Random Graph models of Social Networks

ENAYAT ULLAH

Overview

- Networks
- Social Network Metrics:
  - Small-World, Scale-free, Centrality
- Social Network examples:
  - Erdos Number, Kevin Bacon, Facebook
- Erdos-Renyi Random Graph
- Phase Transition in ER random graph
- Configuration model.
- Preferential Attachment model.

Networks

- A network is a set of items (nodes or *vertices*) connected by *edges* or links.
- Types of networks:
  - *Technological networks:* Internet, phone networks, power grid.
  - *Social Networks* : Collaboration network, Facebook, Kevin Bacon Number.
  - *Biological Networks*: Protein Interaction network, Neural Network
- Social Networks
  - Vertices are people, and edges are relations between them.

Random Graph

- Random Graph is the general term to refer to probability distributions over graphs.
- Graph Sequence: A graph sequence is denoted by by $(G_n)_{n \geq 1}$, where n denotes the size of the graph G_n, i.e., the number of vertices in G.
- Degree Distribution: $P_k^{(n)}$ denotes the proportion of vertices with degree k in G_n,

$$P_k^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{d_i^{(n)} = k\}}$$

- Two types of degree distributions:
  - $P(d) = ce^{-\alpha d}$: The distribution falls off as fast as an exponential, for some a,c > 0.
  - $P(d) = cd^{-\lambda}$: Power law distribution (Scale-free graph)

Scale-free and Sparse Graph

- Scale-free Graphs:
  - We call a graph sequence $(G_n)_{n\geq 1}$ scale free with exponent τ when it is sparse and when

$\lim_{k\to\infty} \frac{\log[1-F(k)]}{\log(1/k)} = \tau - 1,$ where $F(k) = \Sigma_{l\leq k}\, p_l$ denotes the cumulative distribution function corresponding to the probability mass function $(p_k)_{k\geq 0}$.

  - Or, $\lim_{k\to\infty} \frac{\log p_k}{\log(1/k)} = \tau.$ i.e the log-log plot is linear.

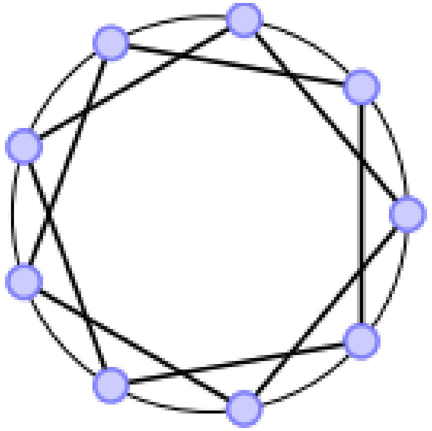
- Sparse Graph:
  - A graph sequence $(G_n)_{n\geq 1}$ is called sparse when $\lim_{n\to\infty} P_k^{(n)} = p_k,$ for some deterministic limiting probability distribution $(p_k)_{k\geq 0}$.

Small-World Graph

- Small-World
  - Vertices are separated by relatively short chains of edges.
  - A Graph Sequence (Gn)n>=1 is a small-world graph sequence, when its typical distances satisfy that there exists a constant K such that:

$$lim_{n\to\infty}P(H_n \leq klogn) = 1$$

where H_n is the average path length.



Clustering

- Clustering measures the degree to which neighbours of vertices are also neighbours of one another.
- *Clustering Coefficient:* Clustering coefficient measures the proportion of wedges for which the
closing edge is also present.
- *CC(g)* = $\frac{3\ x\ Number\ of\ Triangles\ in\ the\ graph}{Number\ of\ connected\ triples\ of\ nodes}$
- Definition: A *graph sequence* $(G_n)_{n \geq 1}$ *is* highly clustered when $lim_{n \to \infty}$CC > 0

Highly Connected

- Highly Connected:
  - A large part of the vertices is in one large connected component.
  - For a graph G = ([n]*,* E) on n vertices and $v \in [n]$, let $C(v)$ denote the *cluster* or *connected component* of $v \in [n]$,
    i.e., $C(v) = \{u \in [n]: \text{dist}G(u, v) < \infty\}$
  - A graph sequence (Gn)n>=1 is called highly connected when :

    $$lim_{n\to\infty}|C_{max}|/n > 0$$
  - Furthermore, for a highly-connected graph sequence, the *giant component* is unique when

    $lim_{n\to\infty}|C_2|/n > 0,$ $|C_2|$ being the size of second-largest component.

