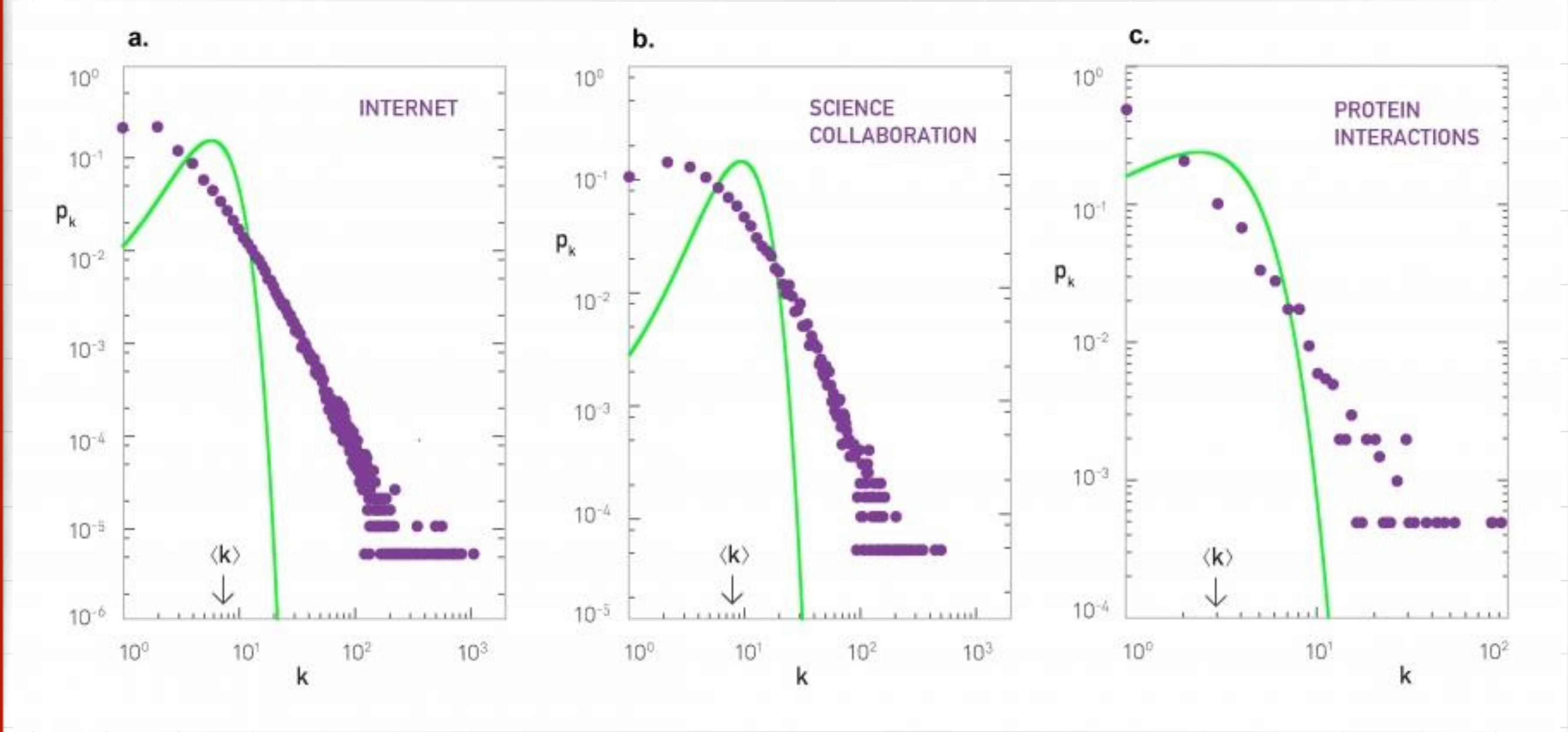
Rozkład stopni węzła w sieciach losowych



