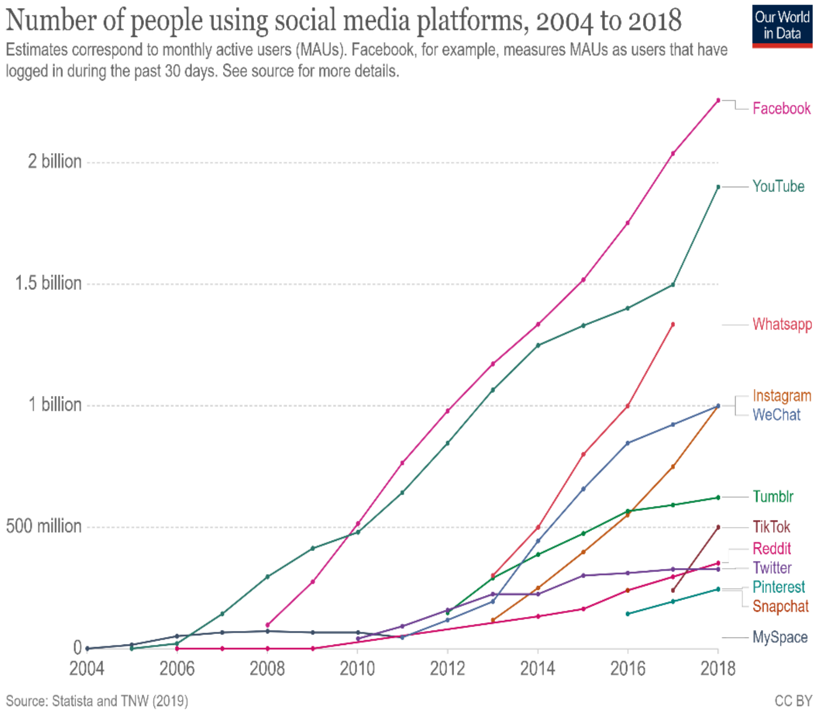
Social Network Analysis

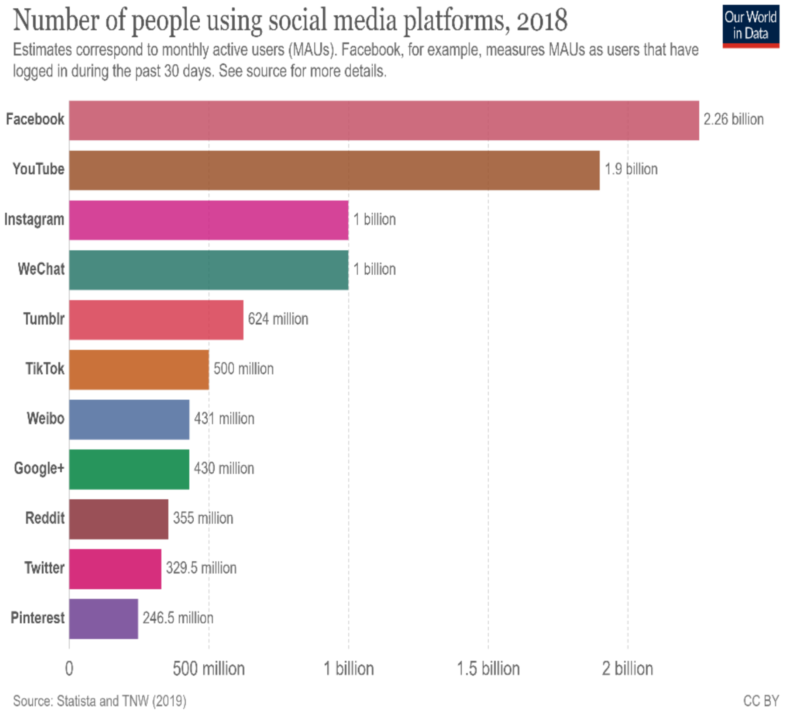
NETWORK GROWTH

Tanmoy Chakraborty

Rise of Online Social Networks



- Online social networks growing rapidly year-by-year
- Their sizes are huge
- In 2014, Facebook possessed 1.39 billion active users and 400 billion friendship links
- March 2015, Twitter has 288 million active users and 60 billion followers



Such huge volume of these online networks restrict the active research with real social networks

https://ourworldindata.org/rise-of-social-media

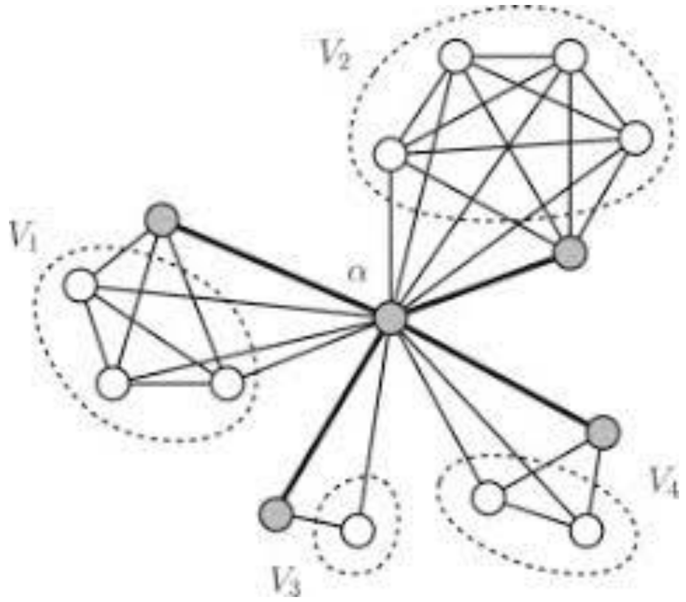
Synthetic Networks

- Generated using theoretical network models
- Often possesses strong underlying mathematical foundation
- Often can simulate important real-world network characteristics
- Help getting insights of the real-life networks
- Allow experimentation through simulation when real networks are unavailable
- Can establish network insights on concrete theoretical foundations

Properties of Real-world Networks

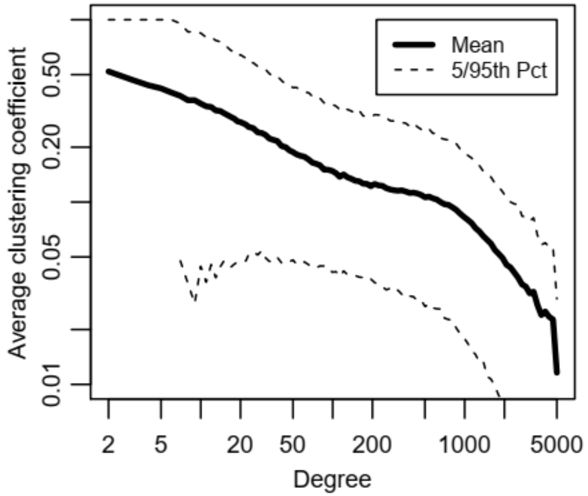
- High average local clustering coefficient
- Small-world property
- Scale-free property

High Average Local Clustering Coefficient



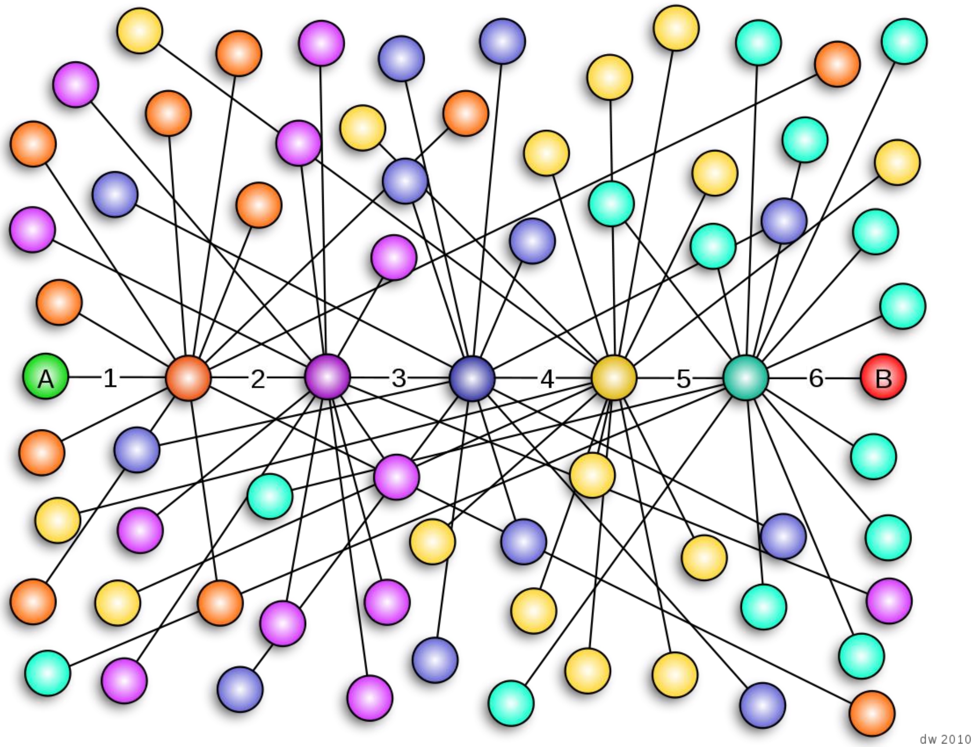
https://www.sciencedirect.com/science/article/pii/S0166218X16304589

