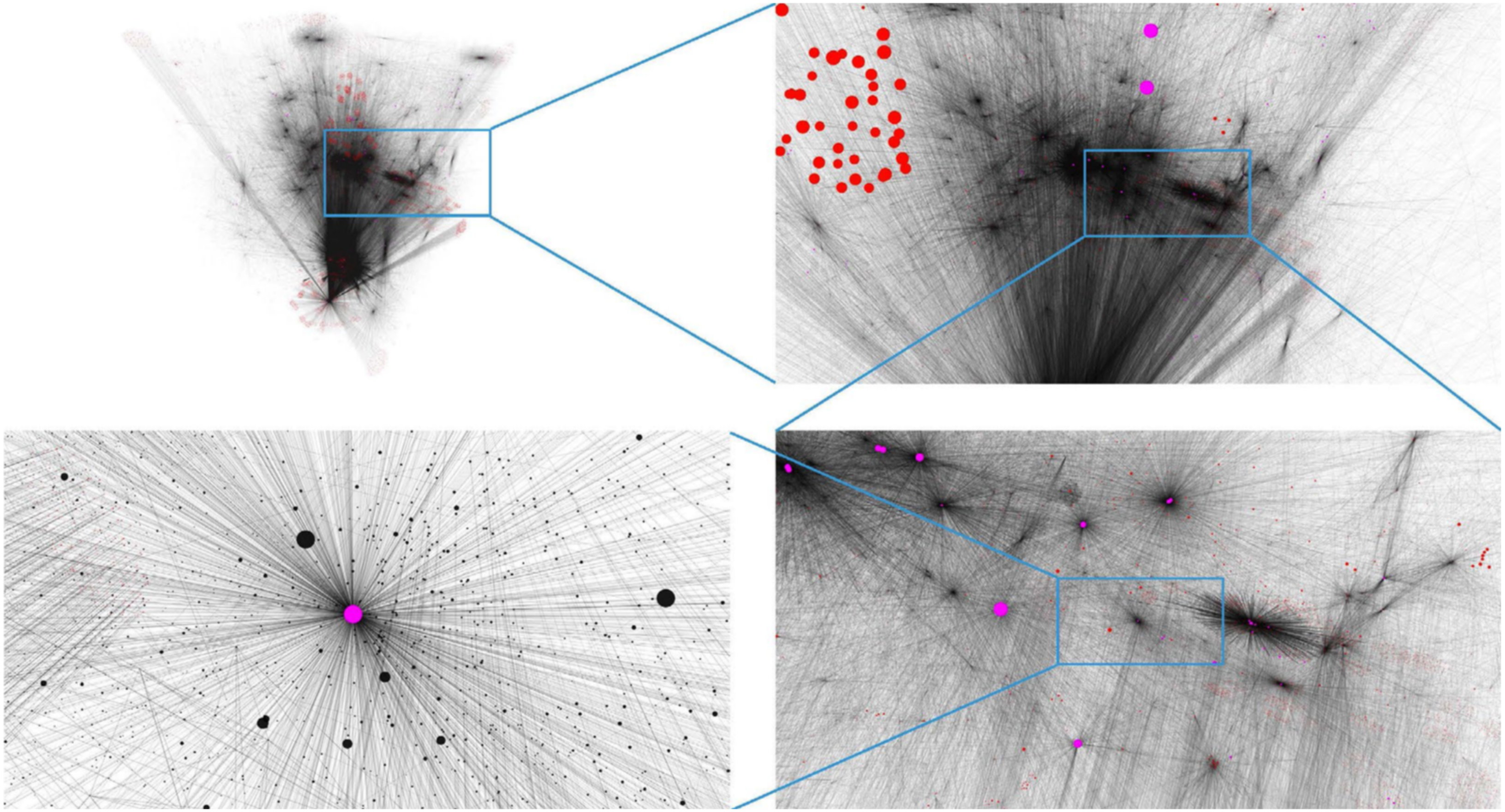
SD212

Scale Free Networks & Small world property

PART 1

Scale Free Networks

World Wide Web



Empirical degree distribution of a node

Graph $G(V, E)$.