Rozkład stopni węzła w sieci losowej a rozmiar sieci



Rzeczywiste sieci nie są losowe



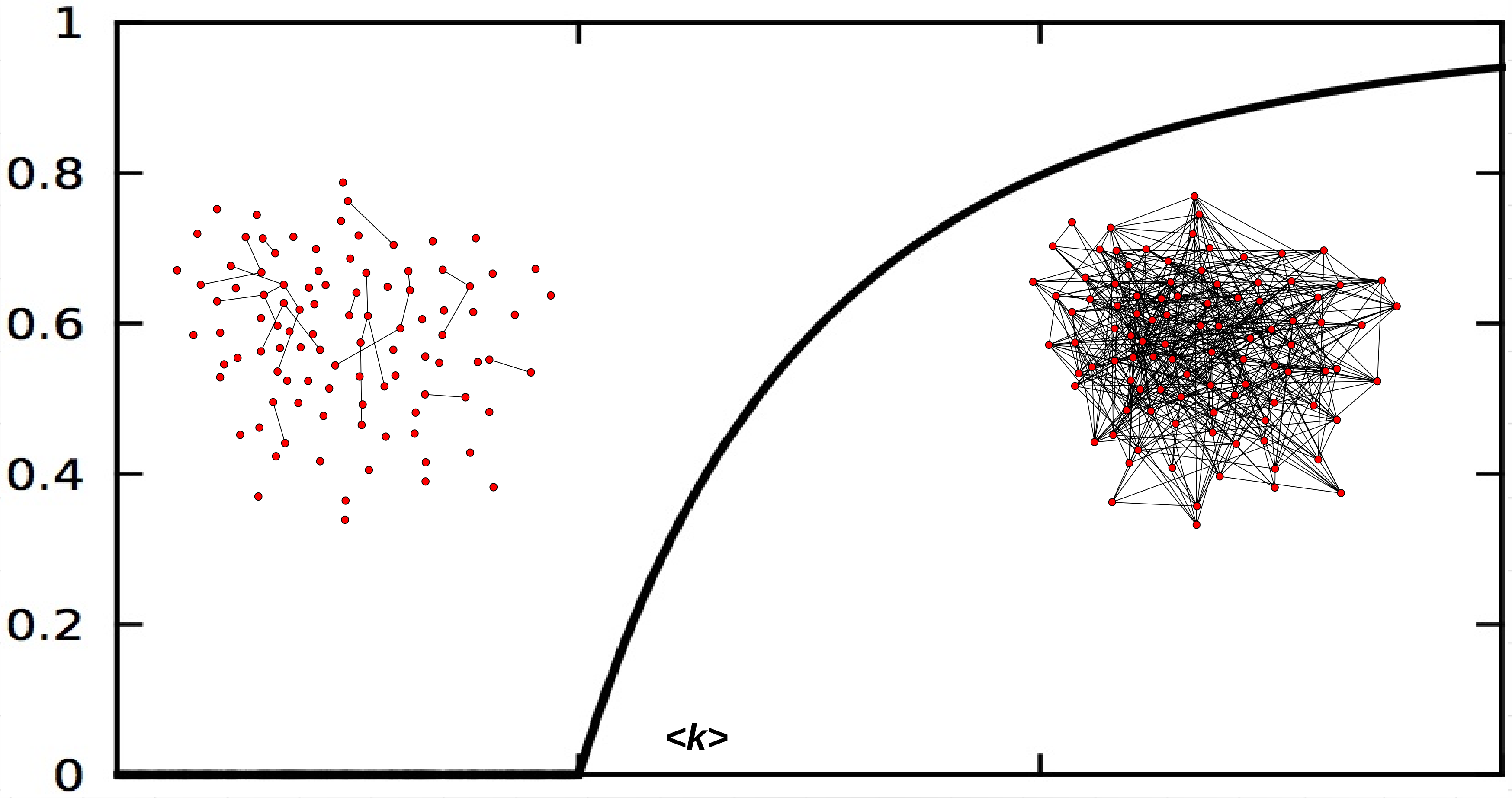
Połączeniowość w sieciach losowych

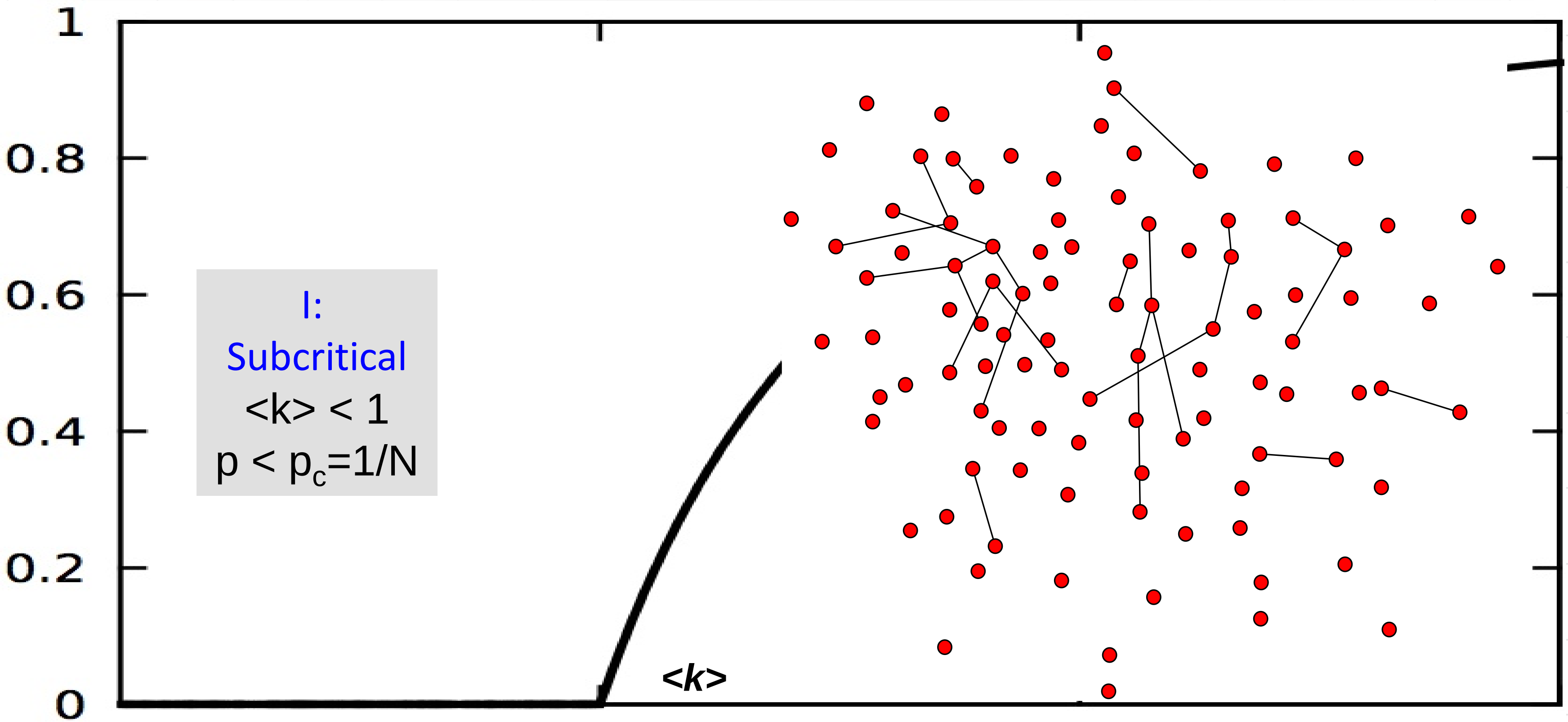
Politechnika Wrocławska

HR EXCELLENCE IN RESEARCH

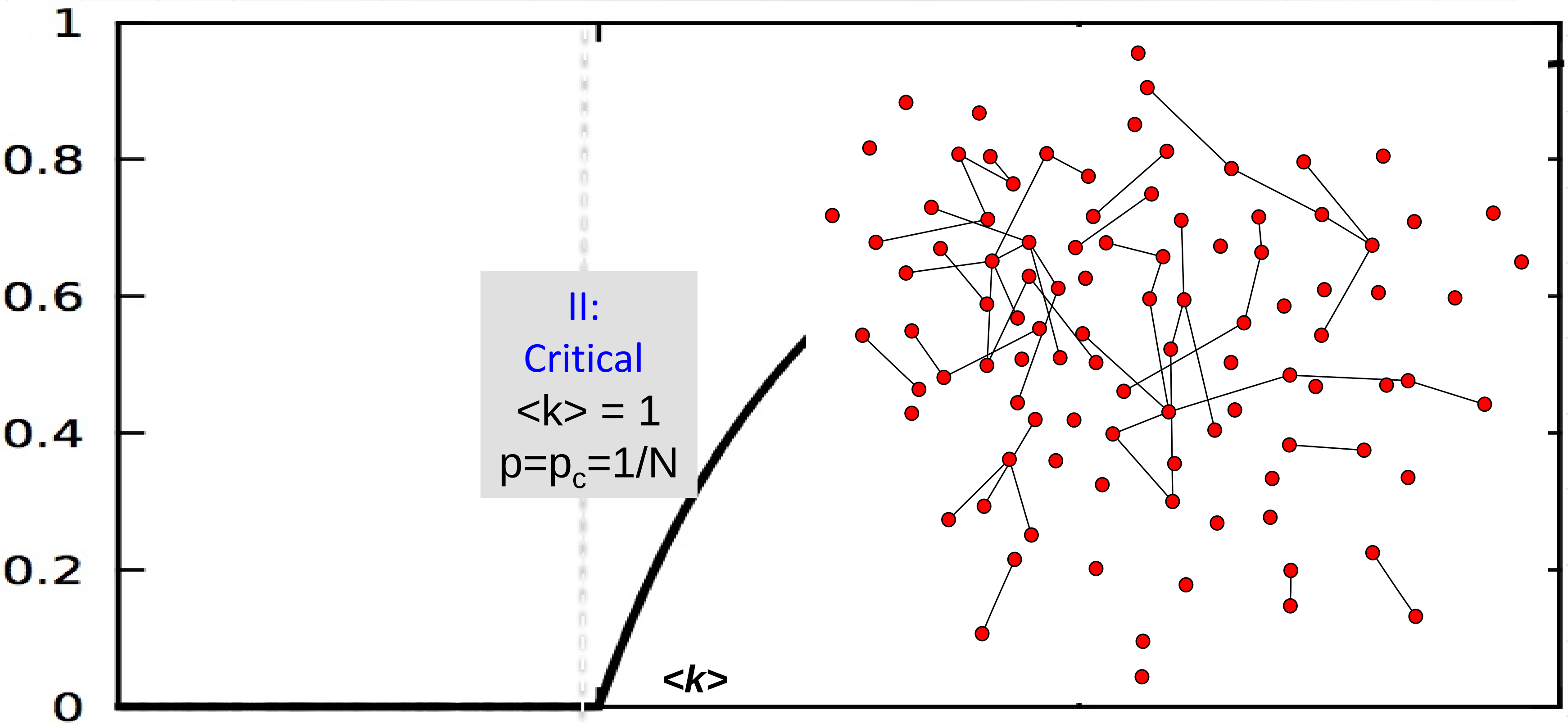
Ewolucja sieci losowej

Od izolowanych węzłów do połączonej sieci

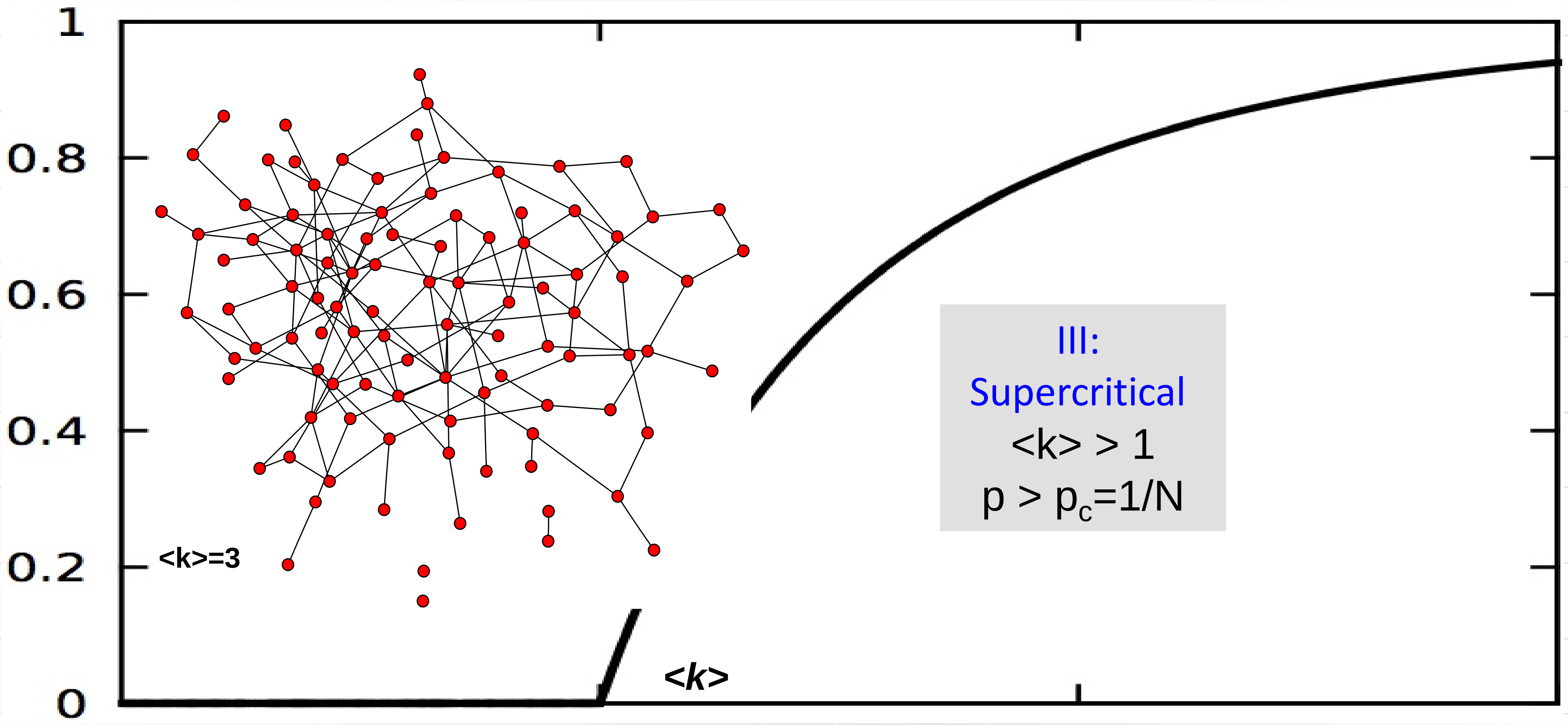




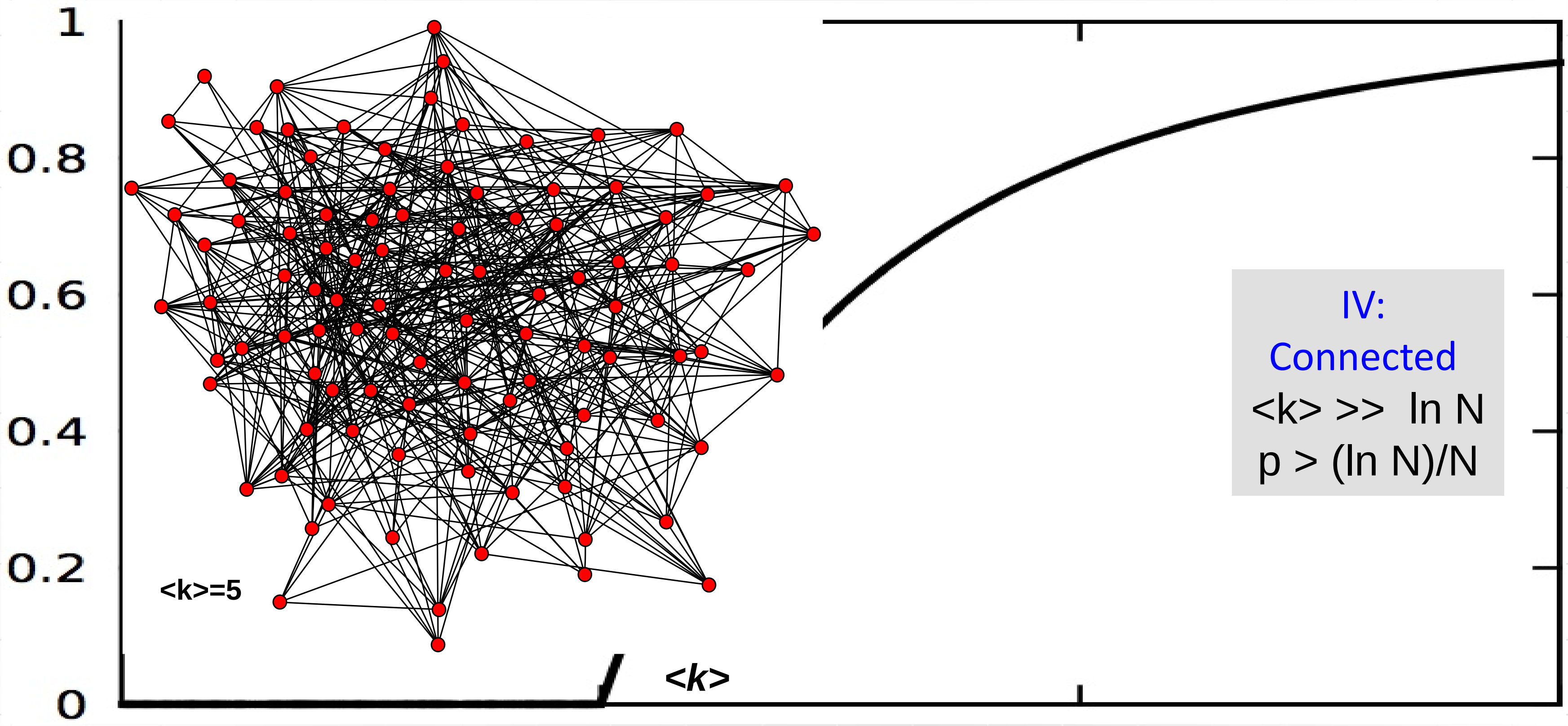
- „Brak” wielkiego komponentu
- N-L izolowanych grup/klastrów
- Największy klaster ma rozmiar ~ *ln N*



- Unikalny największy komponent: $N_G \sim N^{2/3}$ (Barabasi Advanced Topics 3.C)
  - N=1 000 - $ln_N \sim 6.9$; $N^{2/3} \sim 95$
  - N=7 000 000 000 - $ln_N \sim 22$; $N^{2/3} \sim 3,659,250$
- Pozostałe komponenty to zazwyczaj drzewa, natomiast największy komponent zazwyczaj ma cykle