Centrality

- A measure that captures the importance of a node’s position in the network.
  - Closeness Centrality: vertices that are close to many other vertices are deemed to be important.

$$C_i = \frac{n}{\sum_{j \in [n]} \mathrm{dist}_G(i, j)},$$

  - Betweeness Centrality: Vertices lying in the shortest paths between any two vertices are deemed important.

$$b_i = \sum_{1 \le j < k \le n} n_{jk}^i / n_{jk},$$

n_{jk} : number of shortest paths between vertices j and k.

n_{jk}^i: number of shortest paths between vertices j and k containing i.

Empirical Data

- Empirically, the properties exhibited by real-world social networks are:
  - Small- world networks
  - Scale-free networks
  - Existence of a giant component.

Social Networks: Examples

- Six Degrees of Separation [http://www.stanleymilgram.com/milgram.php]
  - Average length between vertices is 6.
  - Established Small-World network.
- Facebook [Ugander et al]
  - 99.91% of the active Facebook users is in the giant component, so that Facebook is indeed very highly connected.
  - The second largest connected component consisting of a meagre 2000 some users. The assortativity coefficient is equal to 0.226
  - This distribution does not resemble a power law (owing to the limit of 5000 friends per person).

Social Networks: Examples

- Kevin Bacon Number [http://www.cs.virginia.edu/oracle/]
  - The vertices are movie actors, and two actors share an edge when they have been cast in the same movie.
  - Turn out Kevin Bacon is not the most central vertex in the graph. A more central actor is Sean Connery.
  - A Scale-free distribution, with the power law exponent equal to 2.3.

- Erdos Number[http://www.ams.org/msnmain/cgd/index.html]
  - The vertices are mathematicians, and there is an edge between two mathematicians when they have co-authored a paper.
  - One giant component consisting of about 268,000 vertices, so that the graph is highly connected.
  - The average number of collaborators per person is 3.36.

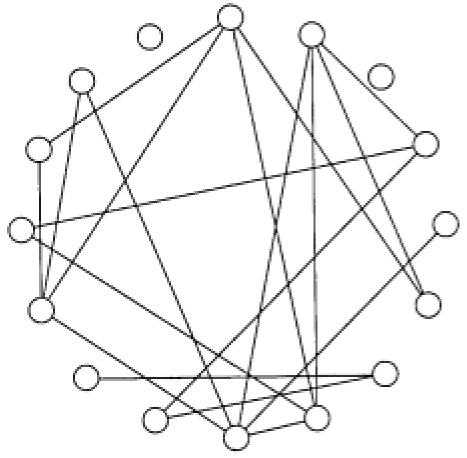
Erdős-Renyi Random Graph Model

- G(n, p) denotes an undirected Erdős-Renyi graph.
- Every edge is formed with probability p ∈ (0, 1) independently of every other edge.
- Let $I_{ij} \in \{0, 1\}$ be a Bernoulli random variable indicating the presence of edge {i, j}
- For the Erdos-Renyi model, random variables I_{ij} are independent and

$$l_{ij} = \begin{cases} 1 & with\ probability\ p \\ 0 & with\ probability\ 1-p \end{cases}$$

- E[number of edges] = $E[\sum l_{ij}] = \frac{n(n-1)}{2} p$

N=16, p=1/7



Erdős-Renyi Random Graph Model

- Let *D* be a random variable that represents the degree of a node.
- *D* is a binomial variable with E[D] = n(n-1)p
- $P(D = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d}$
- As n → ∞, D can be approximated with a Poisson random variable with $\lambda = (n - 1)p$.
  - Since this degree distribution falls off faster than an exponential in d, hence it is not a power-law distribution.

