Advanced Techniques for Mining Structured Data:

Graph Mining

Basics/ Measures/ Models

Dr C.Loglisci

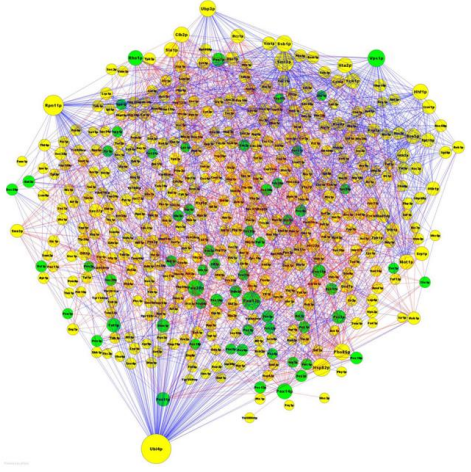
PhD Course in Computer Science and Mathematics XXXII cycle

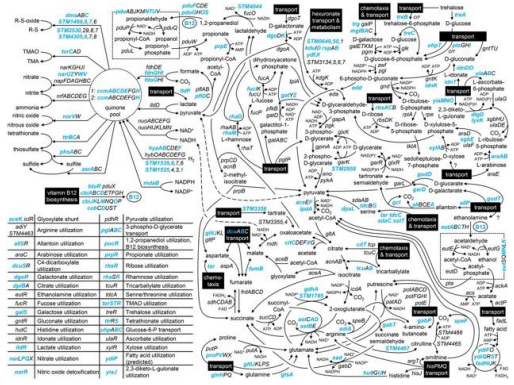
Networks

Social graphs, Knowledge graphs, Biological networks, Metabolic networks

They are complex: groups, links, preferences, attributes

- Connect entities such as persons, organizations, countries, objects through explicit relationships
- Protein-protein interaction networks: proteins, physical interactions
- Metabolic networks: metabolites and enzymes,chemical reactions





facebook

- 1.5 Bln users
- 450 Bln Relationships
- 600 Mln groups
- 10.5 USD per user

Twitter

- 313 Mln users
- 500 Mln Tweets/day
- Avg 208 followers/user

WIKIDATA

- 20Mln entities
- 100Mln relationships
- 2500 types of relationships

Other knowledge graphs:

- YAGO
- DBPedia
- DBLP
- Pubmed
- Linkmdb
- ...

Anything that involves relationships can be modeled as a graph!

Networks

Why Networks?

Describe complex data with a simple structure

Nature, social, concepts, roads, circuits …

Same representation for many disciplines

Computer science, biology, physics, economics, ...

Availability of (BIG) data

Large networks are now available and require complex algorithms

Networks are evolving over time (e.g., new users/friends in Facebook)

Usefulness

Analysis will discover non trivial patterns, and allow simple smooth explorations

They reveal user behaviors

They are valuable (Facebook, Twitter, Amazon ... All of them based on graphs!!!)

Networks

Networks or Graphs?

Network refers to real systems

Web, Social, Biological, …

Terminology: Network, node, link/relationship

Graph is an abstract mathematical model of a network

Web graph, Social graph

Terminology: Graph, vertex/node, edge

BUT

we often use both without distinction

Networks and Their Representations

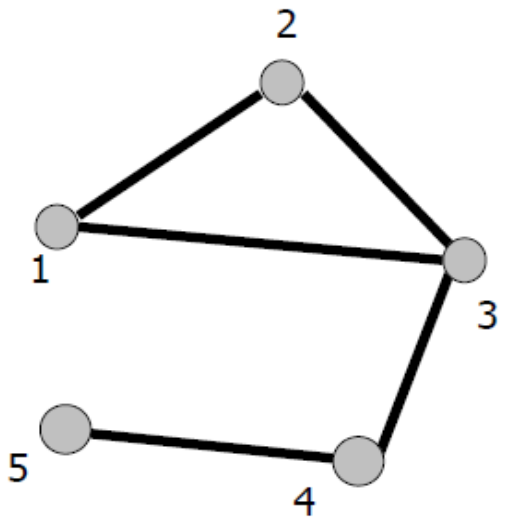
A network/graph: G = (V, E), where V: vertices/nodes, E: edges/links

E: a subset of $V \times V$, n = |V| (order of G), m = |E| (size of G).

Often, we have sets of labels, each associated to nodes and edges.