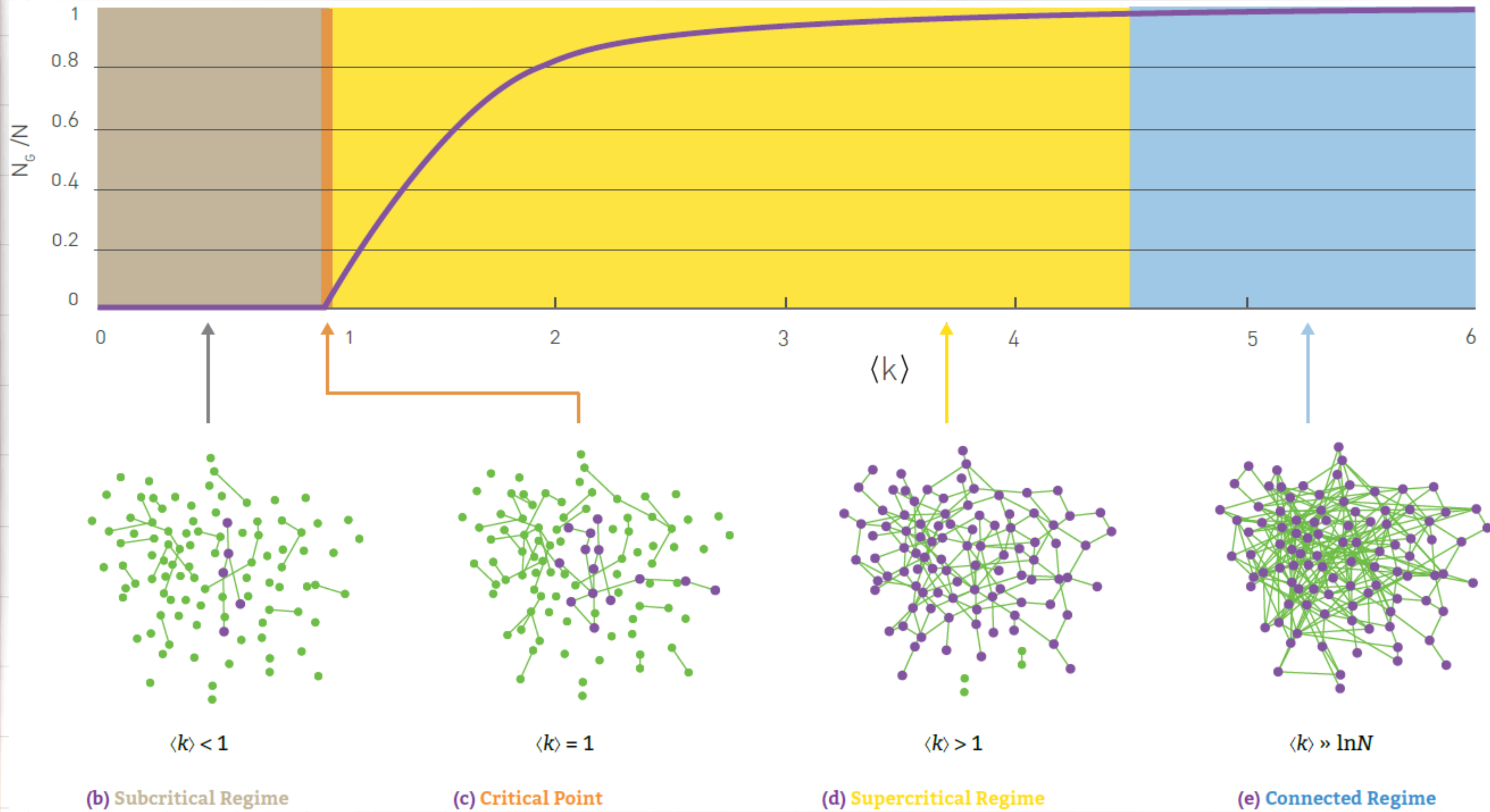


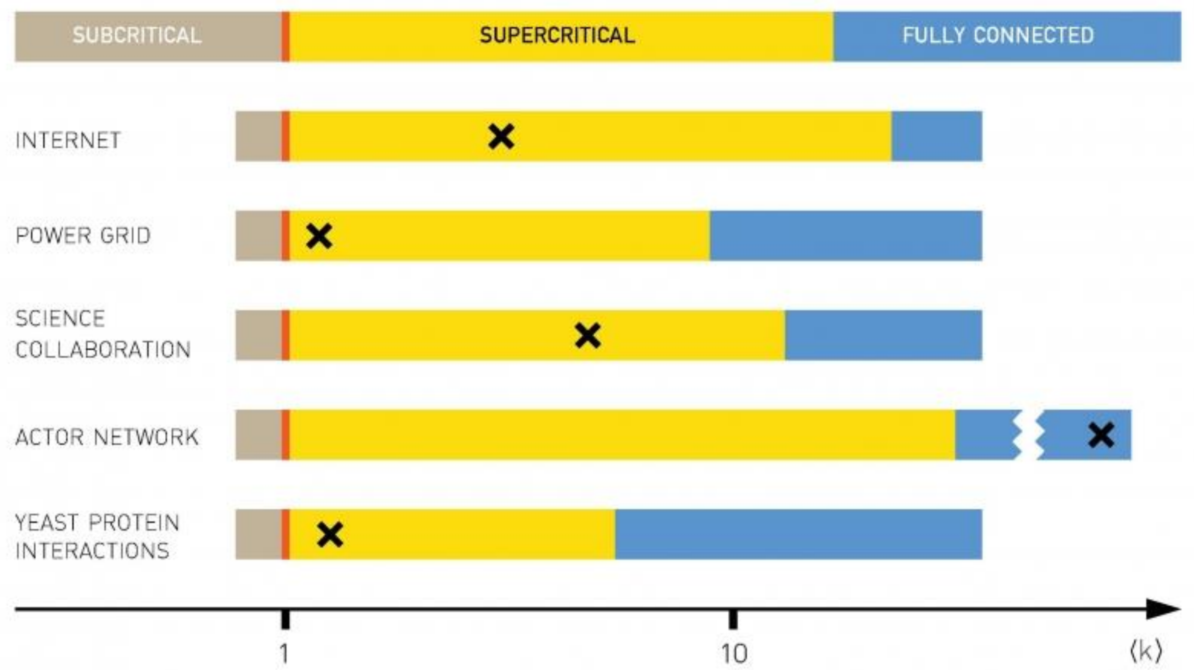
- Unikalny największy komponent: $N_G \sim (p - p_c)N = pN-1$
- Pozostałe komponenty to zazwyczaj drzewa, natomiast największy komponent zazwyczaj ma cykle



- Jeden największy komponent: $N_G = N_G = N$
- Brak izolowanych wierzchołków lub komponentów

Czy sieci rzeczywiste są zgodne z modelem ewolucji sieci losowej?





Sieć	N	L	<k>	ln_N
Internet	192,244	609,066	6.34	12.17
Sieć energetyczna	4,941	6,594	2.67	8.51
Współpraca naukowa	23,133	93,437	8.08	10.05
Sieć aktorów	702,388	29,397,908	83.71	13.46
Interakcje pomiędzy proteinami	2,018	2,930	2.90	7.61

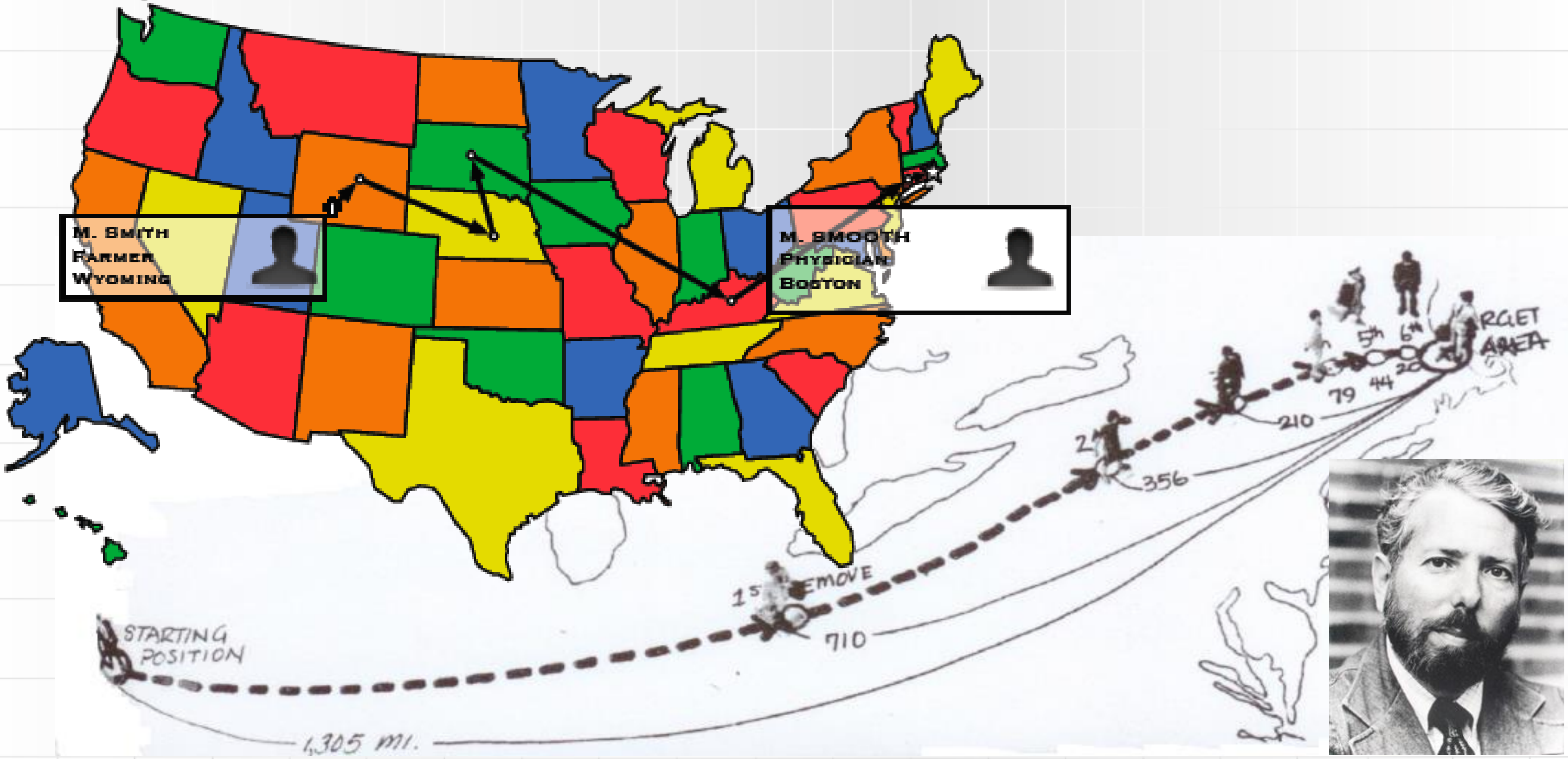
Dystans w sieciach losowych

Politechnika Wrocławska

J. Travers and S. Milgram. An experimental study of the small world problem. Sociometry, 32(4), 1969

Fenomen małego świata

- 1967, makler giełdowy mieszkający pod Bostonem
- 300 listów z Kansas i Nebraski, 64 dotarło, Średnia długość ścieżki: 5,2



Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). **An experimental study of search in global social networks**. *Science*, *301*(5634), 827-829.

Fenomen małego świata - powtórka

- 2003 Columbia University,
- 18 celów z 13 krajów,
- email,
- 98 847 (61 168) uczestników ze 166 krajów,
- 24 163 ścieżki
- 384 dotarło do celu
- $\langle L \rangle = 4.05$
- $5 \leq L_{*} \leq 7$

Politechnika Wrocławska

Leskovec J., Horvitz E.: Planetary-scale views on a large instant-messaging network. WWW 2008, 915-924

Fenomen małego świata - powtórka

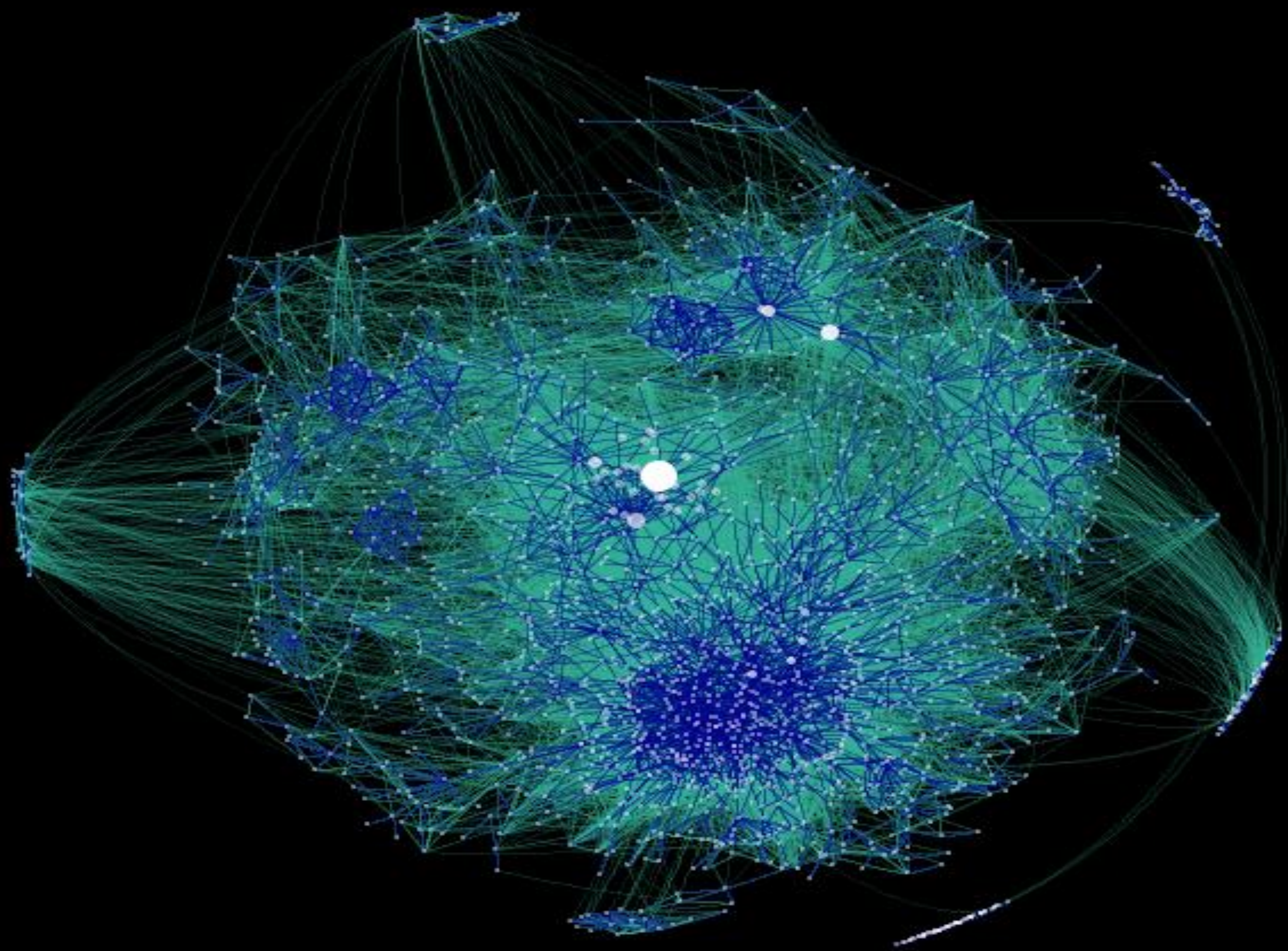
- Jure Leskovec (Carnegie Mellon University, teraz Stanford), Eric Horvitz (Microsoft Research)
- Microsoft Instant Messenger (taki dzisiejszy WhatsApp)
- Czerwiec 2006, 30 miliardów konwersacji, 240 miliony ludzi
- Sieć N=180 000 000, E = 1 300 000 000
- <L> = 6,6 „*We investigate [...] report that people are separated by "six degrees of separation" and find that the average path length among Messenger users is 6.6"*

Fenomen małego świata – Facebook i Twitter

- Facebook
  - Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012, June). **Four degrees of separation**. *4th Annual ACM Web Science Conference* (pp. 33-42).
  - 721 milionów kont, 69 miliardy „przyjaźni”
  - <L> = 4,74
- Twitter
  - Bakhshandeh, R., Samadi, M., Azimifar, Z., & Schaeffer, J. (2011, July). Degrees of separation in social networks. In *Fourth Annual Symposium on Combinatorial Search*.
  - 1 500 losowych par na Twitterze
  - <L> = 3,43

Albert, R., Jeong, H., & Barabási, A. L. (1999). Diameter of the world-wide web. *nature*, *401*(6749), 130-131.