Phase Transition

- For a given property A (e.g. connectivity), we define a threshold function t(n) as a function that satisfies:

$$P(property\ A) \to 0\ if\ \frac{p(n)}{t(n)} \to 0\text{, and}$$

$$P(property\ A)\ \to 1\ if\ \frac{p(n)}{t(n)} \to \infty$$

- This definition makes sense for “monotone or increasing properties,” i.e., properties such that if a given network satisfies it, any super network (in the sense of set inclusion) satisfies it.
- When such a threshold function exists, we say that a phase transition occurs at that threshold.

Phase Transition Example

- Let property A be = {number of edges > 0}
- We are looking for a threshold *t(n)* for the emergence of the first edge.
- E[number of edges] = $\frac{\mathrm{n(n-1)}}{2} p(n) \approx \frac{n^2}{2} p(n)$
- Assuming $\frac{p(n)}{n^2} \to 0$ as n $\to \infty$. Then, E[number of edges]$\to$ 0, which implies that P(number of edges > 0) $\to$ 0.
- Similarly, we next sssume that $\frac{p(n)}{n^2} \to \infty$ as n $\to \infty$. Then, E[number of edges]$\to \infty$.
- Since, the number of edges can be approximated by a Poisson distribution, we have

$$\mathbb{P}(\text{number of edges} = 0) = \left.\frac{e^{-\lambda}\lambda^k}{k!}\right|_{k=0} = e^{-\lambda}$$