- ❑ Neighbors of a node tend to be highly connected with each other at individual level
- ❑ In Facebook social graph, for users with 100 friends have average local clustering coefficient of 0.14
- ❑ For a graph of 150 million nodes, the number is unexpectedly high



Clustering coefficient of Facebook Social Graph
Ugander et al. (2011)

Small-world Property



The "six degrees of separation" model

https://en.wikipedia.org/wiki/Small-world_experiment

- ❑ An outcome of Stanley Milgram's small-world experiment (1967) to measure the probability of two random persons being known to each other
- ❑ Can also be viewed in light of the average path length between two randomly chosen nodes in a network
- ❑ A network G is said to follow the small-world property if the average path length of the network is logarithmically proportional to the network size.

$$Average\ Path\ Length\ \propto \log(Network\ Size)$$

Six Degrees of Separation: Milgram's Small-World Experiment

A possible path of a letter

https://en.wikipedia.org/wiki/Small-world_experiment

- Source cities: Omaha, Nebraska, and Wichita, Kansas
- Destination city: Boston, Massachusetts
- Information packets sent initially to random persons in Omaha/Wichita
- Recipient was to forward the letter
  - Directly to target, if she knew the target personally
  - Else to some friend/relative who more likely to know the target
- 64 out of 296 letters eventually reach the target contact
- Number of intermediates among the traversal path chains:

Maximum: 11 Minimum: 1 Median: 5.2 (≈6)

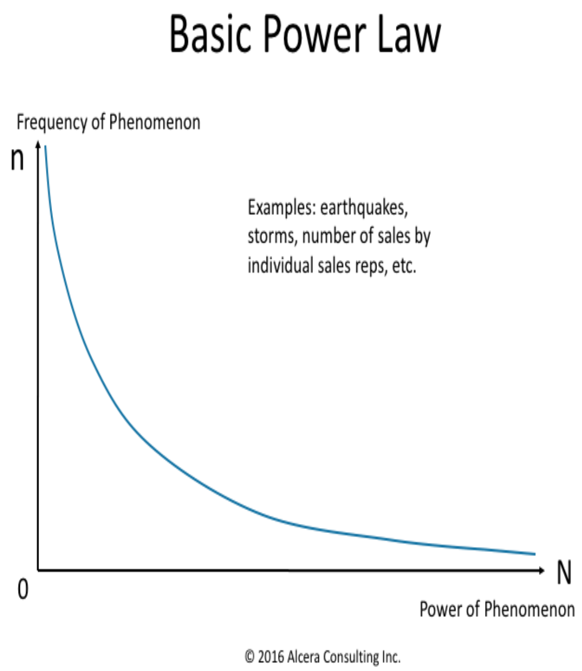
Small-world of Social Media!!!

- Average chain of contacts in Microsoft Messenger was 6.6 people [Leskovec and Horvitz 2007]

- Average distance between two random Facebook users in 2011 was 4.74 with 3.74 intermediaries [Four Degree of Separation by Backstrom et al. (2012)]

- Average distance between two random Facebook users in 2016 was 4.57 with 3.57 intermediaries [Repeat Experiment by Facebook in 2016]

Scale-free Property: Power Law



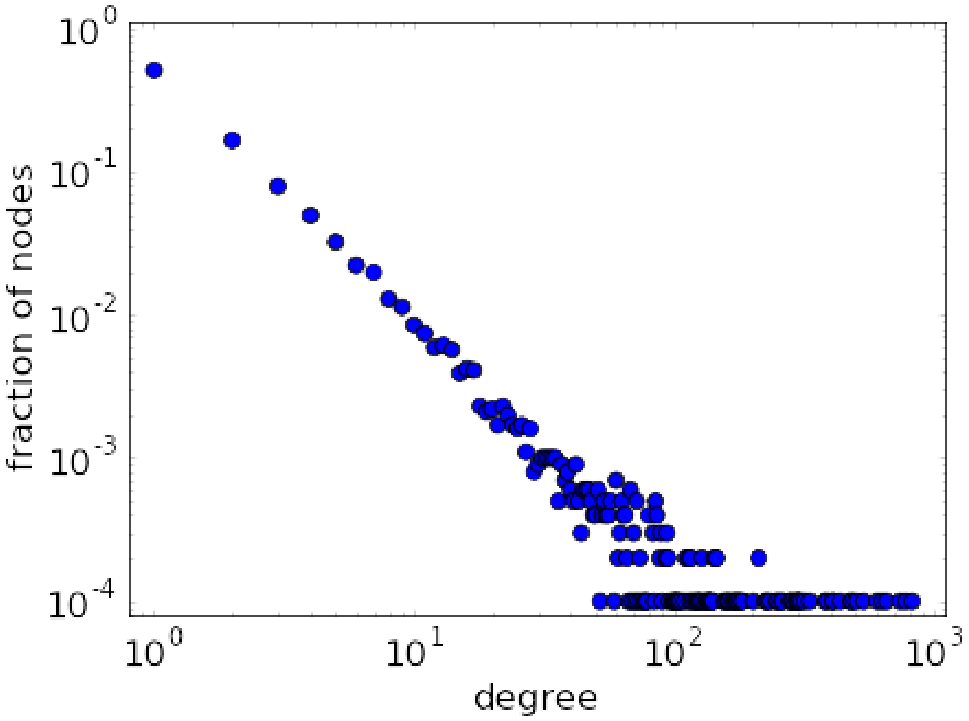
https://exploitingchange.com/2016/09/14/another-powerful-idea-the-power-law/

- Power Law: a relative change in the value of one variable leads to the proportional change in the value of other variable
- Independent of the initial values of both the variable
- Mathematically,

$$y \propto x^{-b}, b \in \mathbb{R}$$

- Functions that follows power-law are scale-invariant
- Pareto Principle (or the 80/20 rule) in Economics: 80% of the outcomes are results of 20% of the causes
- Power law principle is also coined as Pareto Distribution

Scale-free Property: Real Networks

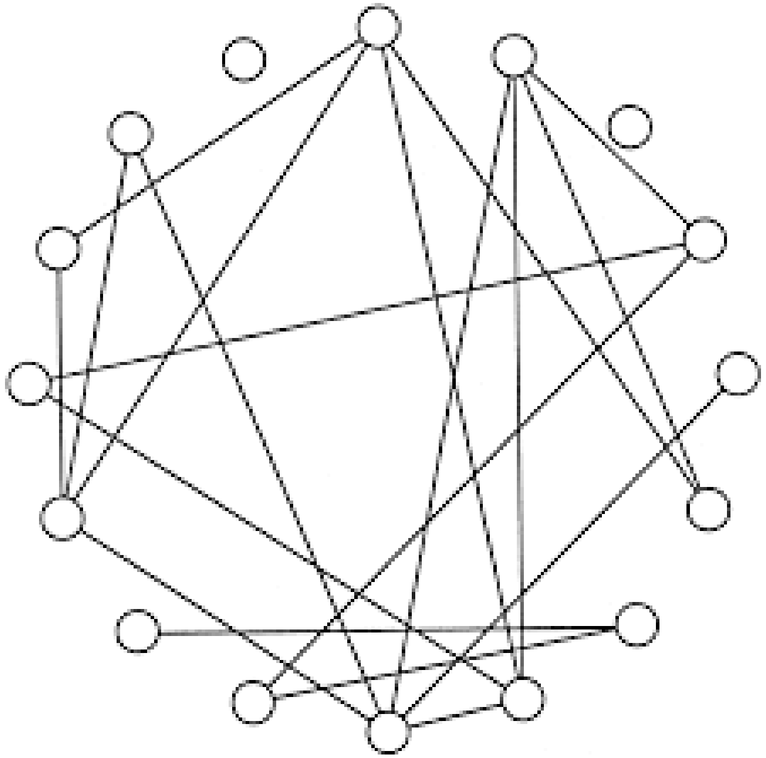


Degree distribution of a scale-free network
https://mathinsight.org/scale_free_network

- ❑ Networks whose degree distribution follows power-law are known as scale-free networks