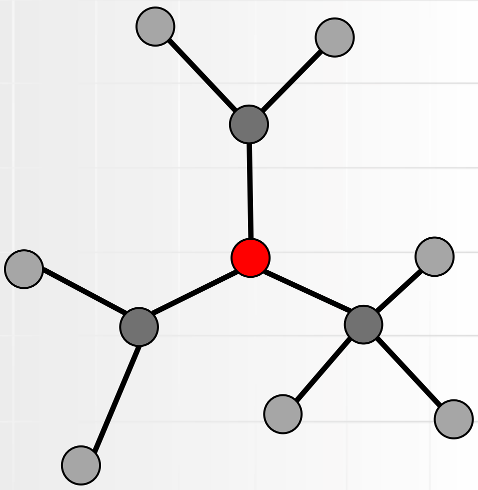
WWW: 8 degrees of separation



http://networksciencebook.com/chapter/3#small-worlds

Dystans w sieciach losowych

- Grafy losowe mają zwykle topologię podobną do drzewa z prawie stałymi stopniami węzłów.
  - <k> wierzchołków w odległości jeden d=1
  - $<k>^2$ wierzchołków w odległości dwa d=2
  - …
  - $<k>^n$ wierzchołków w odległości *n* d=n



$$d_{max} = \frac{lnN}{ln < k >}$$

Politechnika Wrocławska

Dystans w sieciach losowych

$$d_{max} = \frac{lnN}{ln < k >}$$

- Dla większości sieci to równanie dokładniej wyznacza oczekiwaną średnią długość najkrótszych ścieżek w sieci.

$$< d >= \frac{lnN}{ln < k >}$$

- Zjawisko małego świata to właściwość polegająca na tym, że średnia długość ścieżki lub średnica sieci zależy logarytmicznie od rozmiaru systemu. Stąd „mały” oznacza, że ⟨d⟩ jest proporcjonalne do ln N, a nie N.
- Wyrażenie $\frac{1}{ln<k>}$ oznacza, że im gęstsza sieć, tym mniejsza będzie odległość między węzłami.

Dystans w sieciach losowych

Sieć	N	L	<k>	<d>	d_{max}	lnN/ln‹k›
Internet	192,244	609,066	6.34	**6.98**	26	**6.58**
WWW	325,729	1,497,134	4.60	**11.27**	93	**8.31**
Sieć energetyczna	4,941	6,594	2.67	**18.99**	46	**8.66**
Rozmowy telefoniczne	36,595	91,826	2.51	**11.72**	39	**11.42**
Emaile	57,194	103,731	1.81	**5.88**	18	**18.4**
Współpraca naukowa	23,133	93,437	8.08	**5.35**	15	**4.81**
Sieć aktorów	702,388	29,397,908	83.71	**3.91**	14	**3.04**
Sieć cytowań	449,673	4,707,958	10.43	**11.21**	42	**5.55**
Metabolizm Ecoli	1,039	5,802	5.58	**2.98**	8	**4.04**
Interakcje pomiędzy proteinami	2,018	2,930	2.90	**5.61**	14	**7.14**

Politechnika Wrocławska

Współczynnik grupowania w sieciach losowych

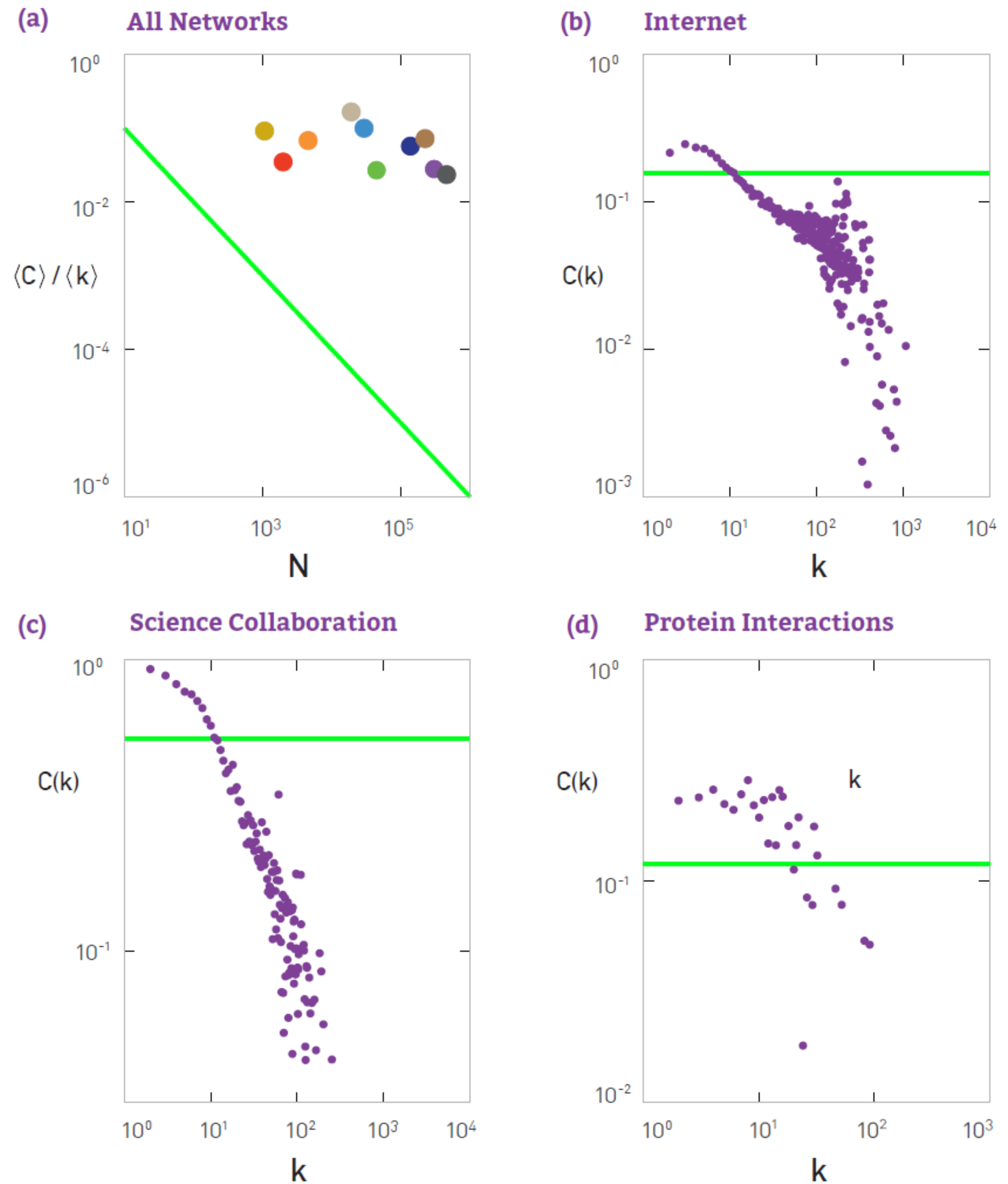
HR EXCELLENCE IN RESEARCH

Współczynnik grupowania

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \qquad \langle L_i \rangle = p\frac{k_i(k_i - 1)}{2}$$

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}$$

Wspó



(a)
All Networks
10^0
10^{-2}
10^{-4}
10^{-6}
⟨C⟩ / ⟨k⟩
10^1
10^3
10^5
N
(b)
Internet
10^0
10^{-1}
10^{-2}
10^{-3}
C(k)
10^0
10^1
10^2
10^3
10^4
k
(c)
Science Collaboration
10^0
10^{-1}
C(k)
10^0
10^1
10^2
10^3
10^4
k
(d)
Protein Interactions
10^0
10^{-1}
10^{-2}
C(k)
k
10^0
10^1
10^2
10^3
k

Sieci rzeczywiste a grafy losowe

Politechnika Wrocławska

HR EXCELLENCE IN RESEARCH

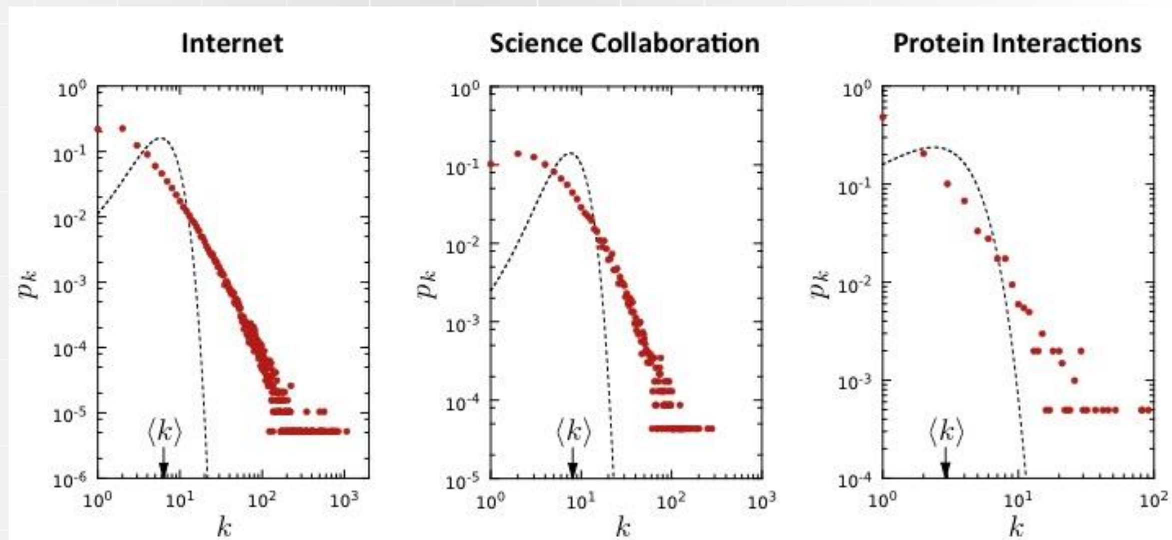
Politechnika Wrocławska

Rozkład stopni węzła

- Powinno być

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

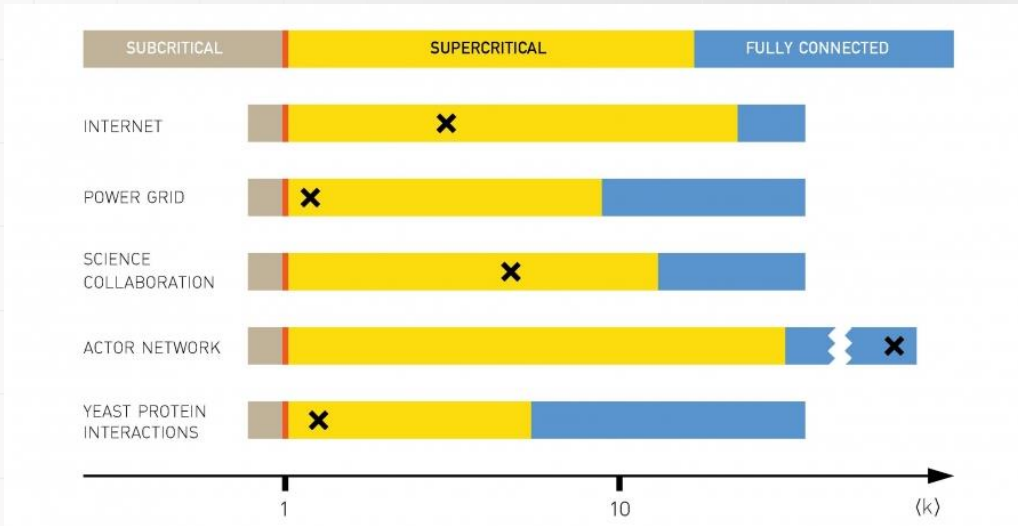
$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$