Adjacency matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$



List of edges

(1,2)
(2,3)
(1,3)
(3,4)
(4,5)

Adjacency list

1: [2, 3]
2: [1, 3]
3: [1, 2, 4]
4: [3, 5]
5: [4]

Networks and Their Representations

Various kinds of networks:

- Simple network: if a network has neither self-loop nor multi-edges
- Multi-edges: if more than one edge between the same pair of vertices
- Self-loop: if an edge connects vertex to itself (i.e., (v_i, v_i))
- Weighted graph: If a weight w_{ij} (a real number) is associated with each edge v_{ij}
- Undirected graphs: e.g., *Co-authorship, Roads, Biological*
- Directed graph (digraph): if each edge has a direction (tail → head) e.g., *Follows*
  - A_{ij} = 1 if there is an edge from j to i; 0 otherwise

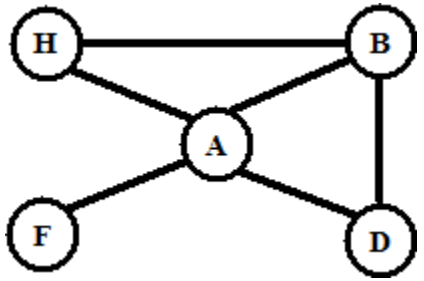
Basic Network Structures

Walk in a graph G between vertices X and Y: sequence of vertices, starting at X and ending at Y, s.t. there is an edge between every pair of consecutive vertices

Hops: the length of the walk

Path: a walk with distinct (non-repeating) vertices. A sequence of vertices that every consecutive pair of vertices in the sequence is connected by an edge in the network

Length of a path: # of edges traversed along the path



Graph G_1

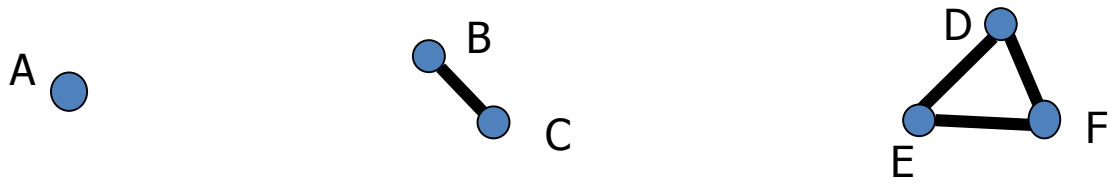
Basic Network Structures

Subgraph: Given G = (V, E) and a subset of vertices V' $\subseteq$ V, the **induced subgraph** G' = (V', E') consists exactly of all the edges present in G between vertices in V'

Connected: two vertices are endpoints of a path

Clique (complete graph): Every node is connected to every other

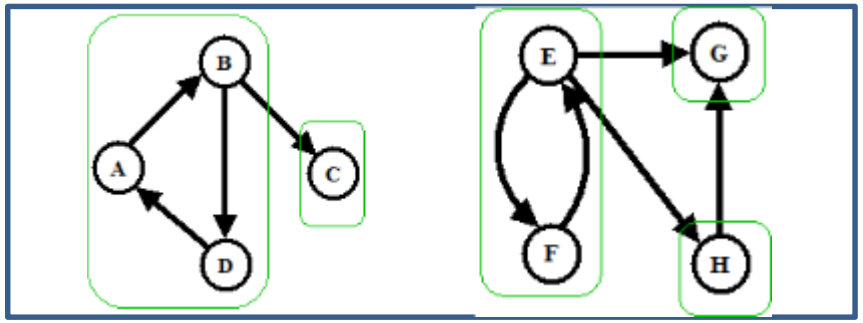
Singleton, dyad (two vertices and their relationship), triad :



Neighborhood of a vertex: the subgraph induced by all vertices that are adjacent (neighbors) to the vertex

Basic Network Structures

Connected Component: A subgraph of a graph such that there exists at least one path from vertex to each other vertex. There are no edges with other vertices of the whole graph



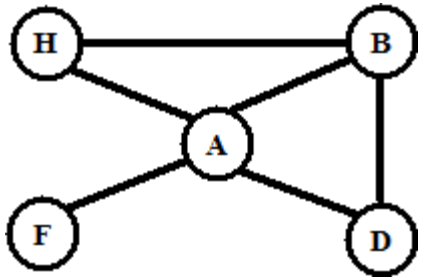
Adjacency matrix of a network with more than one component can be written in block diagraph form

$$A = \begin{pmatrix} [\,] & 0 & \cdots \\ 0 & [\,] & \cdots \\ \vdots & \cdots & \ddots \end{pmatrix}$$