For $u \in V, d_u = \text{degree of node } u$

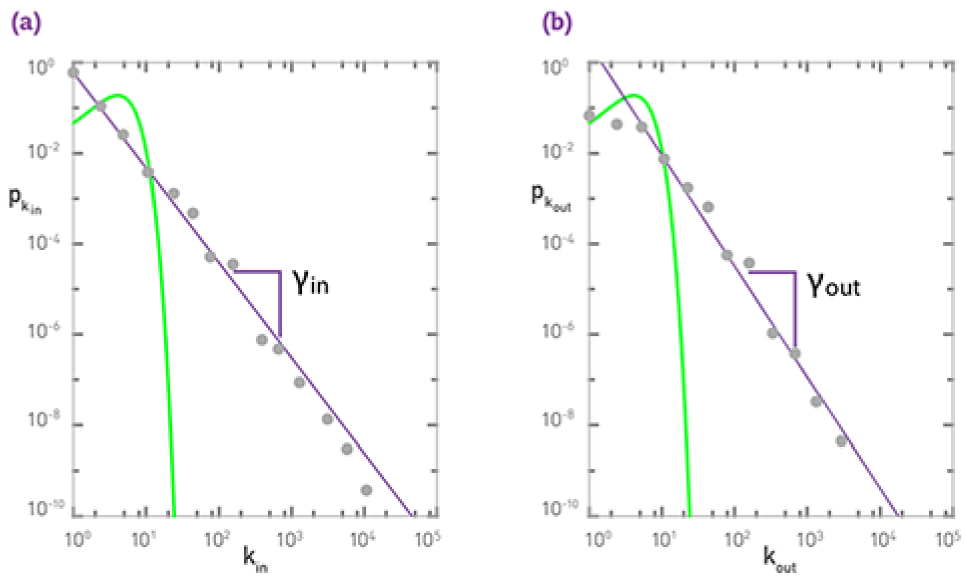
$$p_k = \frac{1}{n} \sum_{u \in V} 1_{\{d_u = k\}}$$

$$p_k = \text{fraction of nodes with degree } k$$

Degree distribution : log-log plot

Degree distribution of the WWW:

k_{in}: in degree / k_{out}: out degree



Scale-free property

For **scale-free networks** we have:

$$\log p_k \sim -\gamma \log k$$

$$p_k \sim k^{-\gamma}$$

The degree distribution follows a ***power law****.*

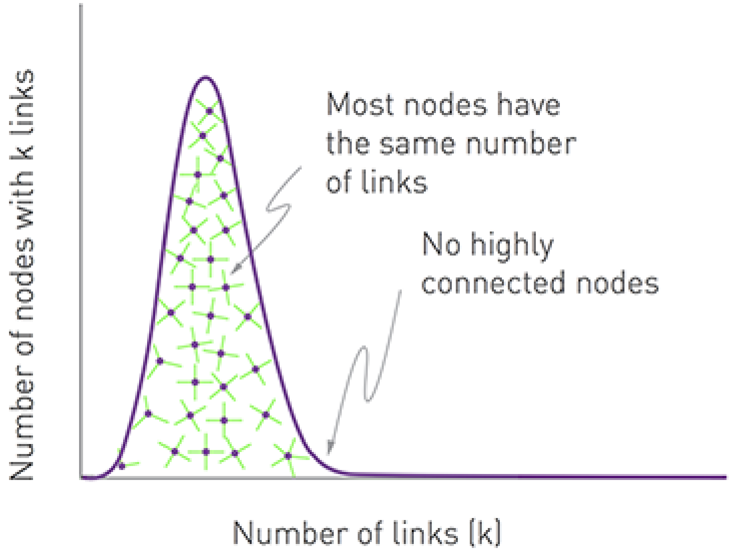
Power laws and scale invariance

$$p(x) = x^{-\gamma}$$

We have:

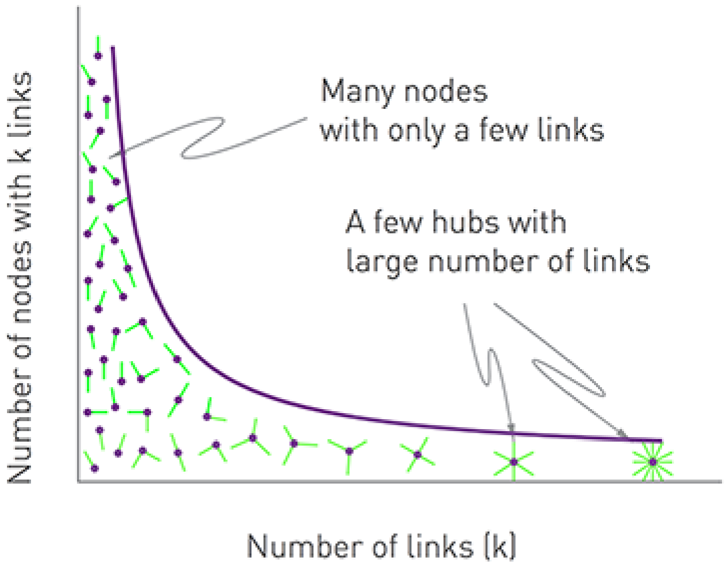
$$p(cx) = c^{-\gamma}x^{-\gamma} \propto p(x)$$

Bell curve vs. power law



Bell curve

Bell curve vs. power law



Power law curve

Example : Wikipedia

In [21]:
```
import networkx as nx
```

In [22]:
```
wikipedia_file = "data/wikipedia-humans/en-wikipedia.humans.ungraph.txt"
```

In [23]:
```
G = nx.read_edgelist(wikipedia_file,
                     comments="#", delimiter="\t", nodetype=int)
G.name = "Wikipedia (Humans)"
```