- Jest

$$P(k) \approx k^{-\gamma}$$

Połączeniowość



Sieć	N	L	<k>	ln_N
Internet	192,244	609,066	6.34	12.17
Sieć energetyczna	4,941	6,594	2.67	8.51
Współpraca naukowa	23,133	93,437	8.08	10.05
Sieć aktorów	702,388	29,397,908	83.71	13.46
Interakcje pomiędzy proteinami	2,018	2,930	2.90	7.61

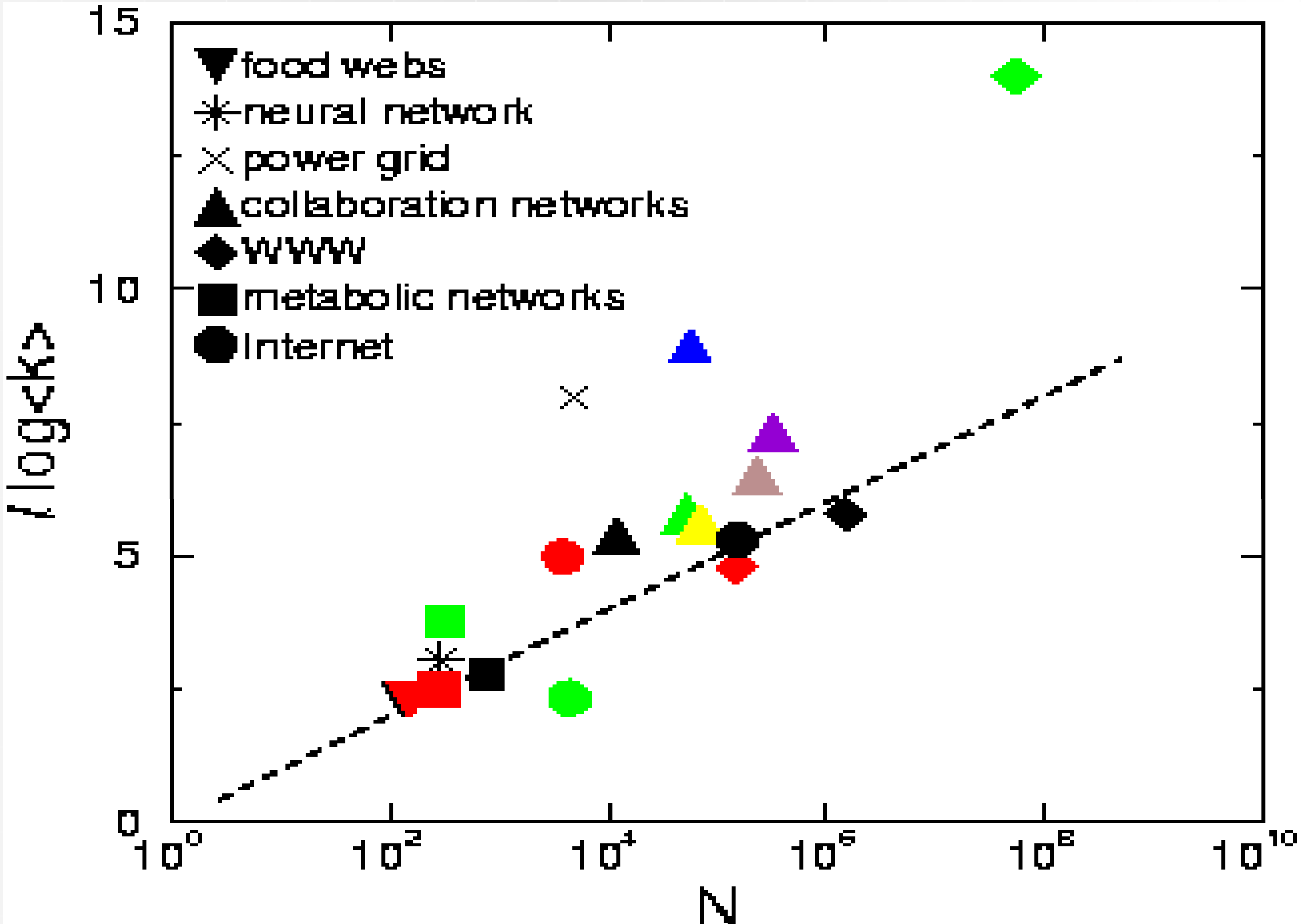
Dystans

- Powinno być

$$< d >= \frac{lnN}{ln < k >}$$

- Jest

$$< d >\approx \frac{lnN}{ln < k >}$$

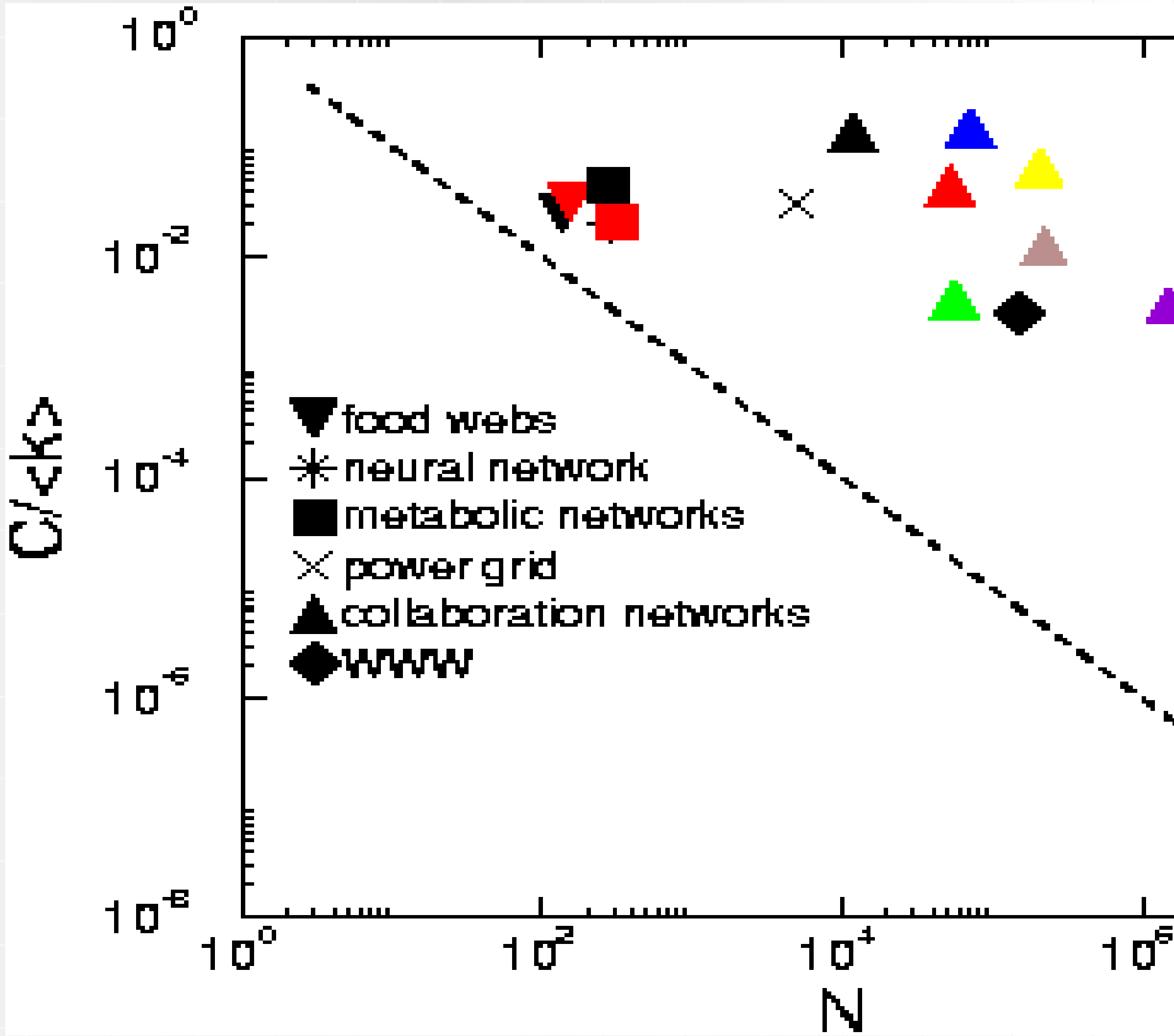


Współczynnik grupowania

- Powinno być

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

- Jest
  - C potrafi być kilka rzędów większe
  - C zależy od k
  - dla sieci o podobnym <k>, C nie zależy od rozmiaru sieci.



Czy prawdziwe sieci są grafami losowymi?

- Gdy 20 lat temu zaczęły pojawiać się dane na temat dużych sieci okazało się że nie.
- Rozkład stopni węzła 
- Połączeniowość 
- Dystans 
- Współczynnik grupowania 

Jeśli jest to błędny model to dlaczego się o nim uczyliśmy?

- Jest to model referencyjny (tzw. null model)
- Pomaga obliczyć wiele wielkości, które można następnie porównać z rzeczywistymi danymi, rozumiejąc, w jakim stopniu dana właściwość jest wynikiem jakiegoś losowego procesu.
  - np. pomagają zidentyfikować cechy sieci rzeczywistych, które są współdzielone przez dużą liczbę rzeczywistych sieci, ale odbiegają od przewidywań losowego modelu sieci.
  - Aby je zidentyfikować, musimy zrozumieć, jak wyglądałaby konkretna właściwość, gdyby była napędzana wyłącznie przez procesy losowe.
- **Chociaż model jest błędny dalej jest przydatny**

Politechnika
Wrocławska

Model Watts i Strogatza

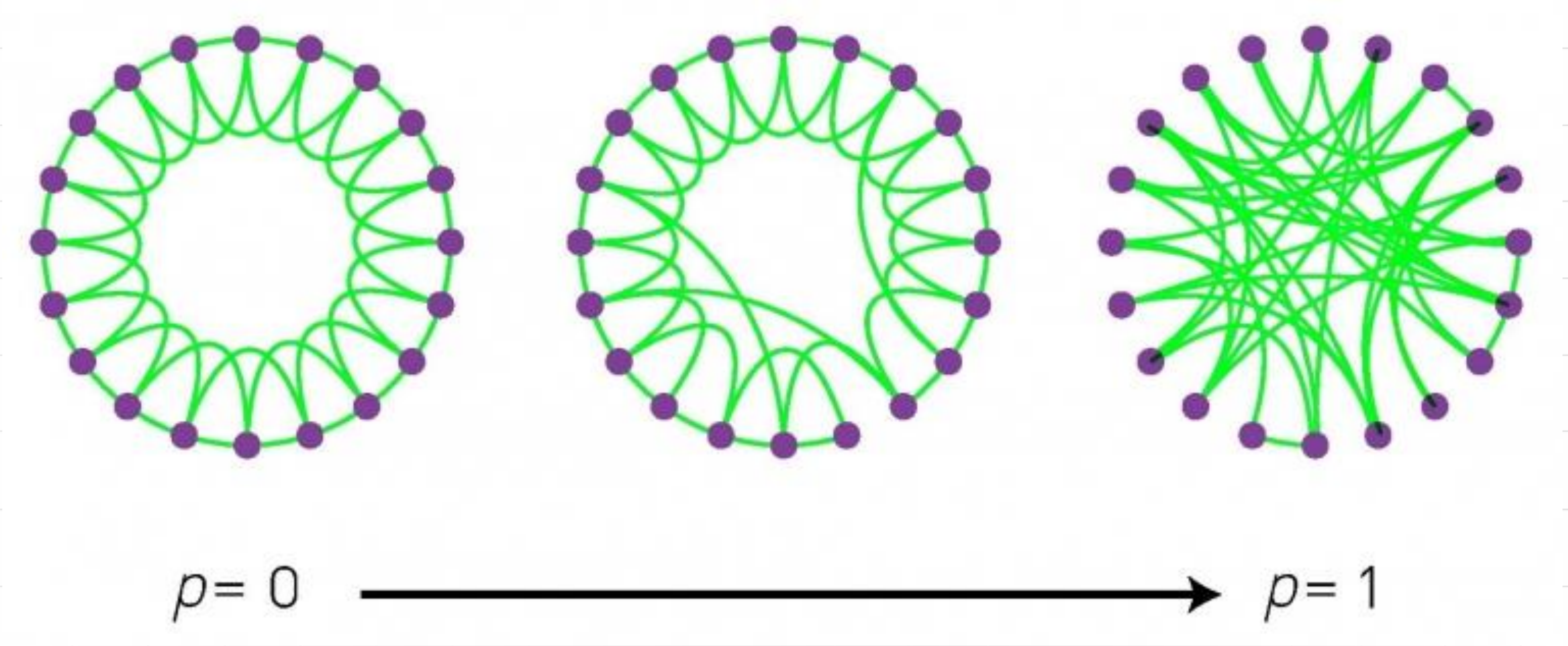
HR EXCELLENCE IN RESEARCH

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’networks. N*ature*, *393*(6684), 440-442.