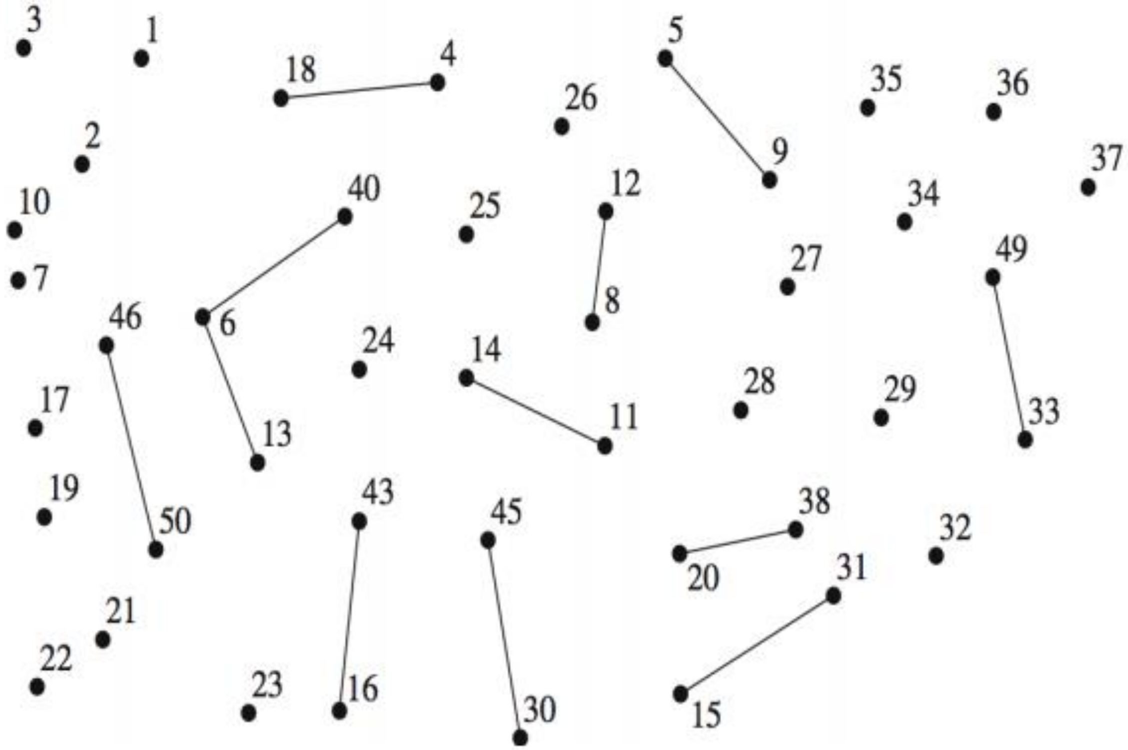
Phase Transition Example

- Since, λ= E[number of edges] $\rightarrow \infty$
- P(number of edges = 0) = $e^{-\lambda} \rightarrow 0$

Phase Transition

- t(n) = $\frac{1}{n^2}$ is the threshold function for the emergence of the first link.
- t(n) = $\frac{1}{n^{\frac{3}{2}}}$ is the threshold function for the emergence of triples in the graph.
- t(n) = $\frac{1}{n^{\frac{k}{k-1}}}$ is the threshold function to start observing a tree with k nodes.
- t(n) = $\frac{1}{n}$ is the threshold function to start observing a cycle.
- Above the threshold of 1/n, a giant component emerges, which is the largest component that contains a nontrivial fraction of all nodes, i.e., at least cn for some constant c.
- The giant component grows in size until the threshold of log(n)/n, at which point the network becomes connected.

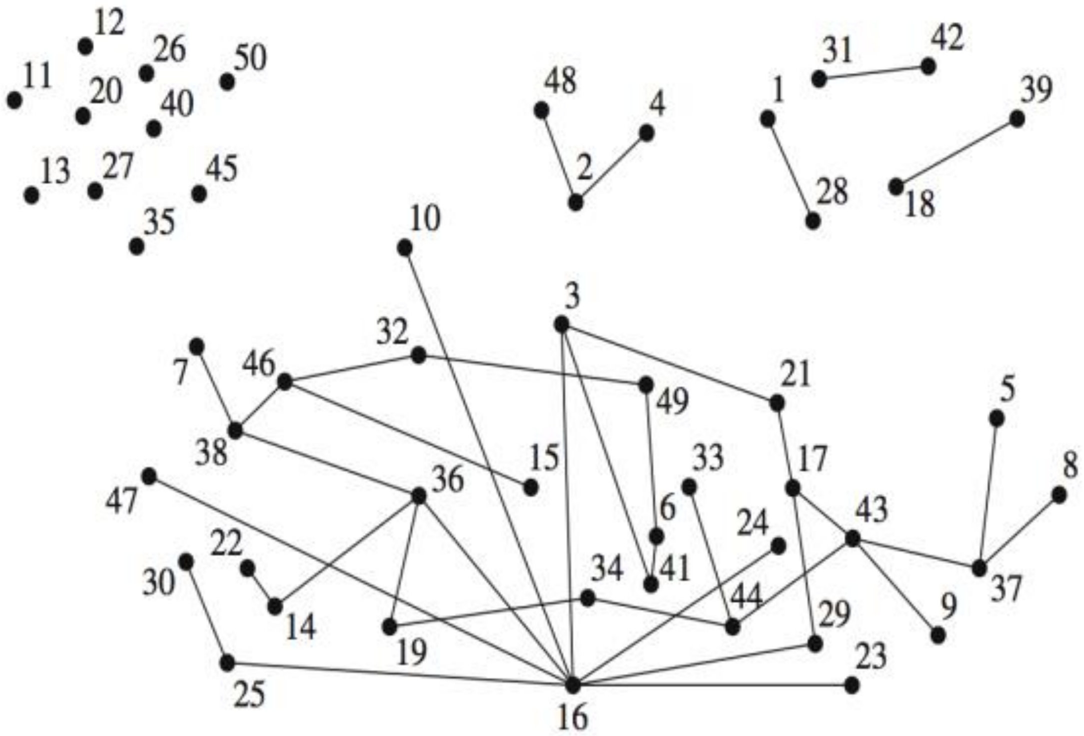
Phase Transition



N = 50
p = 0.01

Emergence of a first component with more than two nodes a random network.