- ❑ Most real-world networks are found to be scale-free

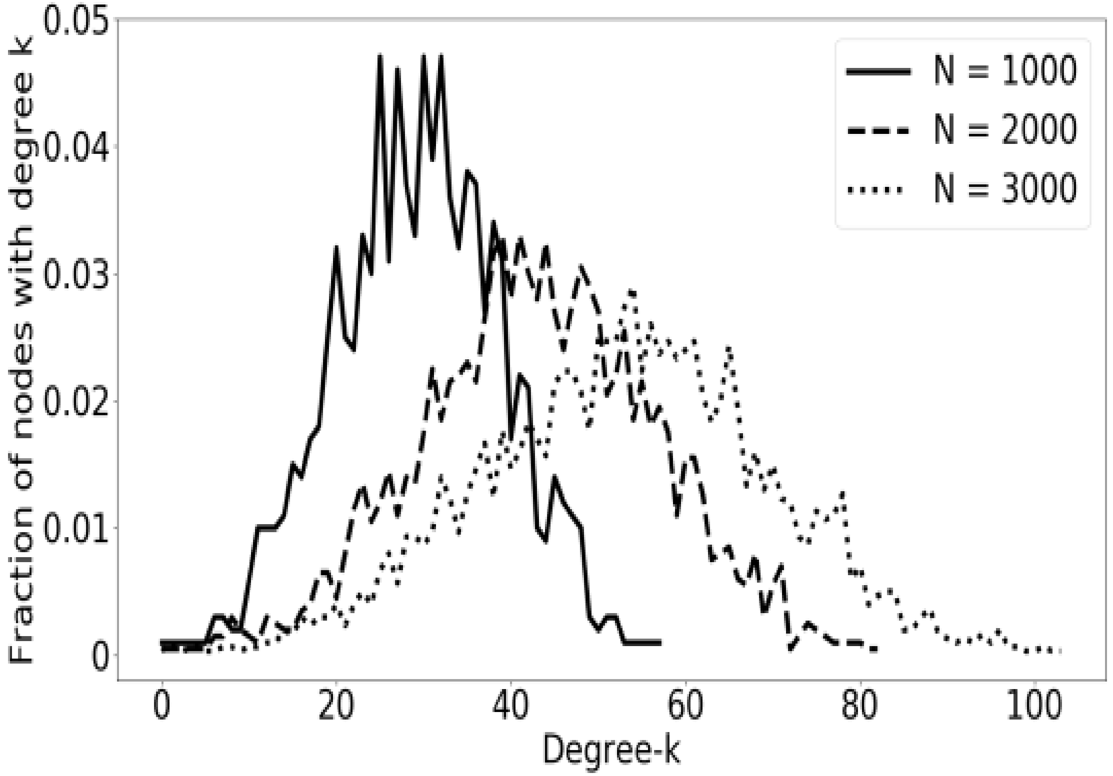
Synthetic Networks: Random Network Model



An instance of $G(16, \frac{1}{7})$ network

- Also popularly known as Erdős and Rényi model (or ER model)
- A number of variants of the model
- Popular variants:
  - $G(N, K)$ model [Erdős and Rényi 1959]: From the set of all networks of N nodes and K edges, a network is chosen uniformly at random
  - $G(N, p)$ model [Gilbert 1959]: Network has N nodes, and any random pair of nodes has a probability p of being adjacent independently with any other pair of nodes in the network
- Both the variants behave identically in the limiting case
- $G(N, p)$ model considered as the standard random network model

Erdős-Rényi Network: Degree Distribution



Comparison of degree distribution

- For a node in $G(N, p)$ to have degree k, the corresponding node must be adjacent to k other nodes of the network
- We can choose these nodes in $\binom{N-1}{k}$ ways
- Probability $P(k)$ of a node to have degree k is given by

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

- Letting $N \to \infty$,

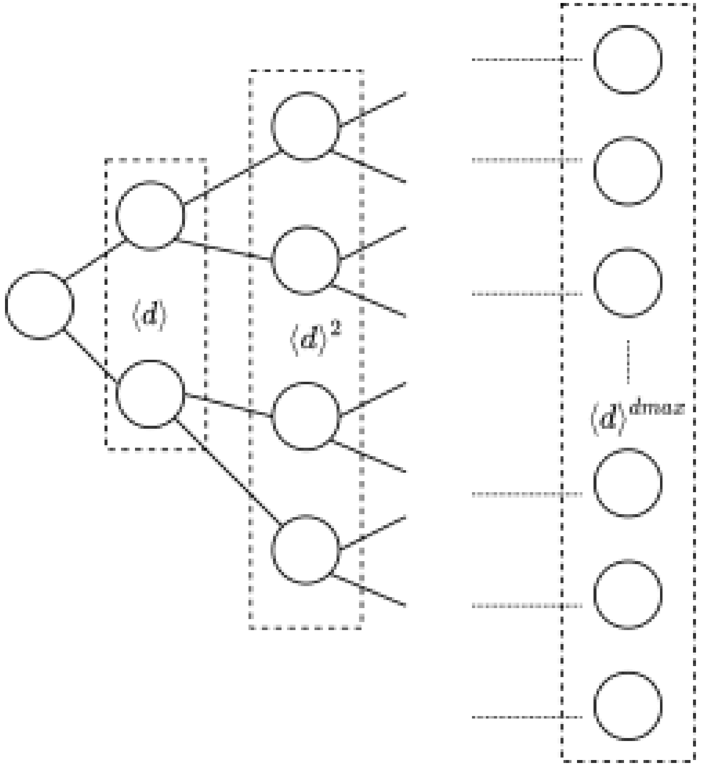
$$P(k) = \frac{e^{-\langle d \rangle} \langle d \rangle^k}{k!}$$

Erdős-Rényi Network: Emergence of Giant Component

Theorem: A giant component emerges in a random network when the average degree of the network is greater than or equal to unity, i.e., $\langle k \rangle \geq 1$.

- For emergence of a giant component, only one link per node on-an-average is sufficient!! The above condition necessary, too
- The emergence is not a smooth, gradual process; it follows a second-order phase transition
- Regimes of evolution:
  - Subcritical Regime ($0 < \langle k \rangle < 1$) $\Rightarrow$ A number of small isolated clusters in the network, as the number of links is much less than the number of nodes
  - Critical point ($\langle k \rangle = 1$) $\Rightarrow$ A distinguishable giant component emerges
  - Supercritical Regime ($\langle k \rangle > 1$) $\Rightarrow$ A growing giant component, and less and less smaller isolated clusters and nodes
  - Connected Regime ($\langle k \rangle > \ln N$) $\Rightarrow$ The giant component absorbs all nodes and components, the network becomes connected

Erdős-Rényi Network: Average Path Length



A depiction of random network in tree format

- When l_{max} represent the maximum path length of $G(N,p)$,

$$1 + \langle d \rangle + \langle d \rangle^2 + \langle d \rangle^3 + \cdots \cdots \cdots + \langle d \rangle^{l_{max}} = N$$

- When $\langle d \rangle \gg 1$, the above yields,

$$l_{max} \approx \frac{logN}{log\langle d \rangle}$$

- Further approximation yields,

$$\langle l \rangle \propto logN$$

Theorem: Erdős-Rényi Networks follow small-world property.

Erdős-Rényi Network: Clustering Coefficient

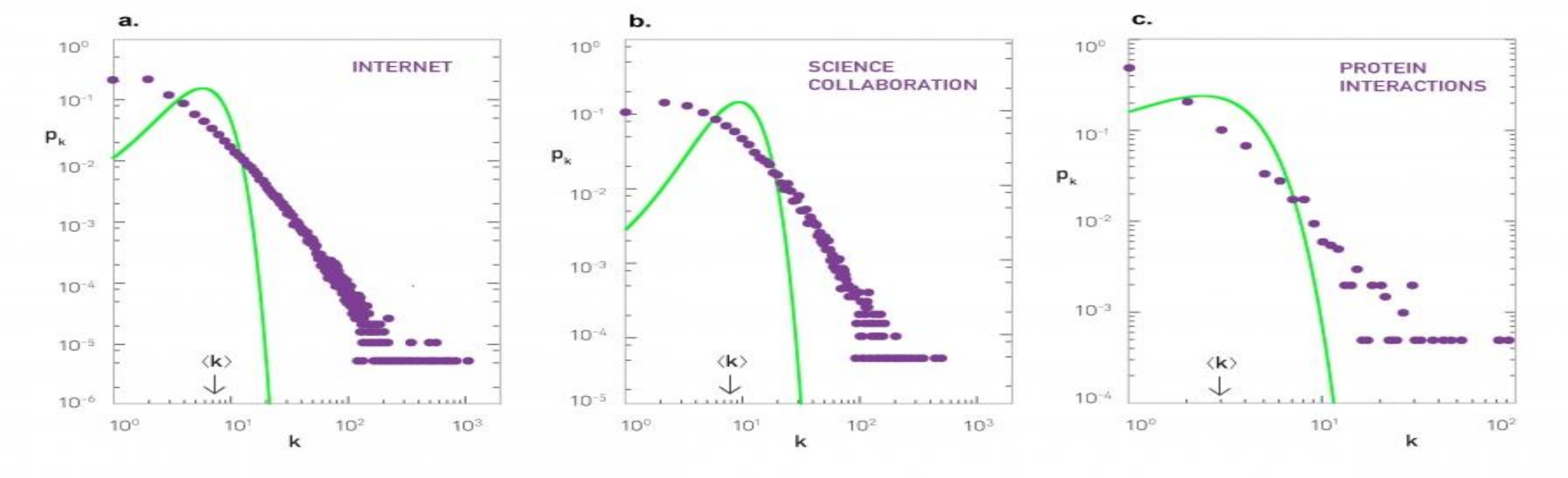
- In $G(N,p)$
  - Number of possible edges between neighbours of a node: $\binom{\langle d \rangle}{2}$
  - Expected number of edges between these nodes: $p \times \binom{\langle d \rangle}{2}$
- The above yields, the local clustering coefficient for a node $v_i \in G$

$$C_i = p \approx \frac{\langle d \rangle}{N}$$

Theorem: The local clustering coefficient for any node in an Erdős-Rényi Network is inversely proportional to the size of the network

Note: The local clustering coefficient for any node in an Erdős-Rényi Network does not depend on the degree of the node