Vertex Degree for Undirected & Directed Networks

Let a network G = (V, E)

Undirected Network $d(v_i) = |v_j| \ s.t. \ e_{ij} \in E \wedge e_{ij} = e_{ji}$



Degree (or degree centrality) of a vertex: $d(v_i)$

# of edges connected to it, e.g., d(A) = 4, d(H) = 2

Directed network

In-degree of a vertex $d_{in}(v_i)$:

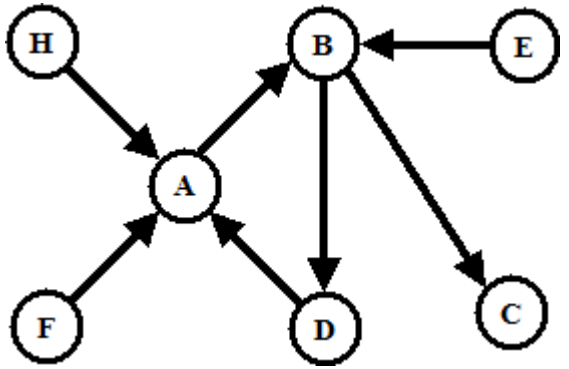
# of edges pointing to v_i $d_{in}(v_i) = |v_j| \ s.t. \ e_{ij} \in E$

E.g., $d_{in}(A) = 3$, $d_{in}(B) = 2$

Out-degree of a vertex $d_{out}(v_i)$:

# of edges from v_i

E.g., $d_{out}(A) = 1$, $d_{out}(B) = 2$



$$d_{out}(v_i) = |v_j| \ s.t. \ e_{ji} \in E$$

Degree Distribution

- Degree sequence of a graph: The list of degrees of the vertices sorted in non-increasing order

  E.g., in graph G_1, degree sequence: (4, 3, 2, 2, 1)

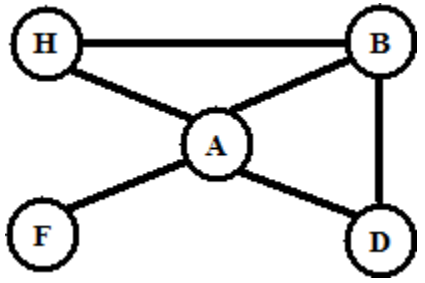
- Degree frequency distribution of a graph: Let N_k denote the # of vertices with degree k(N_0, N_1, …, N_t), t is max degree for a node in G

E.g., in graph G_1, degree frequency distribution: (0, 1, 2, 1, 1)

- Degree distribution of a graph,

probability distribution (f(0), f(1), …, f(t), where $f(k) = P(X = k) = N_k/n$

E.g., in graph G_1, degree distribution (0, 0.2, 0.4, 0.2, 0.2)

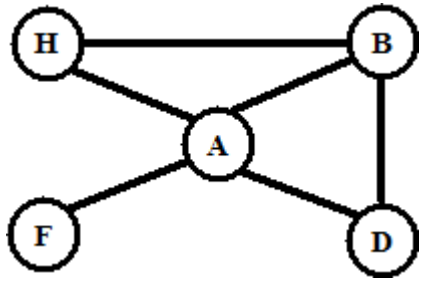


Graph G_1

Various Kinds of Paths

Shortest path (geodesic path, *d*):

Geodesic paths are not necessarily unique: It is quite possible to have more than one path of equal length between a given pair of vertices



For this graph, what is <d> ?

Distance: the length of the shortest path between two vertices

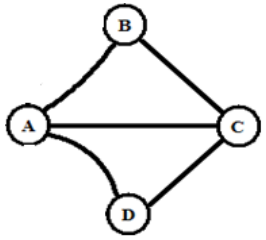
Average path length (<d>):

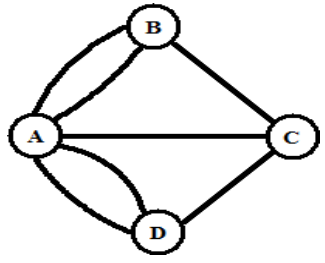
Average of the shortest paths between all pairs of vertices