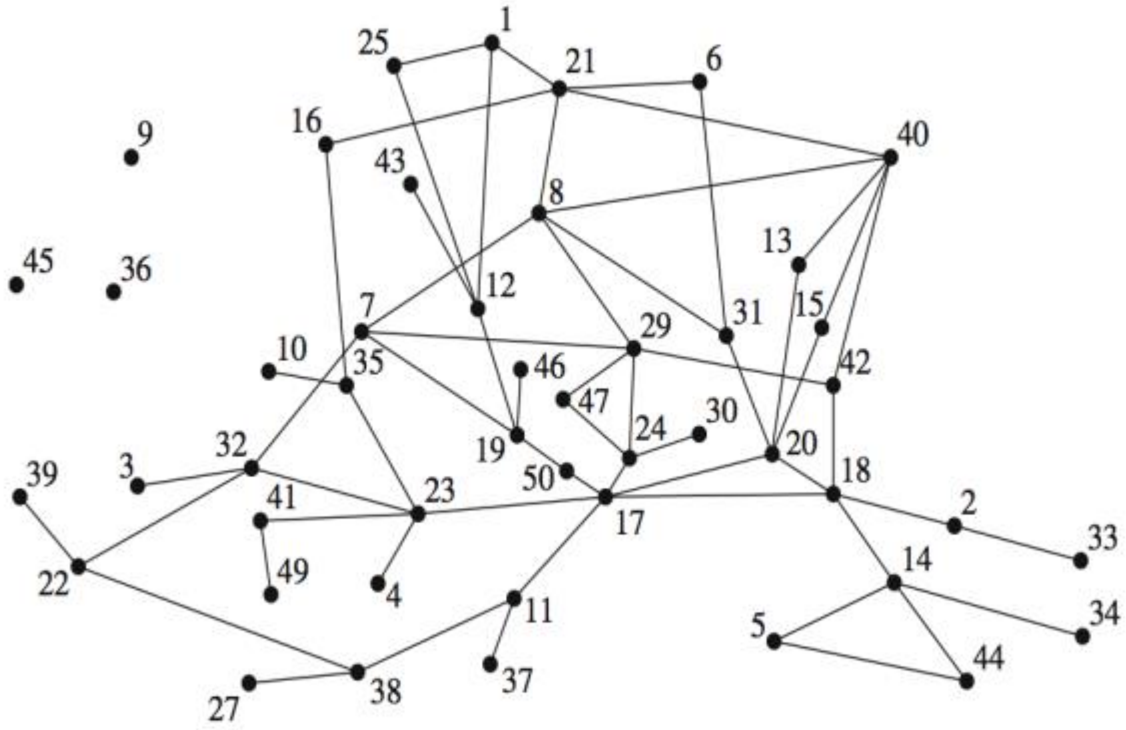
Phase Transition



N = 50
p = 0.03

Emergence of cycles.