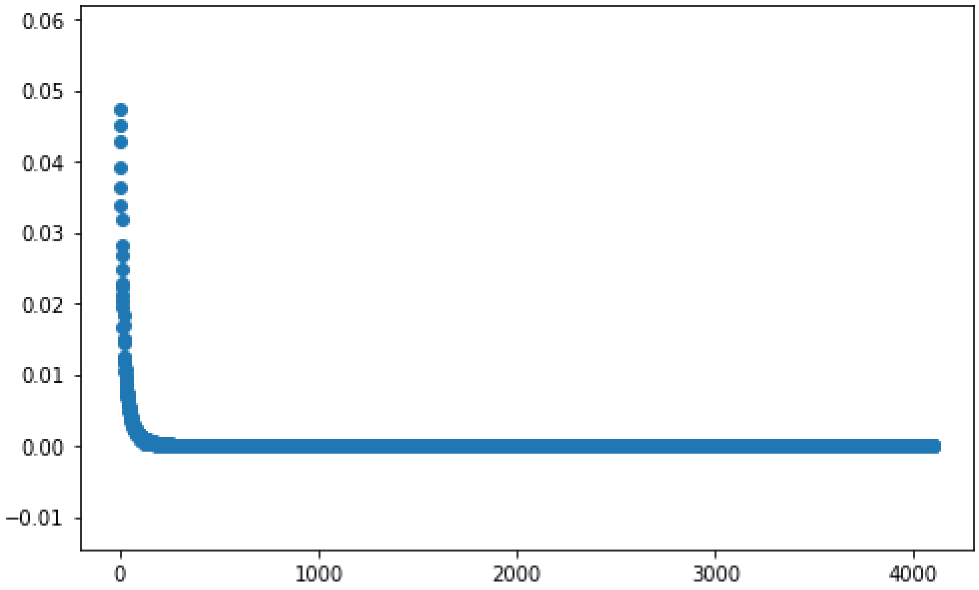
In [24]:
```
print(nx.info(G))
```

```
Name: Wikipedia (Humans)
Type: Graph
Number of nodes: 731293
Number of edges: 3266258
Average degree:   8.9328
```

Wikipedia - Degree distribution

Scatter plot with linear scale

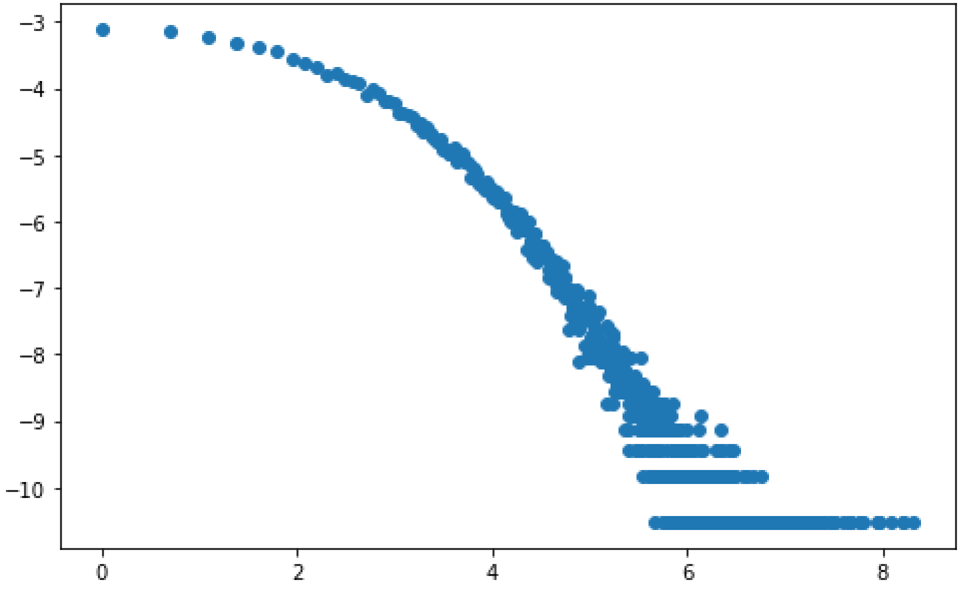
In [29]:
```
plt.figure(figsize=(8,5))
f = plt.scatter(np.arange(len(p)), p)
```