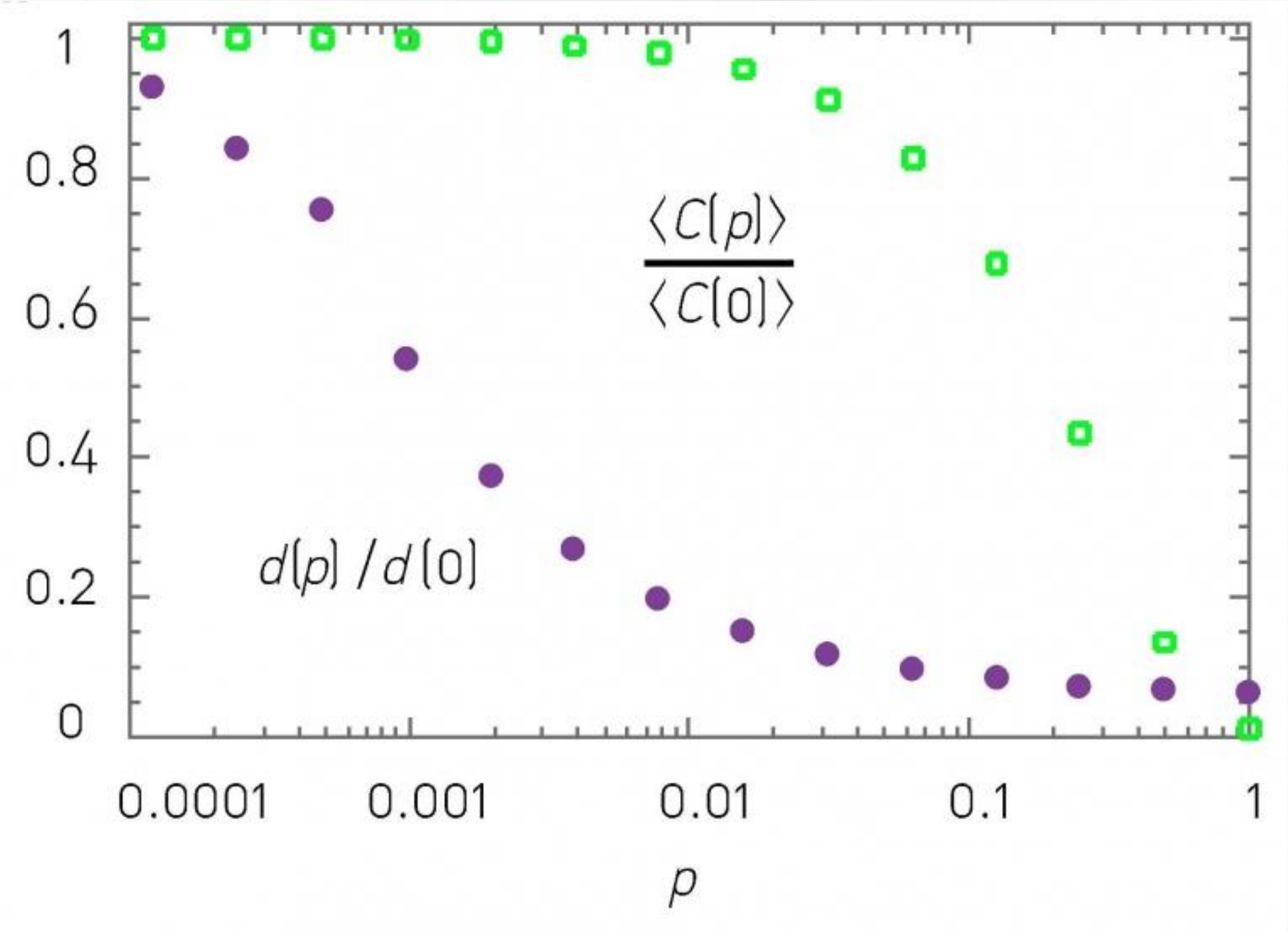
Model Watts i Strogatza

- Watts-Strogatz Model (Small-World model)
- Analiza sieci rzeczywistych
- Dwie kluczowe obserwacje
  - **Właściwość małego świata** (niewielka średnica sieci, krótkie dystanse pomiędzy wierzchołkami)
  - **Bardzo duże zgrupowanie -** wysoki średni współczynnik grupowania w sieciach (o kilka rzędów wyższy niż w sieciach losowych)

Model Watts i Strogatza



Model Watts i Strogatza



Czy prawdziwe sieci są sieciami WS?

- Rozkład stopni węzła
- Połączeniowość
- Dystans
- Współczynnik grupowania

Politechnika Wrocławska

Model Barabási-Albert

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

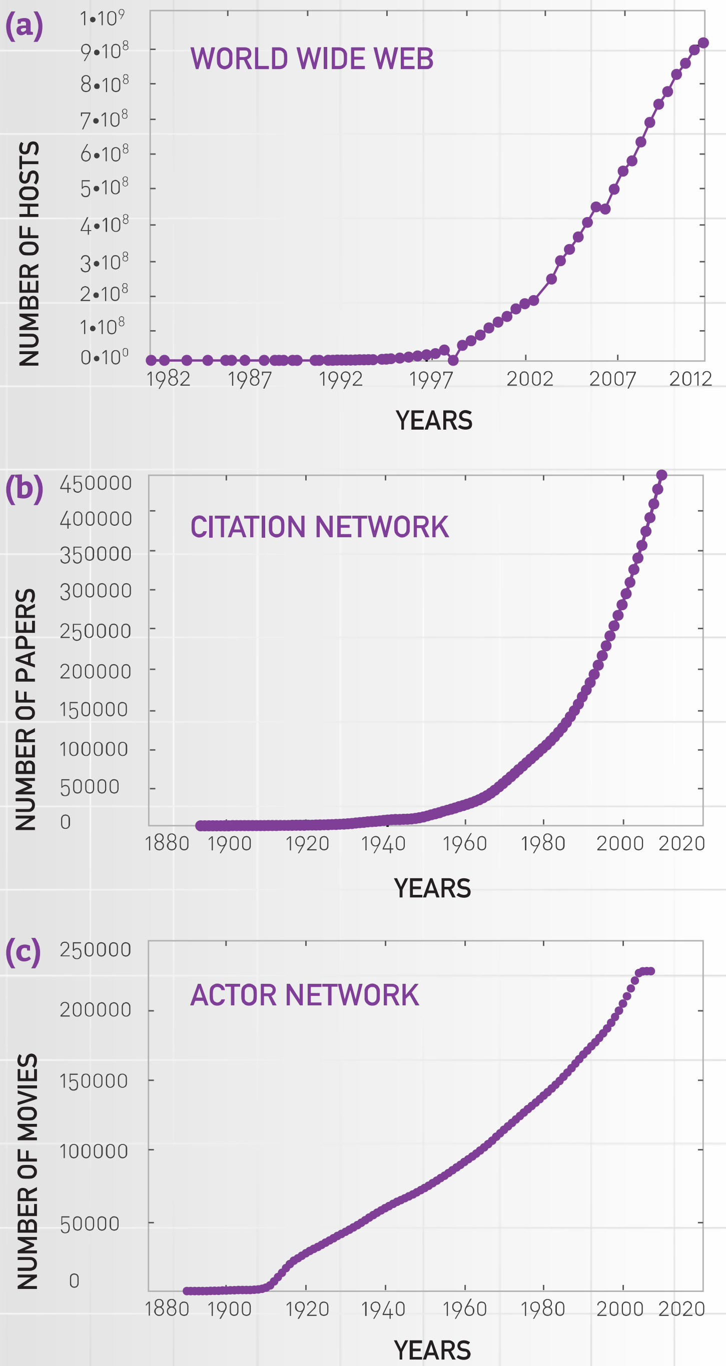
HR EXCELLENCE IN RESEARCH

Huby

- Huby reprezentują najbardziej uderzającą różnicę między siecią losową a siecią bez skalowania. Ich pojawienie się w wielu rzeczywistych systemach rodzi fundamentalne pytania:
  - Dlaczego model sieci losowej Erdősa i Rényi nie przewiduje hubów ani wielu innych cech obserwowanych w wielu rzeczywistych sieciach?
  - Dlaczego tak różne systemy, takie jak WWW, sieć energetyczna lub interakcje pomiędzy białkami, tworzą podobną bezskalową (scale-free) topologię sieci?

Wzrost (Growth)

- Sieć losowa – liczba węzłów (N) jest statyczna.
- Sieci rzeczywiste ewoluują i rosną wraz z pojawianiem się nowych wierzchołków i krawędzi.



Politechnika Wrocławska

Preferencyjne łączenie (Preferential attachment)

- Sieć losowa – krawędzie pomiędzy wierzchołkami są dodawane w losowy sposób

- W sieciach rzeczywistych wierzchołki często „preferują” tworzenie połączeń z istotnymi węzłami w sieci.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

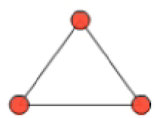
Wzrost i preferencyjne łączenie

1. Sieci stale się rozwijają dzięki dodawaniu nowych węzłów – w każdym kroku dodaj *m* nowych wierzchołków
2. Nowe węzły wolą łączyć się z silnie połączonymi węzłami.