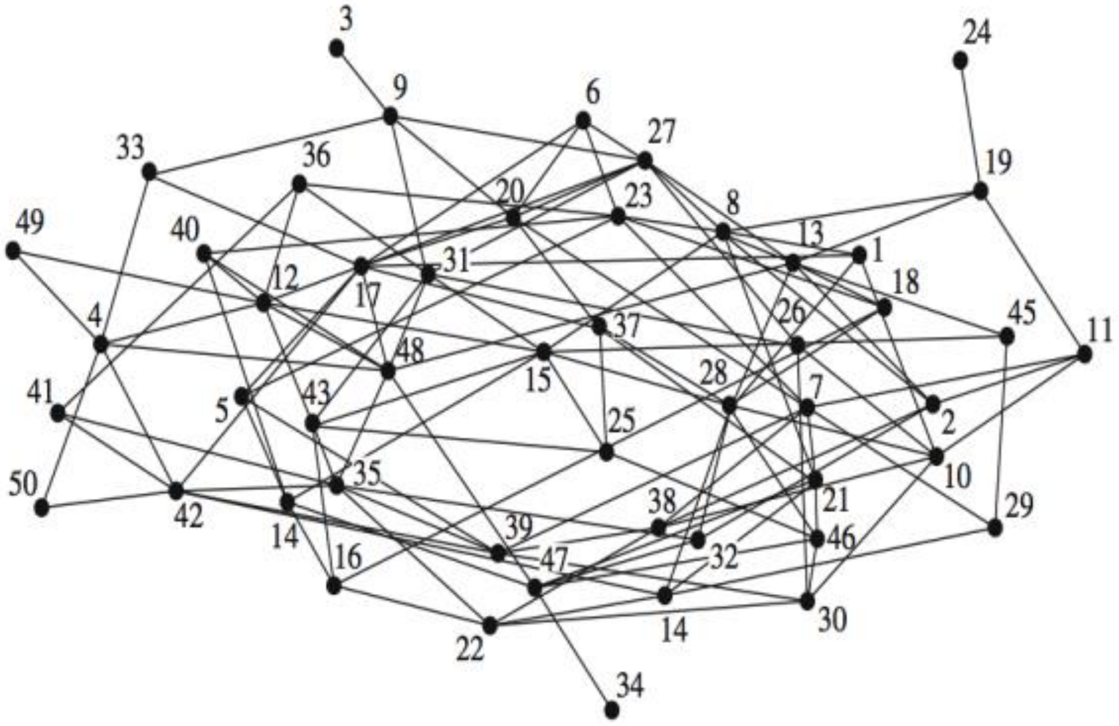
Phase Transition



N = 50
p = 0.05

Emergence of a giant component.

Phase Transition



N = 50
p = 0.10

Emergence of connectedness.

Connectivity

- Theorem: (Erdos and Renyi 1961) A threshold function for the connectivity of the Erdos and Renyi model is $t(n) = \frac{log(n)}{n}$.
- Proof:

  We show that when $p(n) = \lambda \frac{log(n)}{n}$,

  If $\lambda < 1$, P(connectivity) $\rightarrow 0$,

  If $\lambda > 1$, P(connectivity) $\rightarrow 1$
- To prove disconnectedness, it is sufficient to show that the probability that *there exists at least one isolated node* goes to 1.

Connectivity

- Let I_i be a Bernoulli random variable defined as

$$I_i = \begin{cases} 1 & \text{if node i is isolated} \\ 0 & \text{otherwise} \end{cases}$$

- The probability that an individual node is isolated as

$$q = P(I_i = 1) = (1-p)^{n-1} \approx e^{-pn} = e^{-\lambda \log(n)} = n^{-\lambda}$$

(Using $\lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^n = e^{-a}$ to get the approximation)

- Let X = $\sum_{i=1}^{n} l_i$ denote the total number of isolated nodes. We have,