Wikipedia - Degree distribution

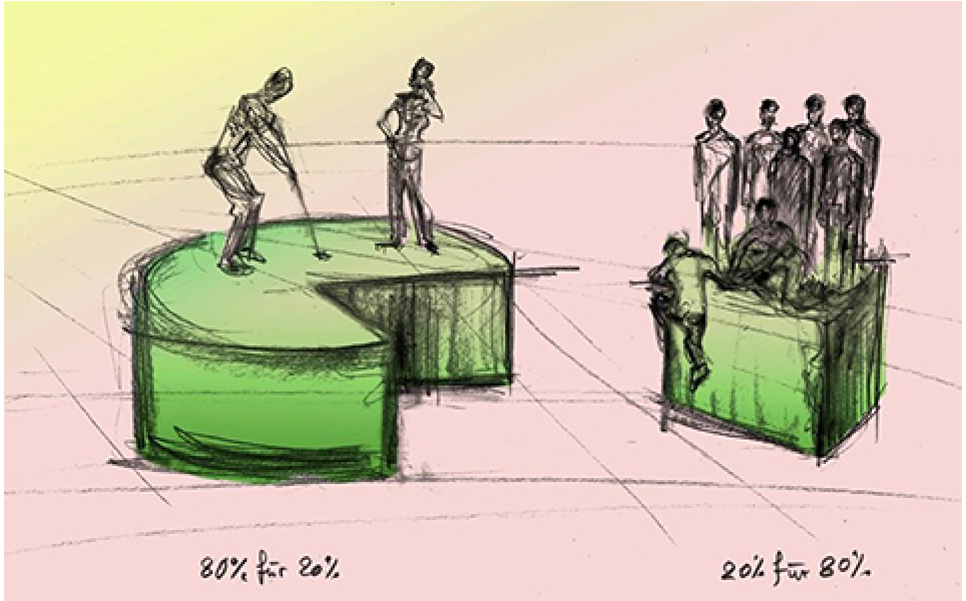
Scatter plot with log-log scale

In [30]:
```
x_values = [x for x in range(1, len(p)) if p[x] > 0]
y_values = [p[x] for x in x_values]
plt.figure(figsize=(8,5))
f = plt.scatter(np.log(x_values), np.log(y_values))
```



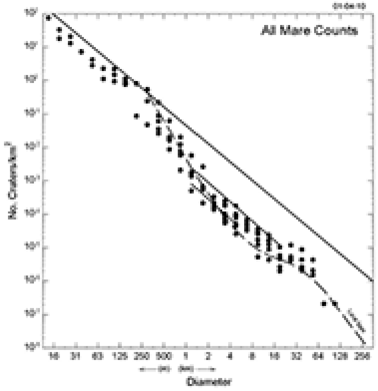
Pareto and the 80/20 rule

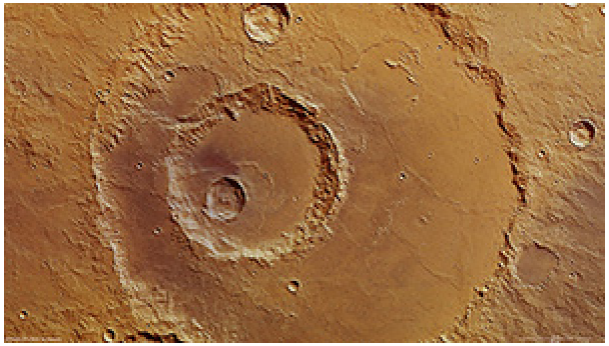
Vilfredo Pareto: Incomes follow a power law



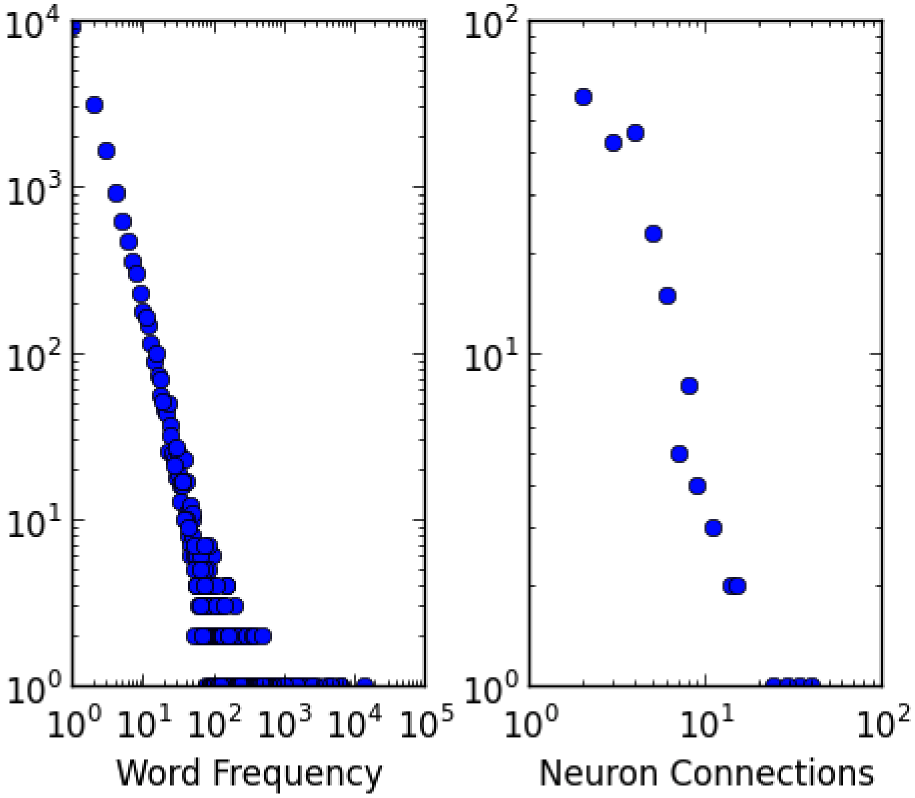
80% of the money is earned by only 20% percent of the population

Martian craters





Neuron connections and word frequency



Why ”scale-free”?