$$\Pi(k_i) = \frac{k_i}{\Sigma_j k_j}$$

Przykład sieć WWW

1. Nowe strony
2. Linki do istniejących znanych stron

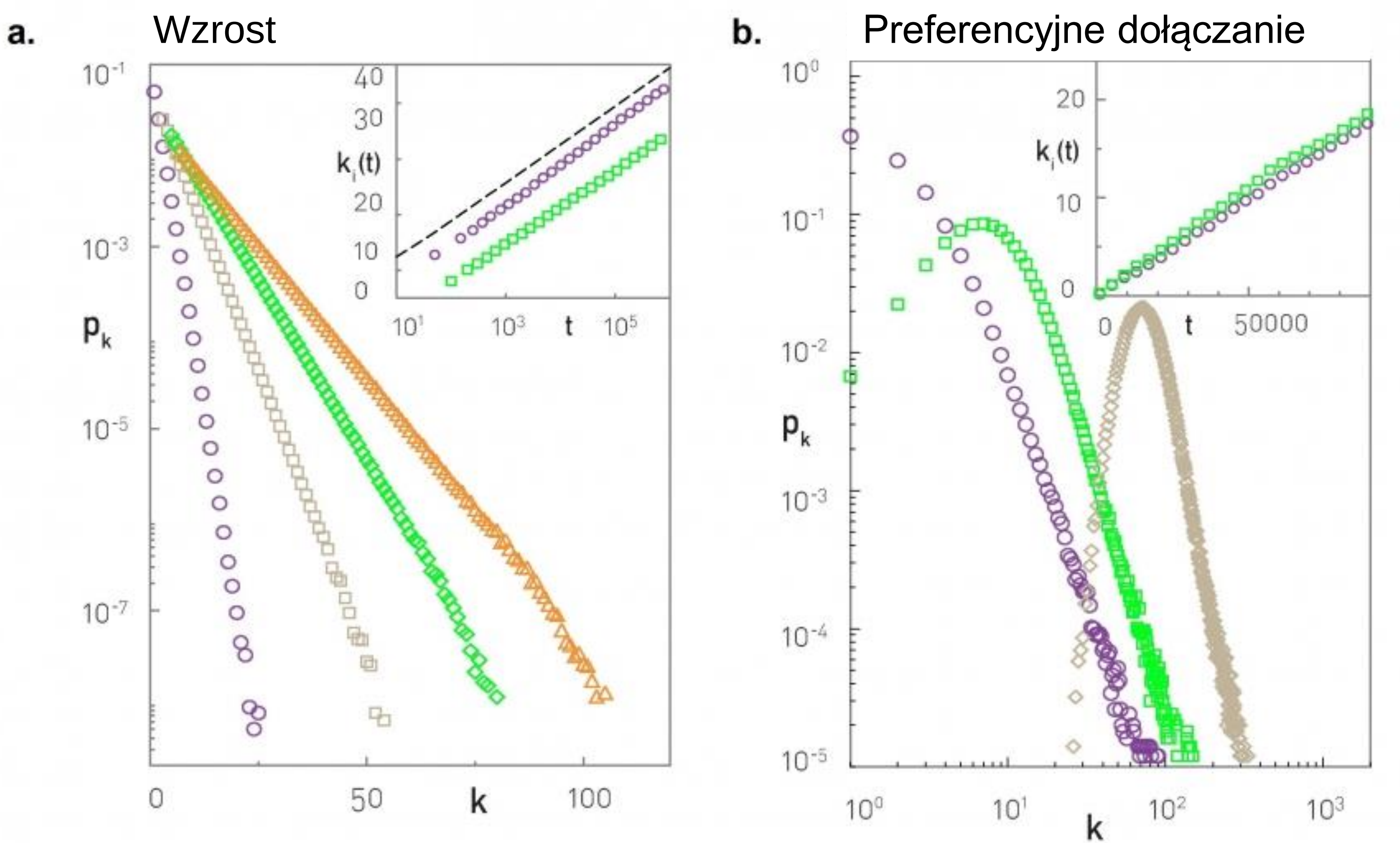


Politechnika Wrocławska

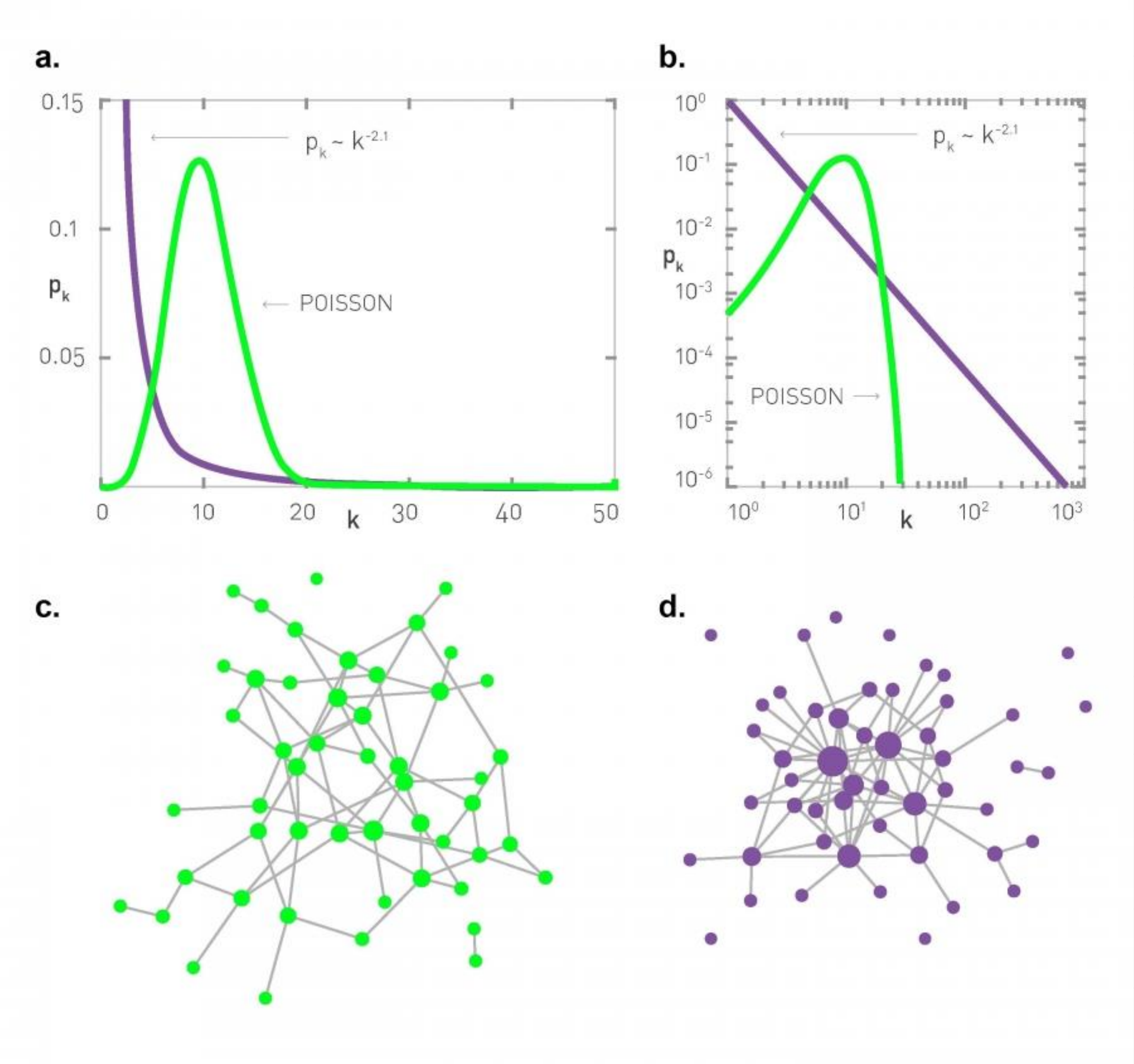
Model Barabási-Albert

- m_0 - początkowa liczba wierzchołków tworzących graf pełny
- m – liczba krawędzi dodawanych wraz każdym nowym wierzchołkiem
- t – liczba dodawanych wierzchołków

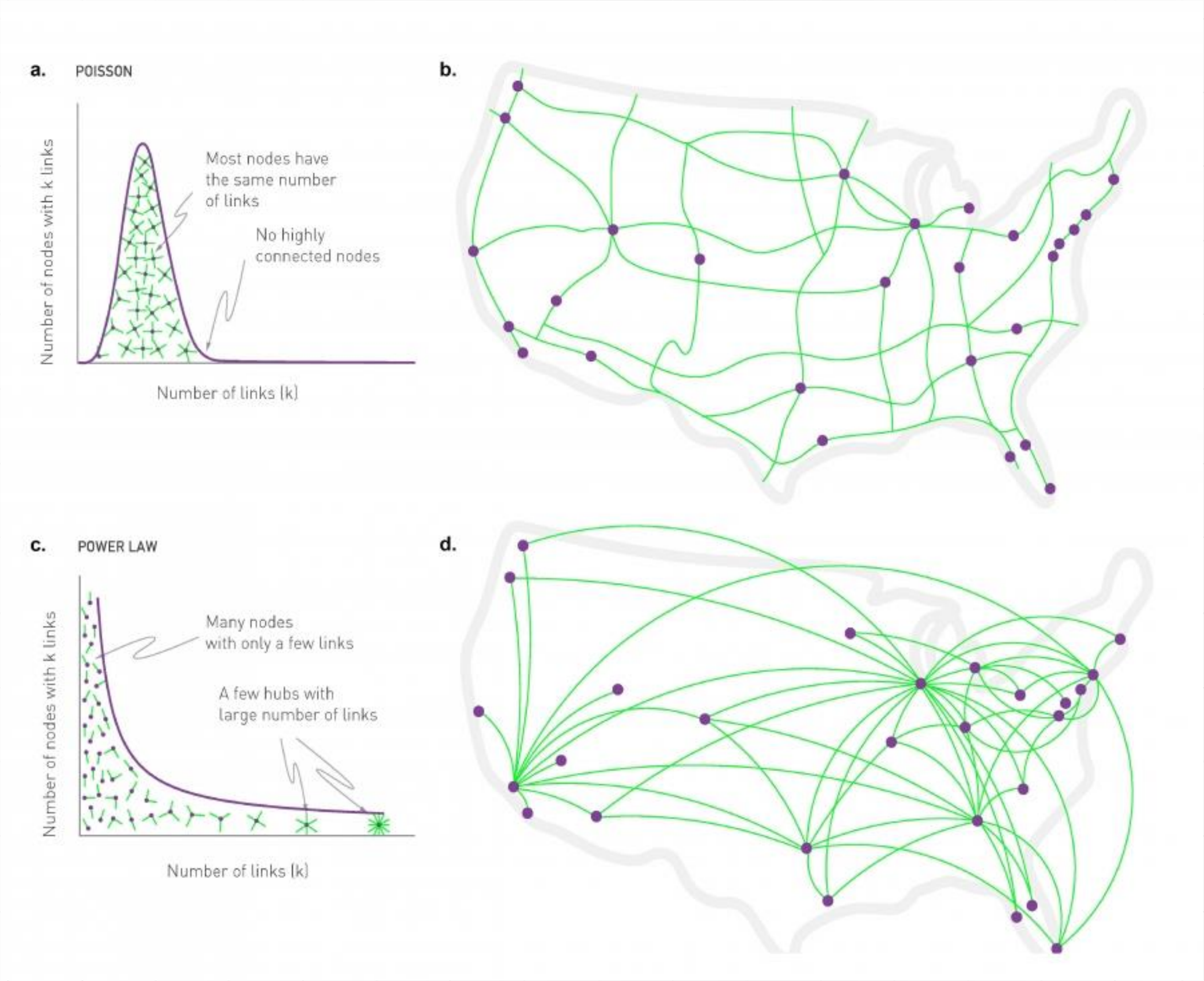
Brak wzrostu lub preferencyjnego dołączania



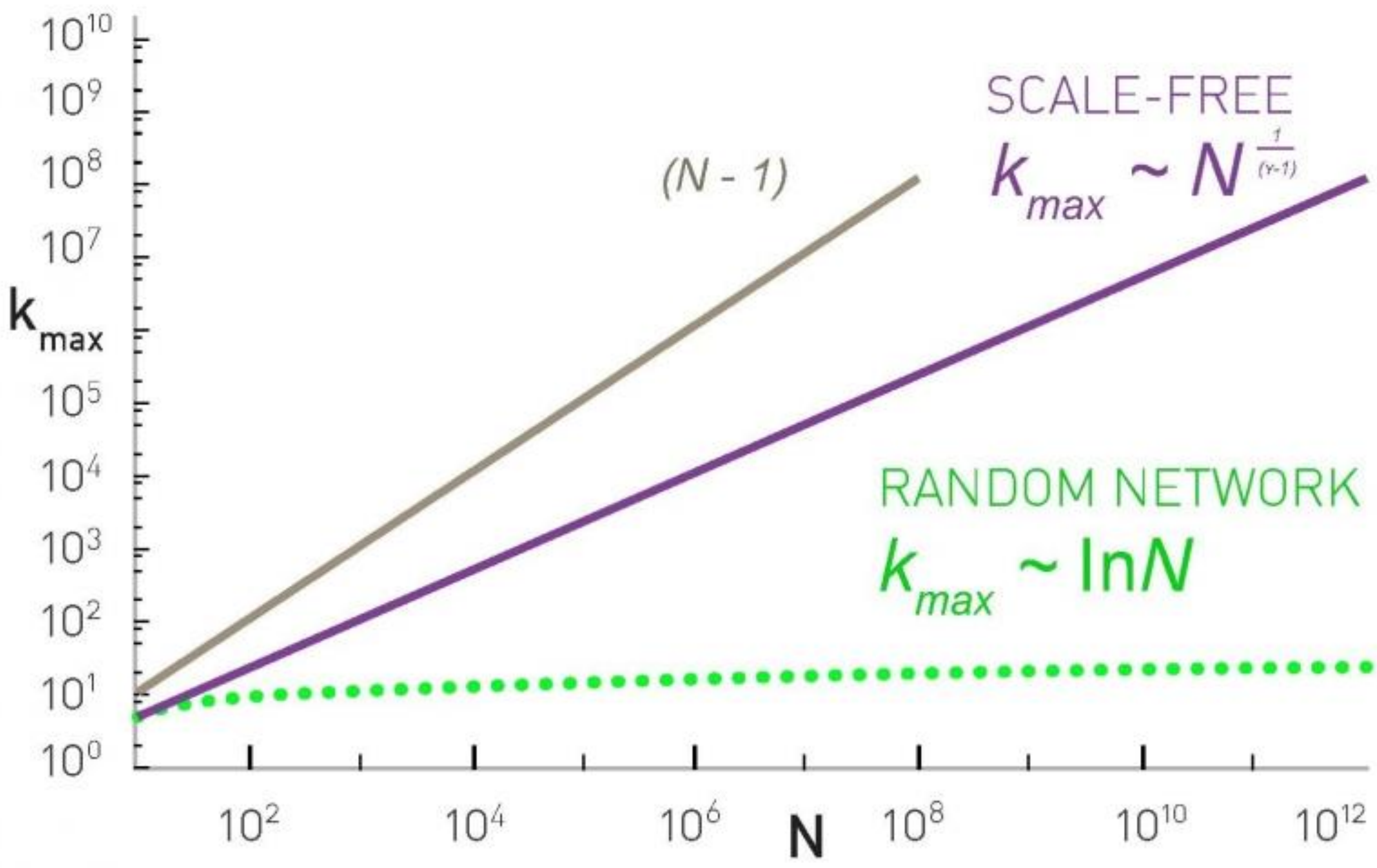
Degree distribution



Sieć losowa a sieć bezskalowa (BA)