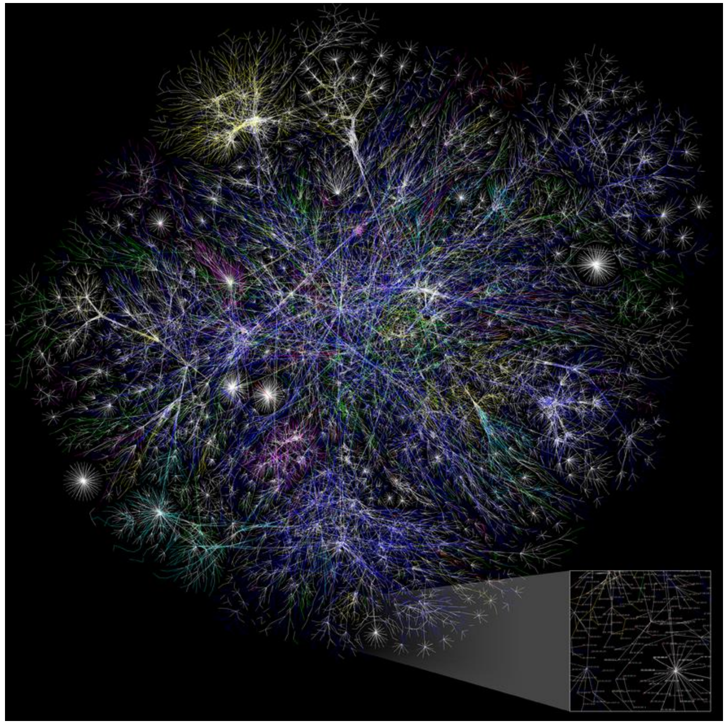
Erdős-Rényi Networks vs. Real-life Networks: Degree Distribution



Real-life networks are often scale-free; however, Erdős-Rényi Networks are not

http://networksciencebook.com/chapter/3#not-poisson

Erdős-Rényi Networks vs. Real-life Networks: Presence of Outliers



Partial map of the Internet based on the January 15, 2005. Hubs are highlighted

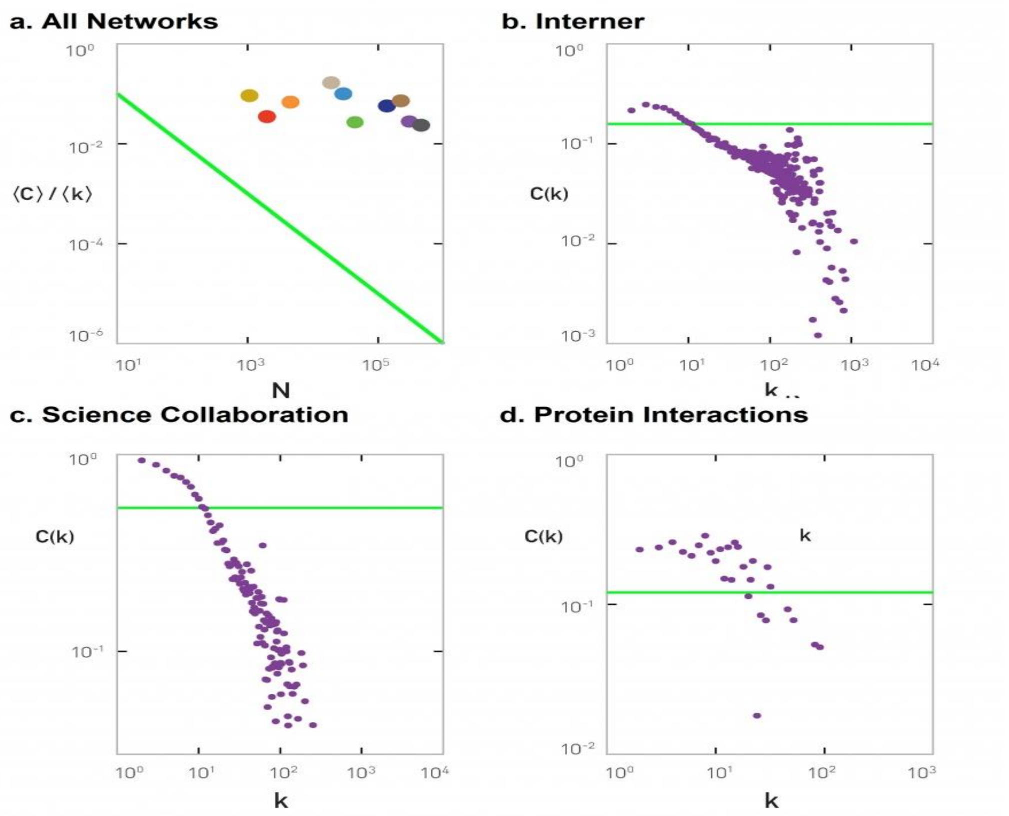
https://en.wikipedia.org/wiki/Hub_(network_science)

- In an Erdős-Rényi Network, probability of having a node with a high degree is extremely low
- probability of a node with 2000 neighbours is 10^{-27}!
- In real-world networks, such nodes exist (Hubs)
- Celebrities in social networks

Erdős-Rényi Networks vs. Real-life Networks: Small-world Property

- Erdős-Rényi Networks follow small-world property as the maximum path length in Erdős-Rényi Networks $\approx \frac{logN}{log\langle d\rangle}$

Erdős-Rényi Networks vs. Real-life Networks: Clustering Coefficient

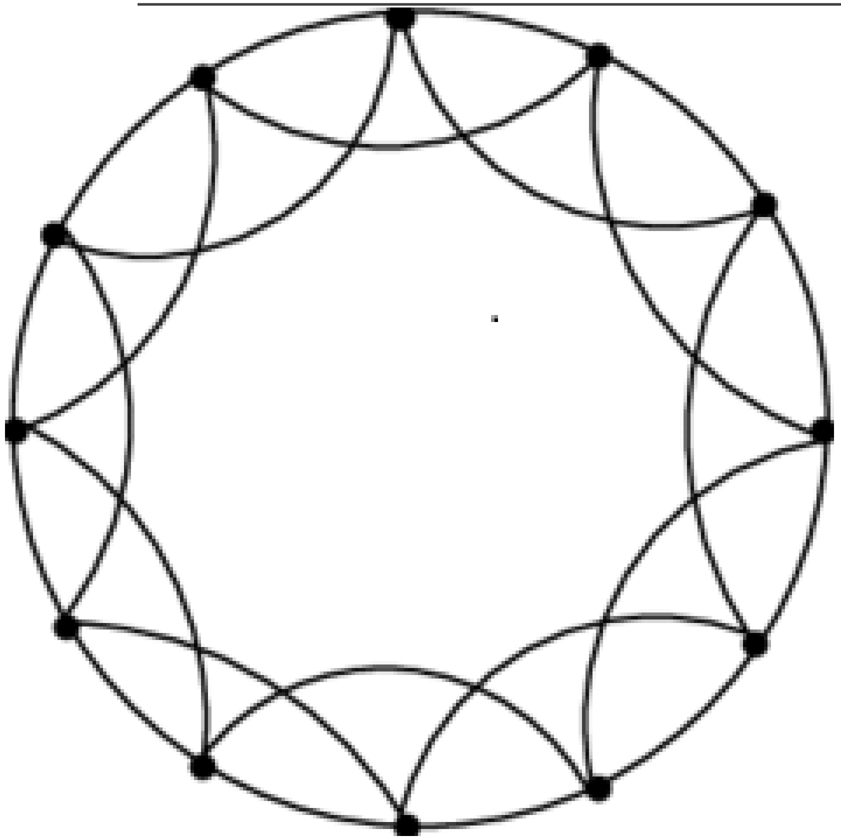


Local Clustering coefficients in real-life networks
http://networksciencebook.com/chapter/3#clustering-3-9

- ❑local clustering coefficients in Erdős-Rényi Networks decreases with the increase in network size
- ❑Nodes having high local clustering coefficient exist in enormously large real-life networks
- ❑Echo chambers in large social networks like Facebook

Network	N	L	‹k›	‹d›	d_{max}	$\ln N / \ln \langle k \rangle$
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
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile-Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	93,437	8.08	5.35	15	4.81
Actor Network	702,388	29,397,908	83.71	3.91	14	3.04
Citation Network	449,673	4,707,958	10.43	11.21	42	5.55
E. Coli Metabolism	1,039	5,802	5.58	2.98	8	4.04
Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

Regular Ring Lattice Network Model

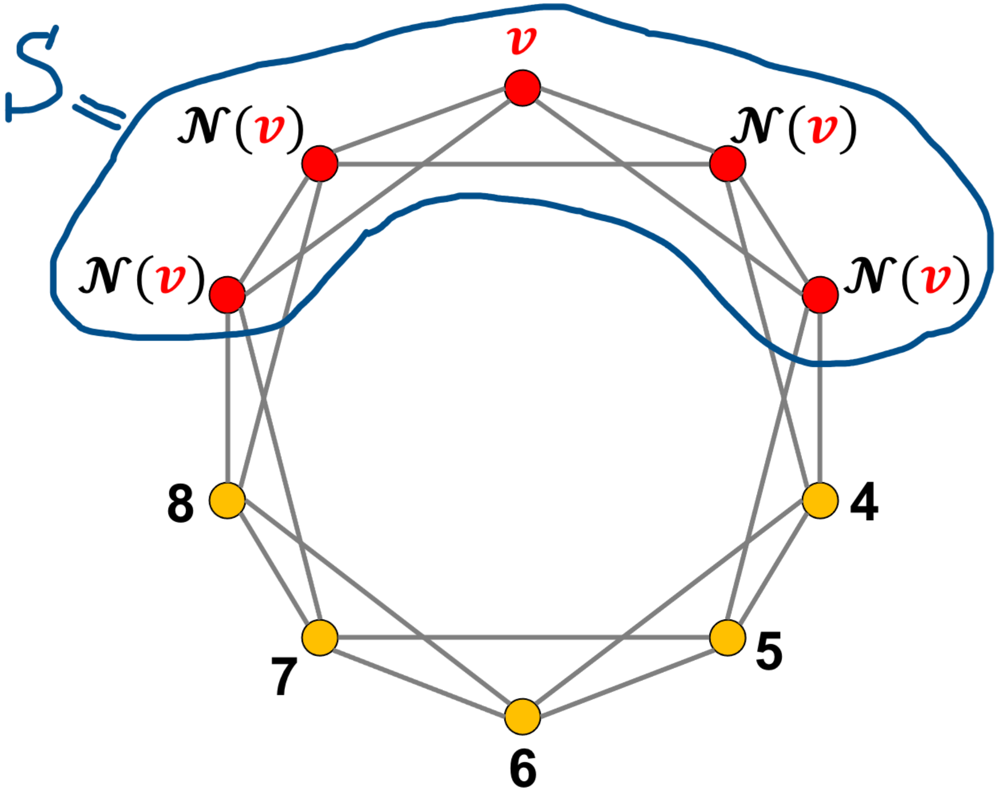


4-Regular Ring Lattice Network

- A ring lattice network consists of N nodes labeled $0, 1, 2, \cdots, N-1$ arranged in circular order