$$\mathbb{E}[X] = n \cdot n^{-\lambda}.$$

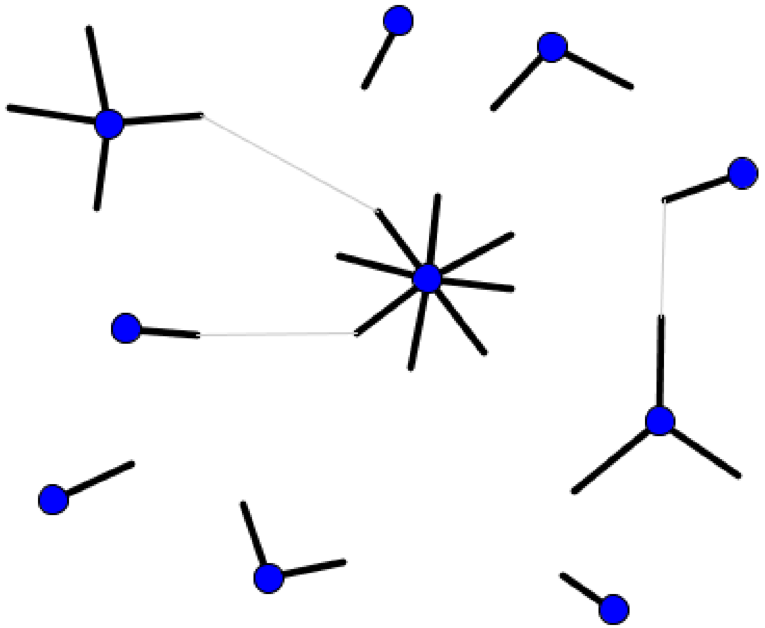
Connectivity

- For $\lambda < 1$, we have $E[X] \to \infty$. This implies $P(X = 0) \to 0$.
- It follows that P(at least one isolated node) $\to 1$ and therefore, P(disconnected) $\to 1$ as $n \to \infty$.

Configuration Model

- Configuration model is a generalized Erdos-Renyi random graph with a “ given degree distribution”(Bender and Canfield, 1978).
- The configuration model is specified in terms of a degree sequence.
- Given (d1, . . . , dn), we construct a sequence where node 1 is listed d1 times, node 2 is listed d2 times, and so on: 1, 1, 1, 1, . . . , 1 | {z } d1 entries 2, 2, . . . , 2 | {z } d2 entries · · · n, n, n . . . , n | {z }
- Each node i in the graph can be thought of as “stubs” sticking out of it, which are ends of edges-to-be.
- We randomly pick two elements of the sequence and form a link between the two nodes corresponding to those entries.
- t(n) = 1/n is the threshold for the emergence of the giant component.

Configuration model



Remarks:

- The sum of degrees needs to be even.
- Self-loops are possible.
- More than one edge between two vertices is possible (multigraph).

- Generating Graphs with arbitrary degree distribution.
- Half-edges(stubs) joined

Preferential Attachment

- Erdos-Renyi, Configuration model are all static models, in which edges among “fixed” n nodes are formed via random rules .
- In a preferential attachment model, nodes are born over time, therefore a dynamic model.
- Each node upon birth forms m edges with pre-existing nodes.
- Let di (t) be the degree of node i at time t. Initially we have m+1 nodes(indexed 0,…m) all connected to each other
- The probability that an existing node i receives a new link to the newborn node at time t is m times i’s degree relative to the overall degree of all existing nodes at time t, i.e

$$\frac{dd_i(t)}{dt} = m\frac{d_i(t)}{\sum_{j=1}^{t} d_j(t)}$$

Preferential Attachment

- Since at a time t, there are 2tm edges. Therefore, $\sum_{j=1}^{t} d_j(t) = \mathbf{2tm}$.

$$\frac{dd_i(t)}{dt} = \frac{d_i(t)}{2t}, \text{ with initial condition } d_i(t) = m$$

- The solution to the equation is : $d_i(t) = m(\frac{t}{i})^{1/2}$
- Let i(d) be $\frac{i(d)}{t} = \left(\frac{m}{d}\right)^2$, has degree d at time t, or $d_{i(d)}(t) = d$
- Therefore,
- For any d and any time t, let i(d) be a node such that $d_{i(d)}(t) = d$The resulting cumulative distribution function then is $F_t(d) = 1 - \frac{i(d)}{t}$.
- In this case, F (d) = 1 − $m^2 d^{-2}$
- P(d) = $2m^2 d^{-3}$, therefore scale-free(Power-law with exponent -3)