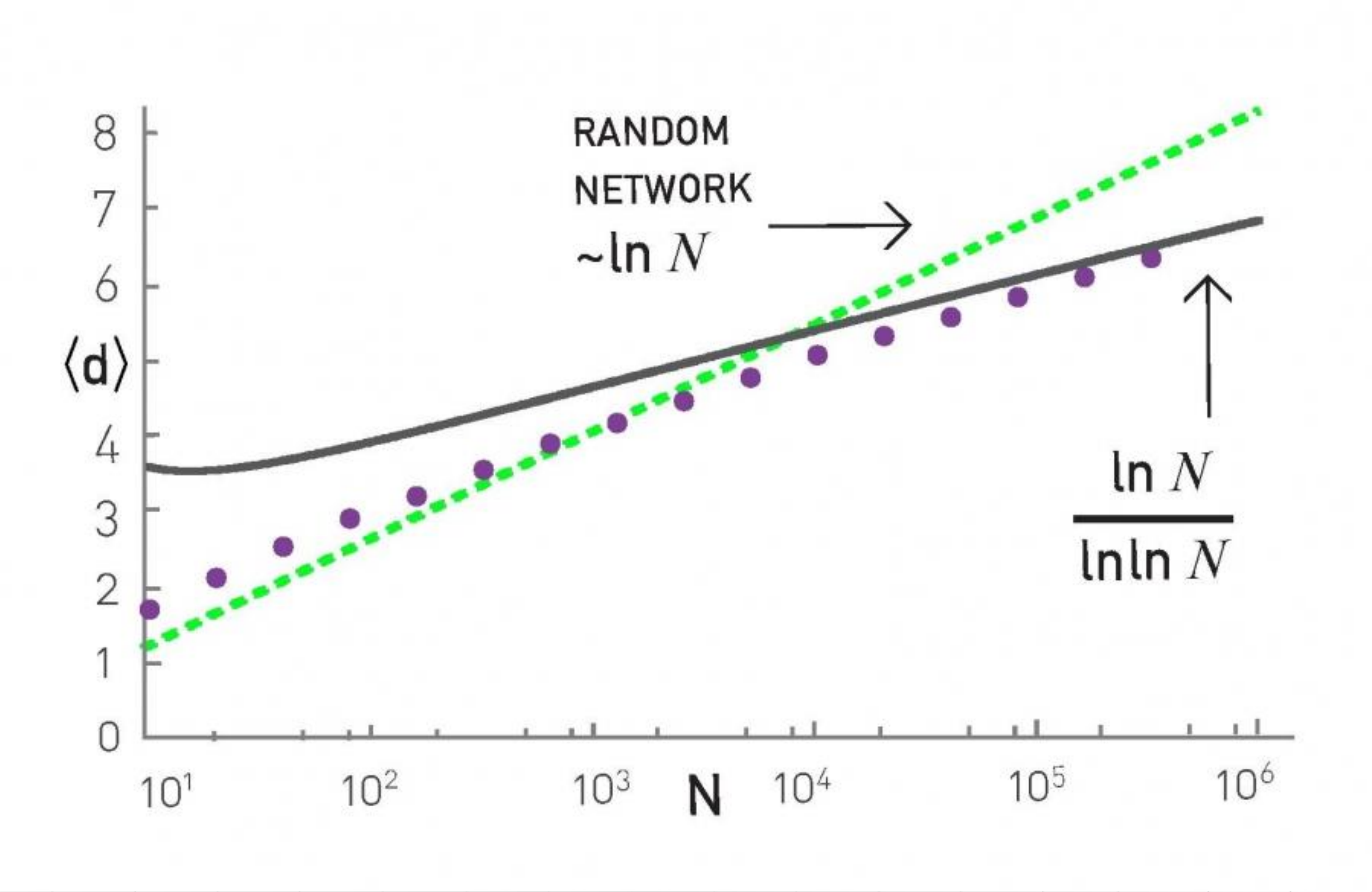


Huby – stopień węzła



Bollobás, B., & Riordan, O. (2004). The diameter of a scale-free random graph. *Combinatorica*, *24*(1), 5-34.

Średnica i dystans



Klemm, K., & Eguiluz, V. M. (2002). Growing scale-free networks with small-world behavior. *Physical Review E*, *65*(5), 057102.

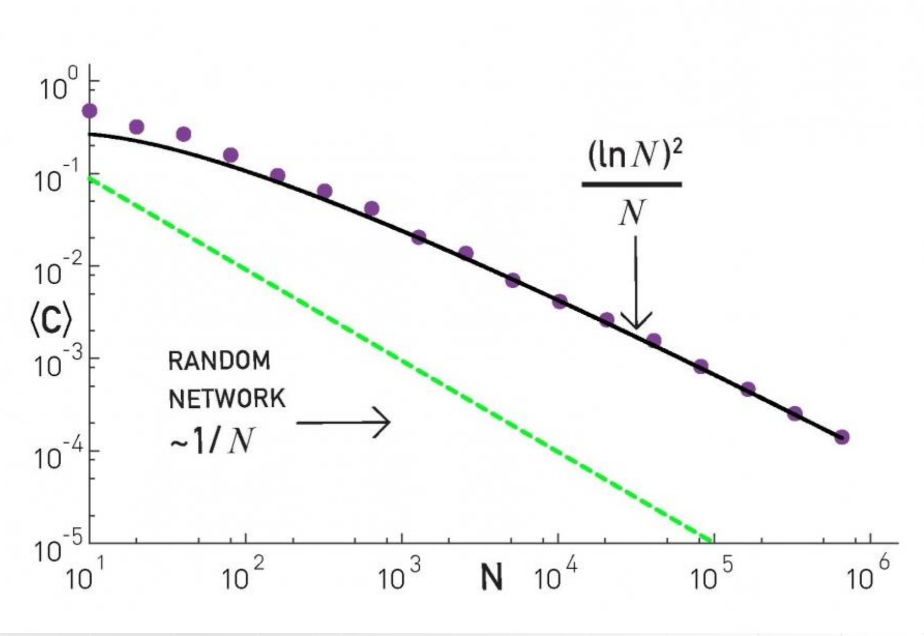
Współczynnika grupowania

- W sieciach losowych

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

- W sieciach BA

$$C = \frac{m}{8} \frac{(\ln N)^2}{N}$$



Politechnika Wrocławska

Model Barabási-Albert

- Liczba wierzchołków
  - $N = m_0 + t \approx t$ (dla dużych t)
- Liczba krawędzi
  - $L = m_0(m_0-1)/2 + mt \approx mt$ (dla dużych t)
- Średni stopień węzła
  - $\langle k \rangle = (m_0(m_0-1) + 2mt)/(m_0+t) \approx 2m$ (dla dużych t)
- Rozkład stopni węzła
  - $p_k \approx k^{-\gamma}$, $2<\gamma<3$
- Średnia długość ścieżki
  - $\langle d \rangle \approx \ln N/\ln(\ln N)$
- Średni współczynnik grupowania
  - $\langle C \rangle \approx (\ln N)^2/N$

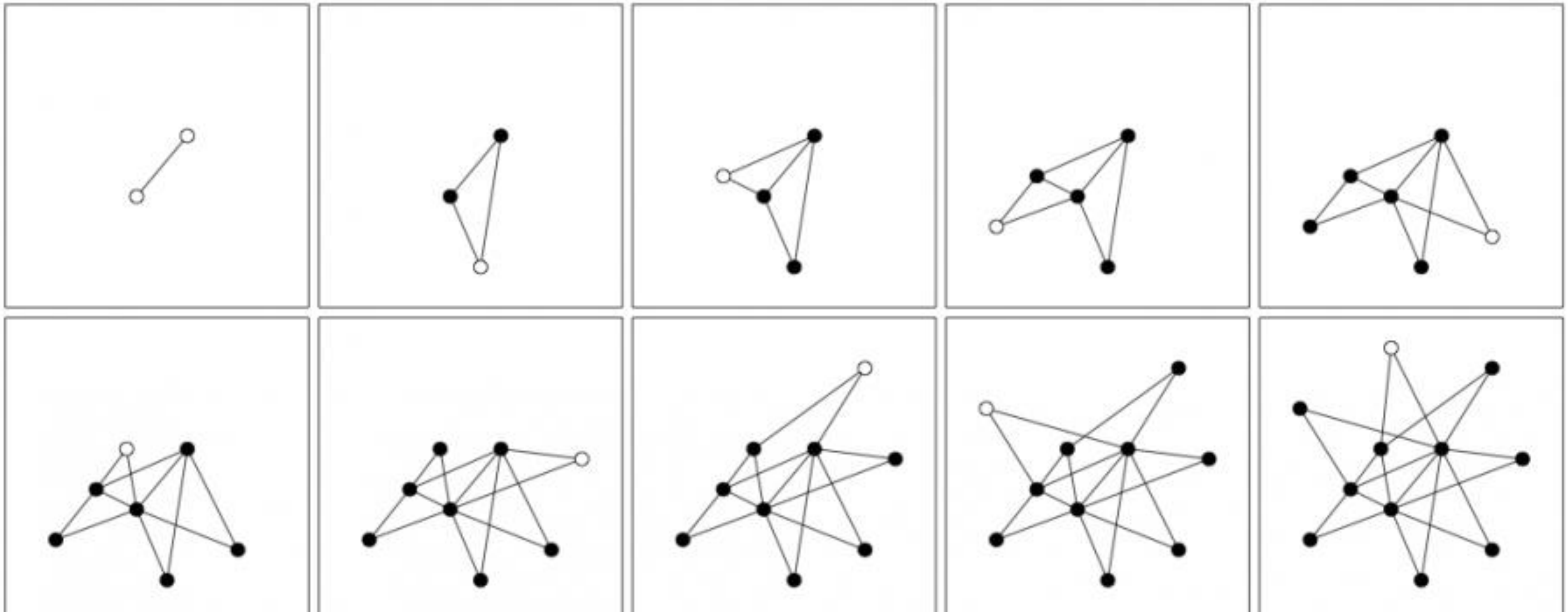
Politechnika Wrocławska

Model Barabási-Albert

- $m_0 = 2$
- $m = 2$
- $t = 9$

- $N = m_0 + t = 11$
- $L = m_0(m_0 - 1)/2 + mt = 19$

- $\langle k \rangle = (m_0(m_0-1) + 2mt)/(m_0+t) = 3{,}45$



Model Barabási-Albert

- $m_0 = 2$
- $m = 2$
- $t = 1000$
- $N = m_0+t = 1002 \approx t$
- $L= m_0(m_0-1)/2+mt = 2001 \approx mt$
- $\langle k \rangle = (m_0(m_0-1) + 2mt)/(m_0+t) = 3,99 \approx 2m$

Model BA nie jest w stanie opisać wielu cech rzeczywistych systemów

- Model przewiduje $\gamma=3$, podczas gdy wykładnik stopnia rzeczywistych sieci waha się od 2 do 5
- Wiele sieci, takich jak WWW czy sieci cytowań, jest skierowanych, podczas gdy model generuje sieci nieskierowane.
- Wiele procesów obserwowanych w sieciach, od łączenia się z już istniejącymi węzłami do zanikania krawędzi i znikania węzłów, jest nieobecnych w modelu.
- Model nie pozwala nam na rozróżnienie węzłów na podstawie pewnych wewnętrznych cech, takich jak nowość artykułu naukowego lub użyteczność strony internetowej.

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, *10*(1), 1-10.

Scale-free networks are rare

- Centralnym twierdzeniem we współczesnej nauce o sieciach jest to, że sieci w świecie rzeczywistym są zazwyczaj „bezskalowe” (scale free), co oznacza, że ułamek węzłów o stopniu k jest zgodny z prawem potęgowym, rozpadając się jak $k^{-\gamma}$, często z $2<\gamma<3$.
- 1 000 sieci społecznych, biologicznych, technologicznych i informacyjnych
- Tylko 4% z nich jest bezskalowa (scale free)

Pytania

?