- Every node in the network is connected to exactly k other nodes, immediate $\frac{k}{2}$ rightmost nodes and $\frac{k}{2}$ leftmost nodes relative to the position of the node in the network

Regular Ring Lattice Network Model: Local Clustering Coefficient

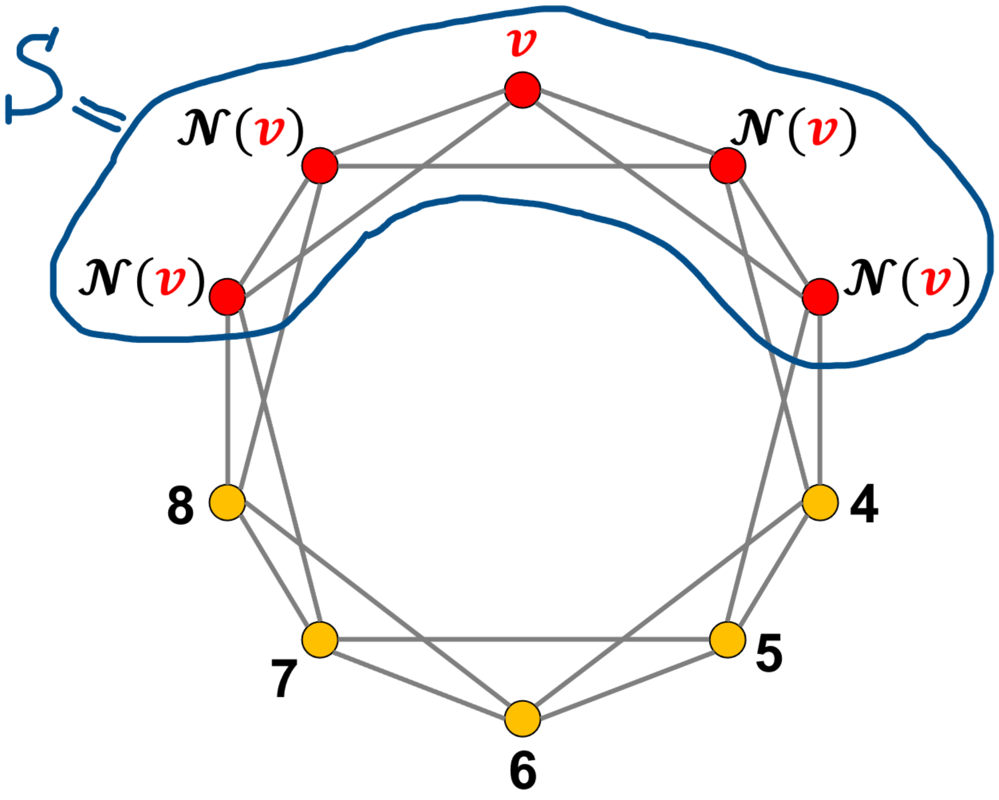


4-Regular Ring Lattice Network

Theorem: Local clustering coefficient of any node v in a $k - Regular\ Ring\ Lattice\ Network$ is $\frac{3(k-2)}{4(k-1)}$

- When k is large, then the above expression boils down to $\frac{3}{4}$
- Regular Ring Lattice networks has high local clustering coefficients

Regular Ring Lattice Network Model: Small-world Property

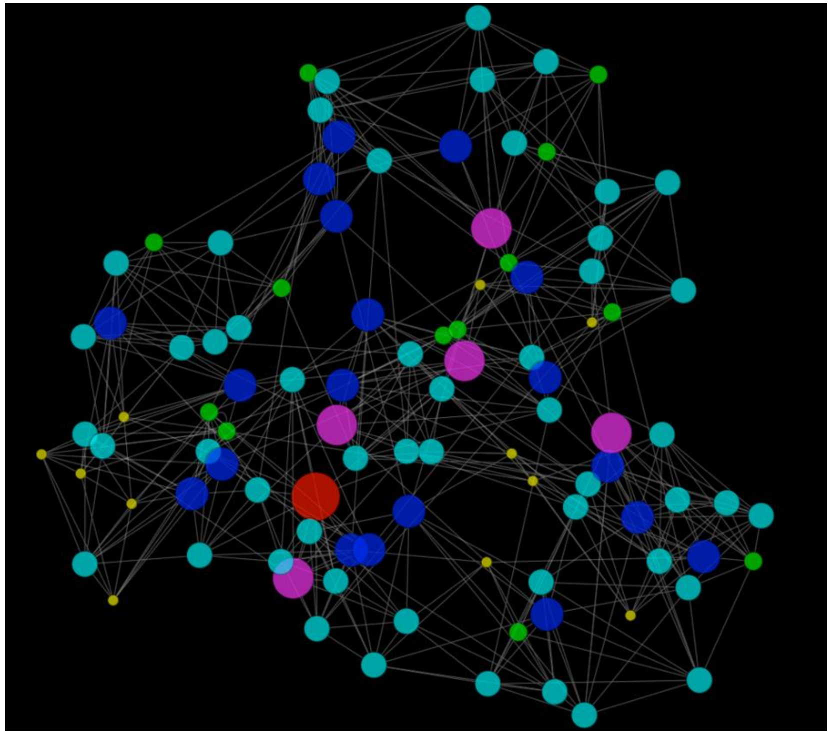


4-Regular Ring Lattice Network

- ❑ All the connection in ring lattice are to nearby nodes
- ❑ Two nodes that are in the opposite sides of the ring will be connected by long chains

Theorem: Regular Ring Lattice networks does not follow small-world property

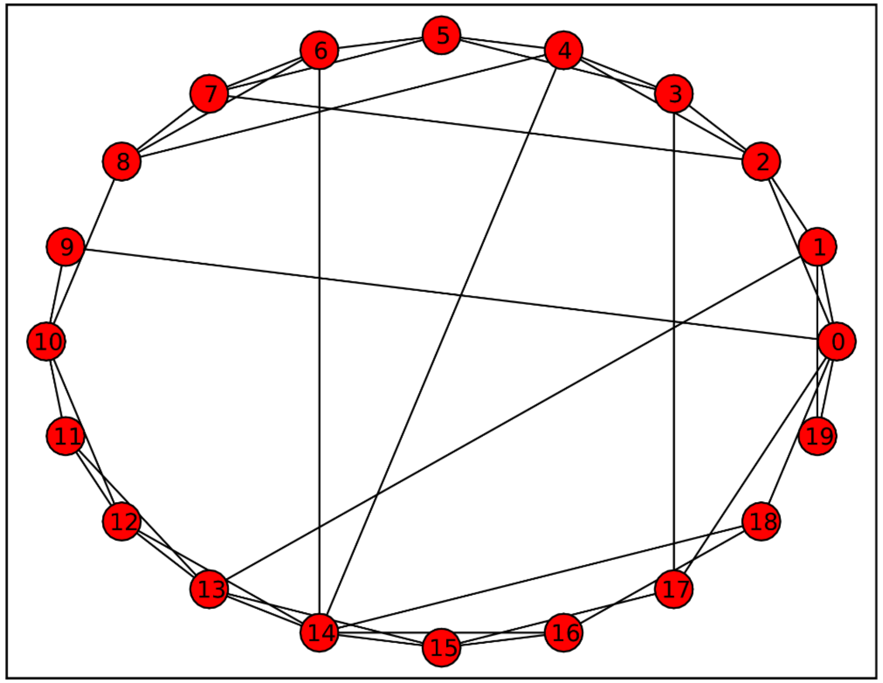
Watts-Strogatz Network Model



Watts–Strogatz network with 100 nodes formed by igraph and visualized by Cytoscape 2.5
https://en.wikipedia.org/wiki/Watts%E2%80%93Strogatz_model

- ❑Erdős-Rényi Networks
  - ❑possesses the small-world property,
  - ❑have small local clustering coefficients for all the nodes when the network size is large
- ❑Regular Ring Lattice networks
  - ❑possess high local clustering coefficients for all the node
  - ❑does not follow small-world property
- ❑Watts-Strogatz Model form networks that has a high local clustering coefficient and possesses the small-world property
- ❑Proposed by Watts and Strogatz in 1998