Why "scale-free"?

Continuum formalism

Density function

$$p(k) = Ck^{-\gamma}.$$

Constant C?

$$p(k) = \frac{\gamma - 1}{k_{min}} \left(\frac{k}{k_{min}} \right)^{-\gamma}.$$

Why "scale-free"?

Average value of $\mathbf{X}$

$$\begin{aligned}
\mathrm{E}(X) &= \int_{k_{min}}^{\infty} kp(k)dk \\
&= C \int_{k_{min}}^{\infty} k^{-\gamma+1} dk \\
&= \begin{cases} \infty & \text{if } \gamma \leq 2 \\ \frac{(\gamma-1)k_{min}^{\gamma-1}}{(\gamma-2)k_{min}^{\gamma-2}} = \frac{(\gamma-1)}{(\gamma-2)} k_{min} & \text{otherwise.} \end{cases}
\end{aligned}$$

Why "scale-free"?

$\mathbf{n^{th}}$ moment of $\mathbf{X}$

$$E(X^n) = \int_{k_{min}}^{\infty} k^n p(k).$$

We have

$$E(X^n) = C \int_{k_{min}}^{\infty} k^{-\gamma+n} = \begin{cases} \infty & \text{if } \gamma - n \leq 1 \\ \frac{\gamma-1}{\gamma-1-n} k_{min}^{n-1} & \text{otherwise.} \end{cases}$$

Why ”scale-free”?

We generally have $2 < \gamma < 3$.

Consequence:

- the average value of the degree of a node is finite
- its variance tends to ∞ (when the number of nodes tends to infinity)

Thus **the fluctuations around the average degree can be arbitrary large**, which explains the term *scale-free*.

Degree distribution in graphs

	γ
INTERNET	3.42
WWW (IN)	2.00
WWW (OUT)	2.31
POWER GRID	4.00
MOBILE PHONE CALLS (IN)	4.69
MOBILE PHONE CALLS (OUT)	5.01
EMAIL-PRE (IN)	3.43
EMAIL-PRE (OUT)	2.03

Degree distribution in graphs

SCIENCE COLLABORATION	3.35
ACTOR NETWORK	2.12
CITATION NETWORK (IN)	2.79
CITATION NETWORK (OUT)	4.00
E.COLI METABOLISM (IN)	2.43
E.COLI METABOLISM (OUT)	2.90
YEAST PROTEIN INTERACTIONS	2.89

Preferential attachment model

- Nodes appear over time (growing network).
- Nodes prefer to attach to nodes with many connections.

https://vimeo.com/53071346 (https://vimeo.com/53071346)

See **Barabasi-Albert model** (Lecture 2)

Part 2

Small-world property

Six degrees of separation

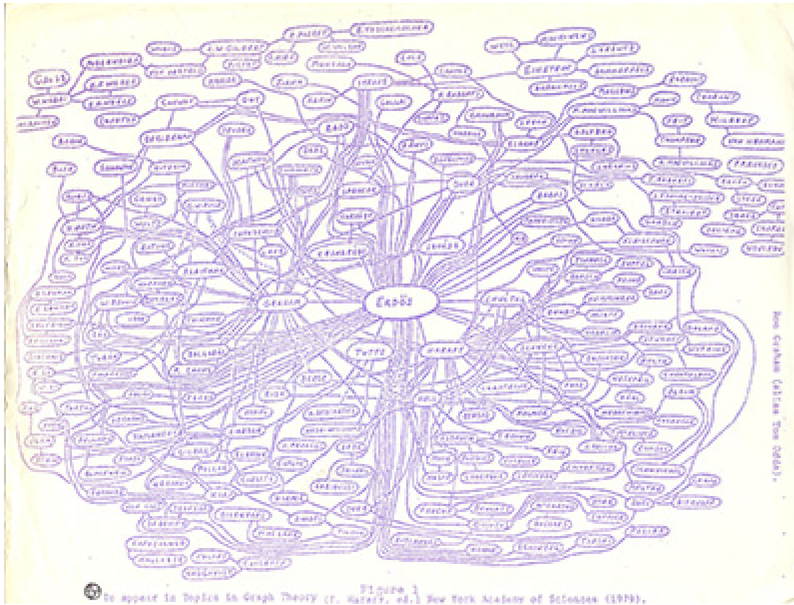
Idea: *"Any two people on Earth are **six or fewer acquaintance links apart**."* Frigyes Karinthy, 1929.

Facebook $\longrightarrow$ The average distance between any two users is 3.56.

Twitter $\longrightarrow$ About 50% of people on Twitter are **only four steps away** from each other.

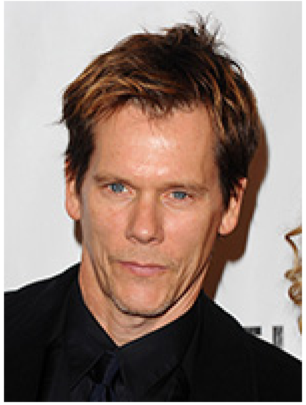
Erdõs Number

Collaborative distance to Paul Erdõs



Bacon number

- Create a graph between actors where you connect two actors if they co-appeared in the movie
- **Bacon number**: number of steps to Kevin Bacon.



- The highest (finite) Bacon number was 8 in 2007.
- About 10% of all actors were not connected to him.

Milgram's small world experiment (1967)

- Picked 300 people in 2 cities in Kansas and Nebraska.
- Ask them to get a letter to a given target in Boston by **passing it through friends**.

Milgram's small world experiment (1967)

- 64 letters reached the target.
- The average length of a chain was 6.2 → *six degrees of separation*

Milgram Experiment: Results

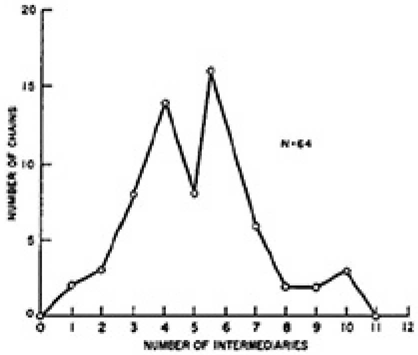
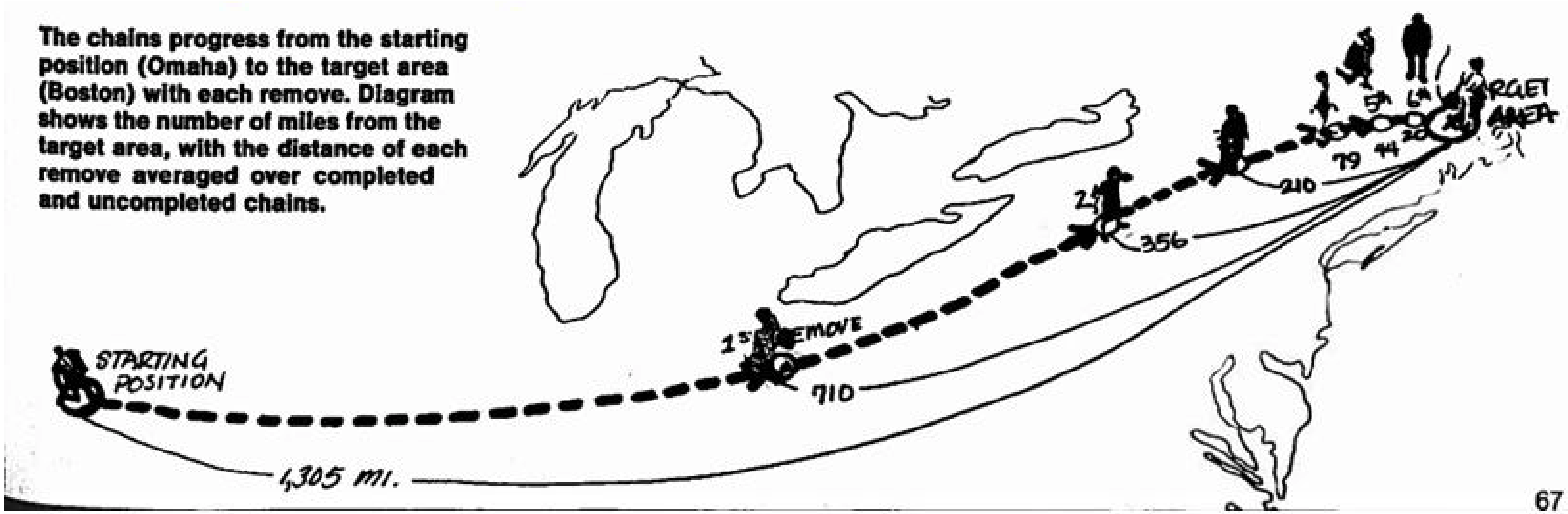


Figure 2.10: A histogram from Travers and Milgram's paper on their small-world experiment [391]. For each possible length (labeled "number of intermediaries" on the x-axis), the plot shows the number of successfully completed chains of that length. In total, 64 chains reached the target person, with a median length of six.

Milgram’s small world experiment (1967)

Observations:

- People who owned stock had shorter paths to the stockbroker than random people: 5.4 vs. 6.7.
- People from the Boston area have even closer paths: 4.4.
- 31 of 64 chains passed through 1 of 3 people as their final step.



Milgram’s small world experiment (1967)

Two conclusions:

1. Short paths are there in abundance.
2. People, acting without any sort of global “map” of the network, are effective at collectively finding these short paths.

Remark:

- $1 \not\Rightarrow 2$
- Ex: “Forward this letter to user number 482285204, using only people you know on a first-name basis.”