Watts-Strogatz Network Model: Network Formation



https://www.hindawi.com/journals/mpe/2014/693743/fig2/

1) Start with a k-regular lattice network of size N
2) List the nodes of the lattice as $1, 2, \cdots, N$
3) Choose i^{th} node from the list, $i = 1, 2, \cdots, N$
4) Select edges that that link i^{th} node to some j^{th} node $(j > i)$
5) With a fixed rewiring probability β, rewire the other end of these edges
6) Avoid formation self-loops and link-duplication
7) Repeat steps 3 through 6 until all the nodes are scanned

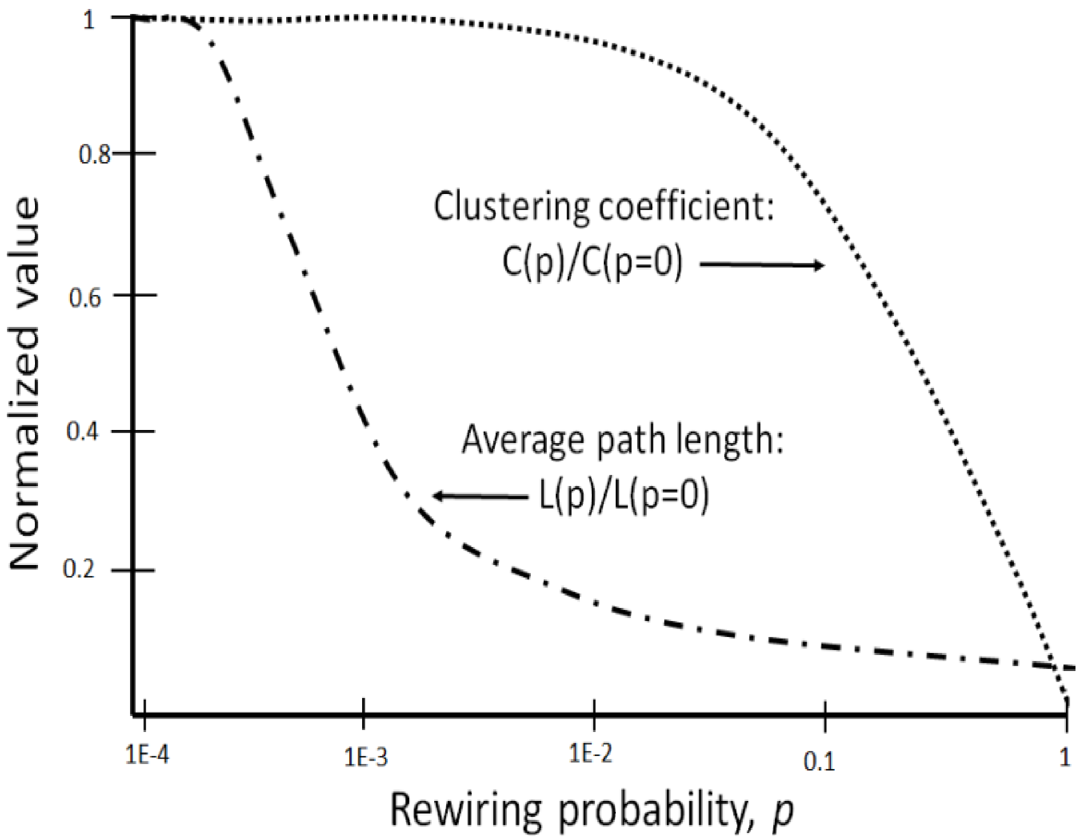
Watts-Strogatz Network Model: Properties

- Locally clustered network due to the underlying ring lattice structure
- Random rewiring of links reduces the average path length
- $\frac{\beta N k}{2}$ number of non-lattice edges introduced due to random random rewiring
- Approximates to ring lattice networks if $\beta \rightarrow 0$
- Approximates to Erdős-Rényi Networks if $\beta \rightarrow 1$

Watts-Strogatz Network Model: Average Path Length

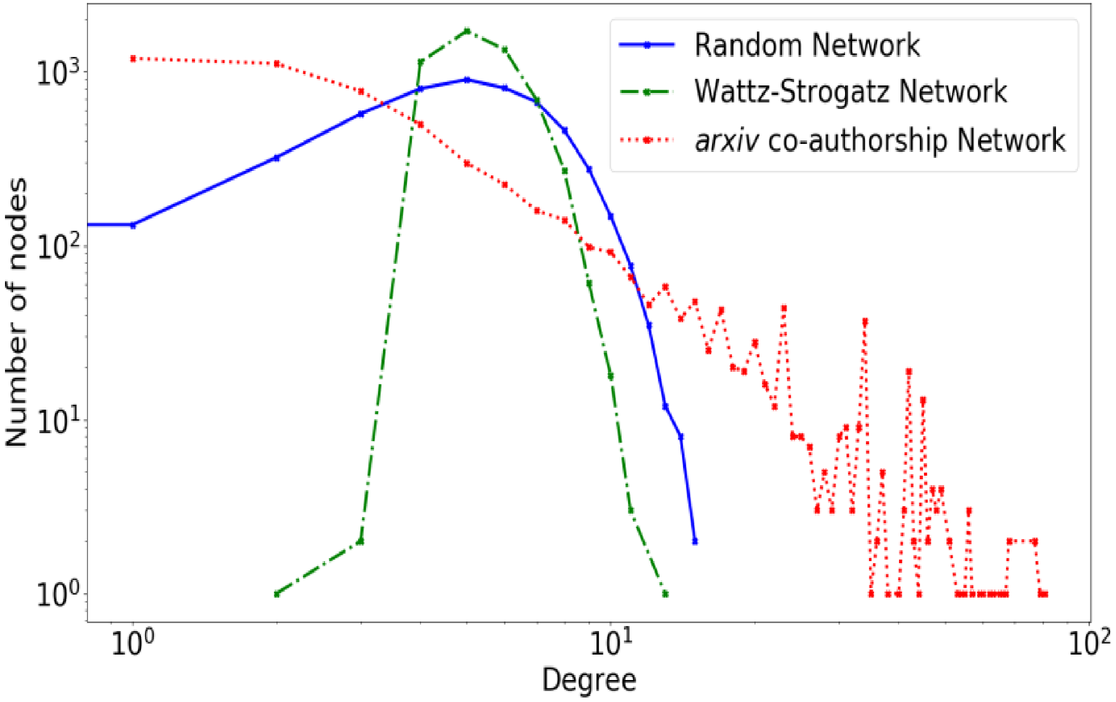
- Average path length in WS network $\approx \frac{N}{2k}$ when $\beta \to 0$
- The above scales linearly with the size of the network
- If $\beta \to 1$, average path length boils down to $\approx \frac{\ln N}{\ln k}$
- If $0 < \beta < 1$, with increase in the value of β, the average path length reduces sharply

Watts-Strogatz Network Model: Clustering Coefficient



- Average path length in WS network $\approx \frac{N}{2k}$ when $\beta \to 0$
- The above scales linearly with the size of the network
- If $\beta \to 1$, average path length boils down to $\approx \frac{\ln N}{\ln k}$
- If $0 < \beta < 1$, with increase in the value of β, the average path length reduces sharply

Watts-Strogatz Networks vs. Real-world Networks



Comparisons of Degree Distribution

- Watts-Strogatz networks have only few outliers, whereas, real-life networks have significantly high number of outliers
- Watts-Strogatz networks follow small-world property if $\beta \rightarrow 1$, however, real-world networks always follow small-world
- Degree distribution in Watts-Strogatz does not follow power law, whereas, real networks often does that
- Both Watts-Strogatz networks and real-world networks have high clustering coefficient

Barabasi-Albert Network Model

- Real-life social networks often evolve with time
- Originates with a small seed network
- The network grows as new nodes and edges gets attached to the network with time
- Barabasi-Albert model follows the same principle of network evolution
- Also known as Preferential Attachment model or rich gets richer model
- Generates scale-free networks

Barabasi-Albert Model: Network Formation

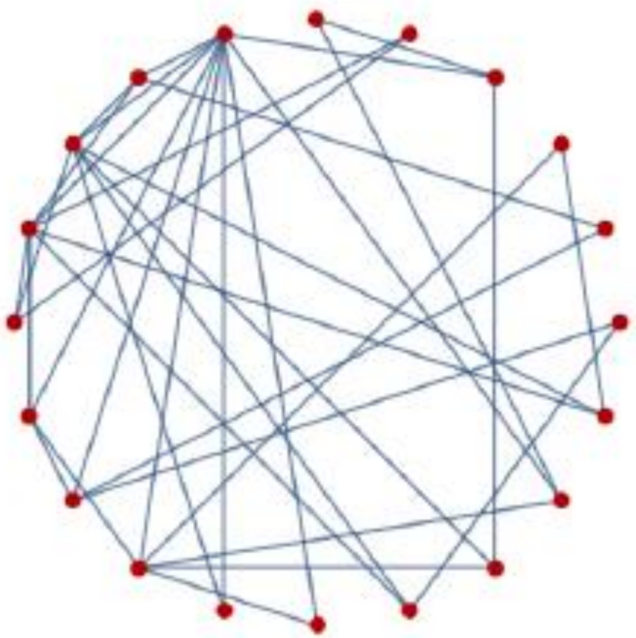
- Initially, the network has a limited number of nodes, m_0
- At every time-step, a new node with m edges enters the network
- New edges gets attached with the existing nodes based on the principle of preferential attachment as follows:
  - the probability that the new edge attaches to an existing node v_i with degree d_i is:

$$P(v_i) = \frac{d_i}{\sum_j d_j}$$

$m_0 = m = 2$

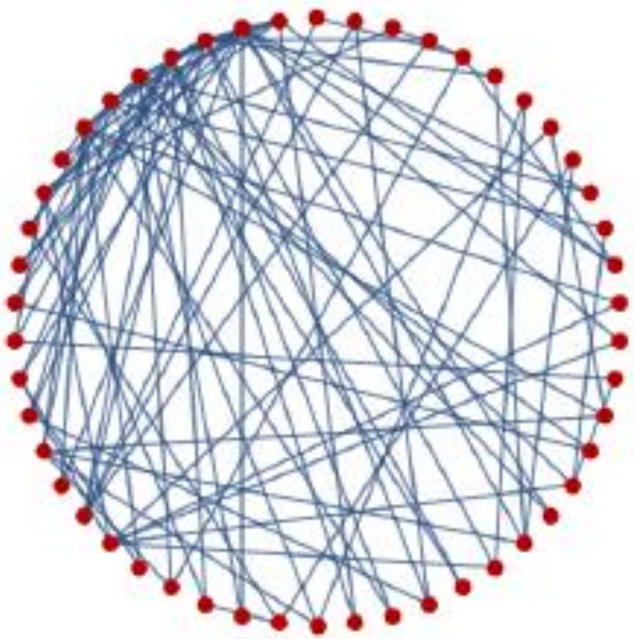
https://en.wikipedia.org/wiki/Barab%C3%A1si%E2%80%93Albert_model

Barabasi-Albert Model: Network Growth



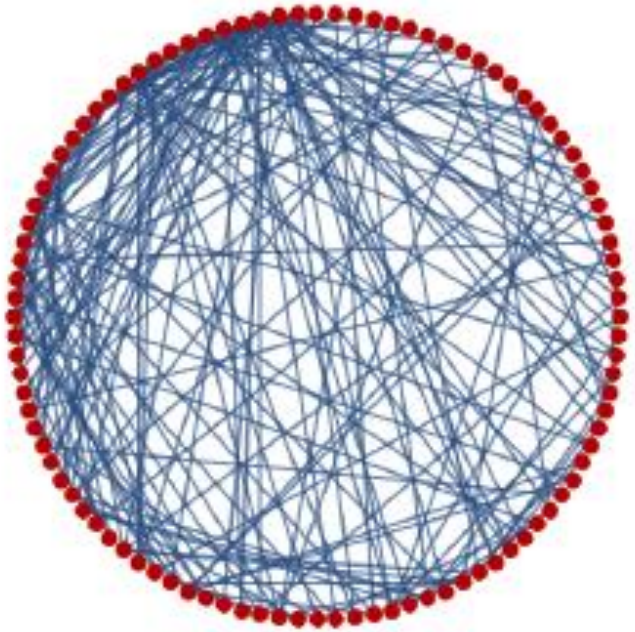
$|G| = 20$

Average Path Length = 2.16



$|G| = 50$

Average Path Length = 2.69



$|G| = 100$

Average Path Length = 3.02

https://www.researchgate.net/publication/259742981_Information_Theory_Kolmogorov_Complexity_and_Algorithmic_Probability_in_Network_Biology

Barabasi-Albert Model: Degree Dynamics

- Captures the continuous changes in degree distribution with addition of nodes and edges into the network with time
- Let the node v_i joined the network at time t_i and has degree d_i at any instance of time
- Then the degree of the above node at current instance is given by the equation

$$d_i(t) = d_i(t_i)\left(\frac{t}{t_i}\right)^{\beta} \cdots\cdots\cdots (*)$$

Where β is the dynamical exponent, and is set as $\frac{1}{2}$ in general

- Expected number of preferential attachment during joining of a node is usually denoted m
- Then the revised equation assumes the form: $d_i(t) = m\left(\frac{t}{t_i}\right)^{1/2} \cdots\cdots\cdots (**)$

Barabasi-Albert Model: Scale-free Property

- Note the equation $(*)$ as follows:

$$d_i(t) = d_i(t_i)\left(\frac{t}{t_i}\right)^{\beta}$$

- This implies

$$d_i(t) \propto t^{\beta}$$

- The above establishes that the growth model follows power law

Barabasi-Albert Model: Scale-free Property

- It implies further from the equation $(*)$

$$d_i(t) \propto \left(\frac{1}{t_i}\right)^{\beta}$$

- earlier the node joins the network, higher would be its degree (First mover advantage)

- It implies even further from equation $(*)$

$$\frac{\partial d_i(t)}{\partial t} \propto \sqrt{\frac{1}{(t_i \cdot t)}}$$

- Thus, the rate of acquiring of new edges by a node slows down with time

Barabasi-Albert Model: Scale-free Property

- It implies further from the equation $(*)$

$$d_i(t) \propto \left(\frac{1}{t_i}\right)^{\beta}$$

- earlier the node joins the network, higher would be its degree (First mover advantage)

- It implies even further from equation $(*)$

$$\frac{\partial d_i(t)}{\partial t} \propto \sqrt{\frac{1}{(t_i \cdot t)}}$$

- Thus, the rate of acquiring of new edges by a node slows down with time

Barabasi-Albert Model: Scale-free Property

- Note equation $(**)$ as,

$$d_i(t) = m\left(\frac{t}{t_i}\right)^{1/2}$$

- It follows from equation $(**)$,

$$Prob(d_i(t) < k) = Prob\left(t_i > \frac{m^2 t}{k^2}\right) = 1 - Prob(t_i \leq \frac{m^2 t}{k^2})$$

- From distribution of the attachment of edges in the network,

$$Prob(t_i) = \frac{1}{N} = \frac{1}{m_0 + t}$$

Barabasi-Albert Model: Scale-free Property

- Note equation $(**)$ as,

$$Prob\left(t_i > \frac{m^2 t}{k^2}\right) = 1 - \frac{m^2 t}{k^2} \times \frac{1}{m_0 + t}$$

- It follows from above,

$$Prob(d_i(t) < k) = 1 - \frac{m^2 t}{k^2(m_0 + t)}$$

- Then the degree distribution,

$$P(k) = \frac{\partial}{\partial k} Prob(d_i(t) < k) = \frac{2m^2 t}{k^3(m_0 + t)}$$

- Letting $t \to \infty$,

$$P(k) = \frac{2m^2}{k^3} \Rightarrow P(k) \propto \frac{1}{k^3}$$

Barabasi-Albert Model: Limitations

- Model assumes that only one node is added at a time; difficult to apply if more than one node arrive simultaneously
- Assumes a linear growth model; may not be realistic in many applications
- Predicts a fixed exponent in the power-law degree distribution; whereas, across real-world networks, exponent varies between 1 and 3
- Model does not capture the temporal decay of preference of a node
- Model does not consider the competitive characteristics of a real-world node that flourish in short notice

Price's Network Model

- A mathematical model for the growth of citation networks
- Named after the physicist cum information scientist Derek J. de Solla Price
- Inspired by the ideas of the Simon's model (on wealth distribution in a society) that reflects the concept rich gets richer
- Price was the first scientist to apply Simon's model to network science
- The concept is also known a Mathew effect
- The notion was referred to as cumulative advantage, currently known as preferential attachment
- Price's idea has been: *the way an existing paper gets new citations should be proportional to the number of existing citations the paper already has*

Price's Network Model: Network Formation

- Assumption: Each new node has a given out-degree and it is fixed in the long run
- Out-degree of individual nodes may vary, however, their mean value, m (say) is fixed over time
- The mean in-degree of the nodes in also m
- The above implies:

$$\sum_k k.p_k = m$$

where p_k is the fraction nodes having degree k

-

Price's Network Model: Network Formation

- According to Price's Model,

$$Prob(A\ new\ node\ attaching\ an\ old\ node) \propto Indegree(Old\ node)$$

- What would happen at the start of simulation? All the nodes have indegree = 0 at $t = 0$!