Another experiment:

Dodds, Muhamad and Watts (2003)

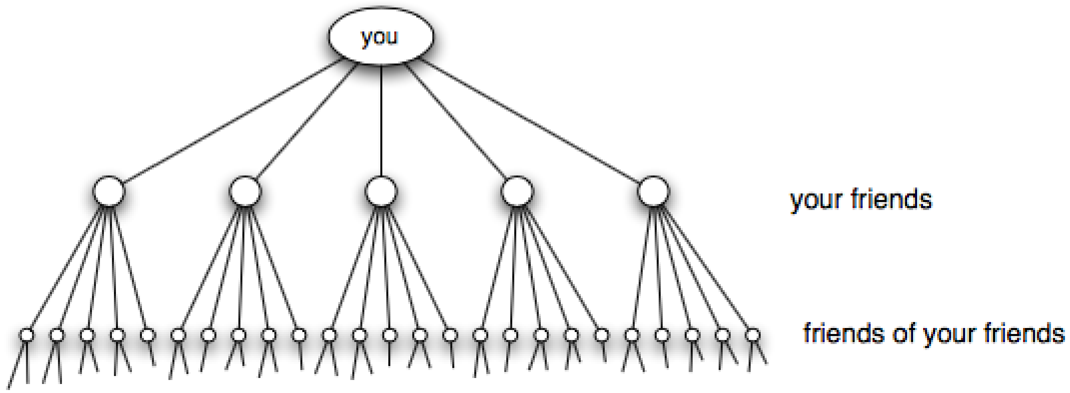
- Experiment using e-mail.
- 18 targets of various backgrounds.
- 24,000 first steps.
- 384 chains completed (1.5%).
- Average chain length = 4.01 (without correction).
- After the correction: **typical path length = 7**.

Small world phenomenon: why?

First intuition

Suppose each of us knows more than 100 other people.

- Step 1: reach 100 people
- Step 2: reach $100 \times 100 = 10,000$ people
- Step 3: reach $100 \times 100 \times 100 = 1,000,000$ people
- Step 4: reach $100 \times 100 \times 100 \times 100 = 100$M people
- In 5 steps we can reach 10 billion people.



Small world phenomenon: why?

First intuition

n: number of nodes, h: number of steps, $\bar{k}$: average degree

Based on the first intuition:

$$\bar{k}^h = n$$

Thus:

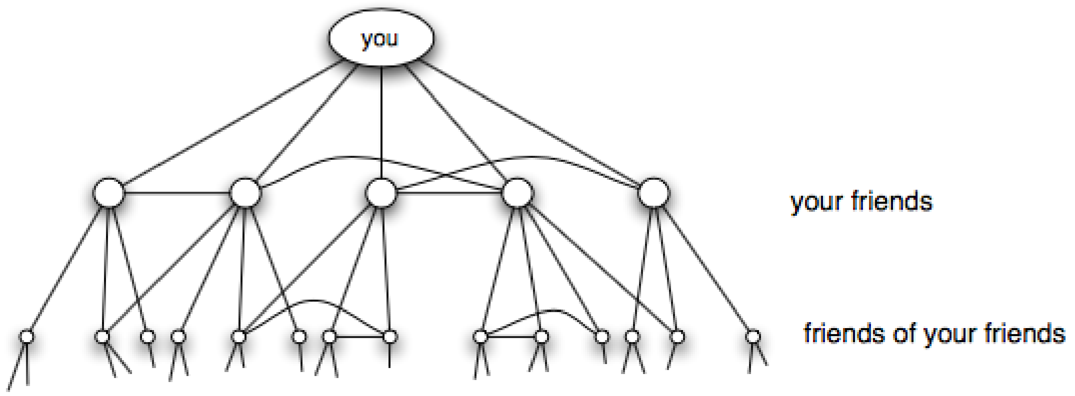
$$h = \frac{\log(n)}{\log(\bar{k})} = O(\log(n))$$

Problem in the reasoning!

Small world phenomenon: why?

Triadic closure: ”the friend of my friend is my friend”

- Problem with the previous reasoning.
- In Step 2: many of your 100 friends will know each other.
- Triangles reduces the growth rate.



Remark: 92% of new FB friendships are to a friend-of-a-friend [Backstom-Leskovec ‘11].

Clustering coefficient

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of vertices}}$$

$$C = \frac{\text{number of closed triplets of vertices}}{\text{number of connected triplets of vertices}}$$

For an Erdõs Rényi random graph $G(n, p)$:

$$C = p = \frac{\text{average degree}}{n} = \frac{\bar{k}}{n} \underset{n \to \infty}{\to} 0$$

For real social network graph: C is much higher!

Real-life networks

h = average path length

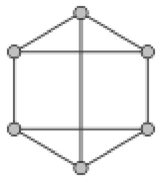
C = clustering coefficient

Network	$\mathbf{N}$	$\bar{\mathbf{k}}$	$\mathbf{C}$	$\mathbf{C_{random}}$	$\mathbf{h}$	$\mathbf{h_{random}}$
Film actors	225,226	61	0.79	0.00027	3.65	2.99
Power grid	4,941	2.67	0.080	0.005	18.70	12.40
Network of neurons	282	14	0.28	0.05	2.65	2.25

Random $\mathbf{k}$-Regular graphs

Regular graph: each node has the same degree.

k-regular graph: each node has k neighbors.



3-regular graph

Random k-regular graph:

- Each node has k half edges.
- These half edges are randomly paired up.

Goal: Compute the path length between two random nodes in a random k-regular graph.

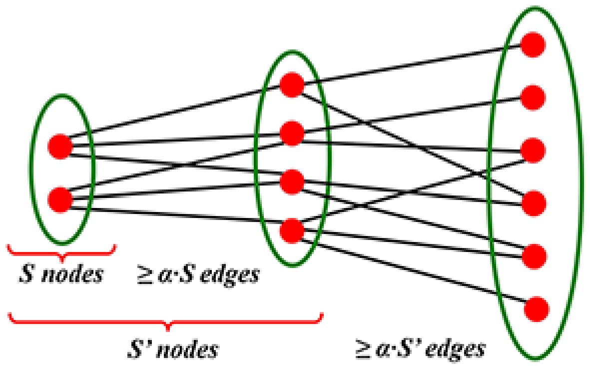
Expansion

A graph has expansion α if:

$$\forall S \in V, \quad \text{nb of edges leaving } S \geq \alpha \cdot \min(|S|, |V \setminus S|)$$

In other words:

$$\alpha = \min_{S \subset V} \frac{\text{nb of edges leaving } S}{\min(|S|, |V \setminus S|)} = \min_{|S| \leq n/2} \frac{|\partial(S)|}{|S|}$$



Random $\mathbf{k}$-Regular graphs

Number of nodes n and expansion α

We consider the breath-first search (BFS) from s.

We denote by S_j **the set of all nodes found within j steps**.

For all $j \geq 0$, we have

$$|S_{j+1}| \geq |S_j| + \frac{|\partial S_j|}{k},$$

because at most k edges of ∂S_j collide at one node of $V \setminus S_j$ (G is k-regular).

Random $\mathbf{k}$-Regular graphs

$$\alpha = \min_{S \in V : |S| \leq n/2} \frac{|\partial S|}{|S|},$$

If $|S_j| \leq n/2$, we have

$$|\partial S_j| \geq \alpha |S_j|,$$

thus

$$\begin{aligned} |S_{j+1}| &\geq |S_j| + \frac{\alpha |S_j|}{k} \geq |S_j| \left(1 + \frac{\alpha}{k}\right) \\ &\geq |S_0| \left(1 + \frac{\alpha}{k}\right)^{j+1} \geq \left(1 + \frac{\alpha}{k}\right)^{j+1}. \end{aligned}$$

Random $\mathbf{k}$-Regular graphs

We want to know the numbers of steps needed to reach more than $\frac{n}{2}$ nodes of V ?

$$|S_j| = \left(1 + \frac{\alpha}{k}\right)^j > n/2$$

Taking the logarithm, we obtain

$$j \log\left(1 + \frac{\alpha}{k}\right) > \log n - \log 2.$$

As $\log(1 + \alpha/k) \leq \alpha/k$, we have

$$j\frac{\alpha}{k} > \log n - \log 2 \Leftrightarrow \frac{j}{\log 2} > \frac{k}{\alpha}(\log_2 n - 1).$$

In particular, if $j = k \log_2 n/\alpha$, then $|S_j| > n/2$, i.e. **the BFS reach more than half of the nodes of $\mathbf{V}$.**

Random $\mathbf{k}$-Regular graphs

Number of nodes n and expansion α

Theorem:

Between all pairs of nodes s and t, there exists a path of length $O(\frac{\log(n)}{\alpha})$.

Watts-Strogatz model

Small world random graph model

Model two properties of social networks.

Homophily

We connect to others who are like ourselves.

Weak ties

We have links with acquaintances that connect us to parts of the network that would otherwise be far away.

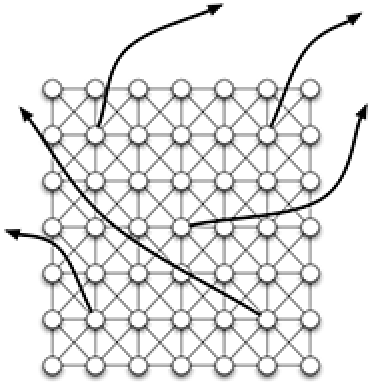
Watts-Strogatz model

Step 1: Low dimensional regular latice

Ex: two-dimensional grid

Step 2: shortcuts

For some other constant value k: connect each node to k other nodes selected uniformly at random



Watts-Strogatz model

QUESTION

For a given pair of nodes s and t, what is the minimal path length?

Decentralized search

One source s and one target t

Task: s must send a message to t

Constraints:

- **Decentralized**: No supervisor and no map of the network.
- Each node knows only the links to its neighbors.
- But each node knows the "location" of t and the "locations" of its friends.

Small-world $\neq$ Searchable

Decentralized search in Watts-Strogatz

1-dim latice where each node has 1 random edge

THEOREM: search time is $\geq O(n^{1/2})$.