- To overcome this, Price proposed:

$$Prob(A\ new\ node\ attaching\ an\ old\ node) \propto k + k_0$$

- In general, k_0 is an arbitrary constant; however, Price let $k_0 = 1$, with the justification that an initial citation is associated with the paper itself

- So, probability of a new edge connecting to any node with a degree k is

$$P(k) = \frac{(k+1)p_k}{\sum_k (k+1)p_k} = \frac{(k+1)p_k}{m+1}$$

Price's Network Model: Network Evolution

- We let a new node with outdegree m join the network
- The mean number of edges connected to nodes of degree k

$$mP(k) = \frac{(k+1)mp_k}{m+1}$$

- Let us measure the net change in the number of nodes with degree k when we add new nodes to the network
- On adding new nodes to the network, some k-degree nodes have new edges, hence becoming (k + 1)-degree nodes; some (k − 1)-degree nodes might get new edges, becoming k degree nodes.
- Then, the net change is

$$(n+1)p_{k,n+1} - np_{k,n}$$

where $p_{n,k}$ denote the fraction of k-degree nodes at a network with n vertices

Price's Network Model: Network Evolution

- The net change in the count of nodes with degree k

$$(n+1)p_{k,n+1} - np_{k,n} = \begin{cases} [kp_{k-1,n} - (k+1)p_{k,n}]\dfrac{m}{m+1}, k \geq 1 \\ 1 - p_{0,n}\dfrac{m}{m+1}, k = 0 \end{cases} \quad \cdots\cdots\cdots (\#)$$

- We arrive at the stationary solution, we let $n \to \infty$, to get

$$p_{k,n+1} \approx p_{k,n} \approx p_k$$

- With the above, equation (#) yields,

$$p_k = \begin{cases} [kp_{k-1} - (k+1)p_k]\dfrac{m}{m+1}, k \geq 1 \\ 1 - p_0\dfrac{m}{m+1}, k = 0 \end{cases} \quad \cdots\cdots\cdots (\#\#)$$

Price's Network Model: Network Evolution

- With little manipulation, and using Beta function, we obtain

$$p_0 = \frac{1 + \frac{1}{m}}{2 + \frac{1}{m}}$$

$$p_k = \left(1 + \frac{1}{m}\right)\beta\left(k + 1, 2 + \frac{1}{m}\right)$$

- Solving,

$$p_k = \left(1 + \frac{1}{m}\right)(k + 1)^{-(2+\frac{1}{m})}$$

$$p_k \sim k^{-(2+\frac{1}{m})}$$

- Theorem: Degrees of the nodes in Price's model follow a power-law distribution

Price's Network Model: Generalization

- Considers the situation when $k_0 \neq 1$
- Mimicking the calculations, we may find

$$p_k = \frac{m + k_0}{m(k_0 + 1) + k_0} \times \frac{\beta(k + k_0, 2 + \frac{k_0}{m})}{\beta(k_0, 2 + \frac{k_0}{m})} \cdots\cdots\cdots (\#\#\#)$$

- On simplification, equation (###) yields to a power law distribution of p_k with exponent

$$\alpha = 2 + \frac{k_0}{m}$$

for large k and fixed k_0

Preferential Attachment Models: Shortcomings

- Preferential attachment models are not the only models that exhibit power-law degree distribution

- Cumulative advantage model is enough to justify protein-protein interactions

- Kleinberg et al. 1999 suggested alternative mechanism, Vertex Copying Model

- The model is a variation of the Price's network model

Vertex Copying Model: Network Formation

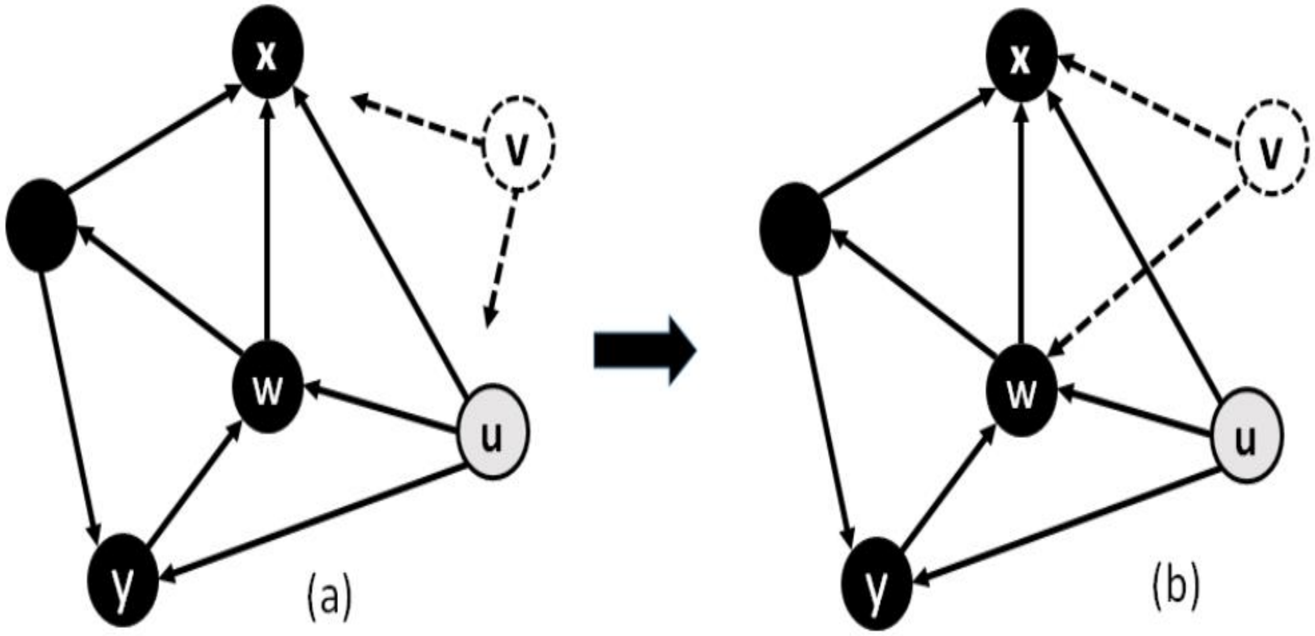


Illustration: Copying a vertex

- The Philosophy of the model: authors are copying references from the bibliographies of papers they read
- To add a vertex of degree m
  - Instead of choosing a node preferentially, an existing node is chosen uniformly at random from the network
  - Chooses uniformly at random from the previously chosen vertex m many links
  - Replicates the chosen links with the target vertex
- If the chosen node u does not have m connections, more than one nodes may be chosen uniformly at random

Vertex Copying Model: Properties

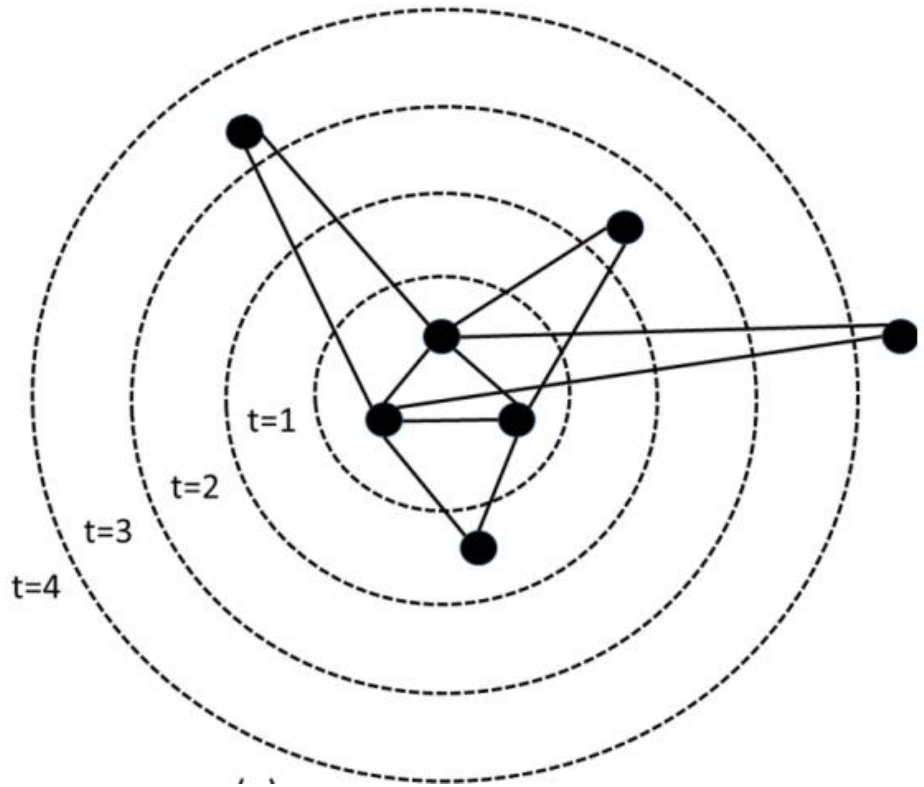
- Model results a power-law degree distribution
- There are a number of variants of the model
- useful to synthesize protein interaction network, citation networks, web networks, etc.

Other models

Local-world Network Growth Model

- All the models introduced so far have a tacit assumption: any newly added point has a global information about the whole network
- Availability of such global knowledges are rare for large networks
- Decisions about links are made in a local world rather than as a global decision
- Such network models are coined as Local-world models
- Discussed first by Li and Chen [2003]
- Principle: while calculating the preferential attachment of a node, we should not consider the global context of a network; rather we should focus on the surrounding context of a node

Network Model with Accelerating Growth: Super-linearity



Network growth process in Preferential Attachment Process

- Growth of the network is linear in Preferential Attachment model
- The assumption of linearity is not at all realistic
- Many real networks often exhibit a kind of accelerating growth feature
- WWW, citation networks, author-collaboration networks are typical examples
- Super-linear growth

Network Model with Accelerating Growth: Aging Effect

- Newer nodes tend to ignore older popular nodes to establish the connections
- Contemporary emerging nodes are preferred over popular older nodes
- Older nodes fail to receive links from newly-added nodes after a while
- A significant deviation from the philosophy of the preferential attachment

Aging in Preferential Attachment Model: Time-dependent Network Models

- Adopted from Barabási-Albert Network Model
- Preferential attachment probability is dependent on the degree and the age of the existing nodes.
- General form of preferential attachment probability:

$$\Pi(k, t)$$

Here k denotes the degree of the node, and t denotes the age of the node

- Assuming a separable functional form:

$$\Pi(k, t) = K(k) \cdot f(t)$$

Short-term Memory Preferential Attachment Mechanism

- The attachment probability

$$\Pi(k,t) \sim k_1$$

where k_1 denotes the number of citations received in the recent one year

- Past citations, except only the most recent past, are considered as useless in the computation

END