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j=1,N(i \neq j)} d_{i,j}$$

Eulerian path: a path that traverses each edge in a network exactly once

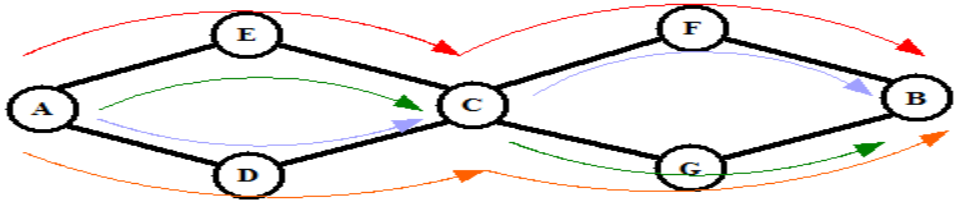
Hamilton path: a path that visits each vertex in a network exactly once

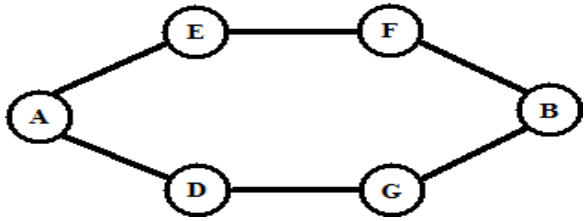




Independent Paths, Connectivity, and Cut Sets

Two path connecting a pair of vertices (A, B) are edge-independent if they share no edges
Two path are vertex-independent if they share no vertices other than the starting and ending vertices





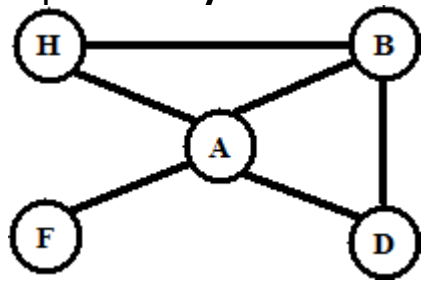
A vertex cut set is a set of vertices whose removal will disconnected a specified pair of vertices
An edge cut set is a set of edges whose removal will disconnected a specified pair of vertices

Radius and Diameter of a Network

Eccentricity: The eccentricity of a node v_i is the maximum distance from v_i to any other vertices in the graph

$e(v_i) = \max_j \{d(v_i, v_j)\}$

E.g., e(A) = 1, e(F) = e(B) = e(D) = e(H) = 2



Graph G_1

Radius of a connected graph G: the min eccentricity of any node in G

$r(G) = \min_i \{e(v_i)\} = \min_i \{\max_j \{d(v_i, v_j)\}\}$

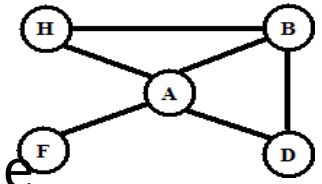
E.g., $r(G_1) = 1$

Diameter of a connected graph G: the max eccentricity of any node in G

$d(G) = \max_i \{e(v_i)\} = \max_{i,j} \{d(v_i, v_j)\}$

E.g., $d(G_1) = 2$

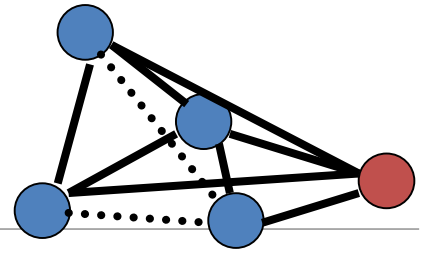
Radius and Diameter of a Network



Graph G_1

- Commonly, it is the length of the longest shortest path between any pair of vertices, that is, the maximum of the distances between pairs of vertices in the graph.
- If the graph has weights on its edges, then it is weighted by the sum of the edge weights along a path.
- Diameter is sensitive to outliers.
- Effective diameter: min # of hops for which a large fraction, typically 90%, of all connected pairs of vertices can reach each other

Clustering Coefficient



- Clustering coefficient of a node v_i (respectively, of a graph) :
  - A measure of the *density of edges* in the neighborhood of v_i (in a graph G)
- Let $G_i = (V_i, E_i)$ be the subgraph induced by the neighbors of vertex v_i, $|V_i| = n_i$ (# of neighbors of v_i), and $|E_i| = m_i$ (# of edges among the neighbors of v_i)
- (Local) Clustering coefficient of v_i for *undirected network* is $C(G) = \frac{1}{n}\sum_i C(v_i)$

$$C(v_i) = \frac{\#\ edges\ in\ G_i}{max\ \#\ edges\ in\ G_i} = \frac{2 \times m_i}{n_i(n_i - 1)}$$

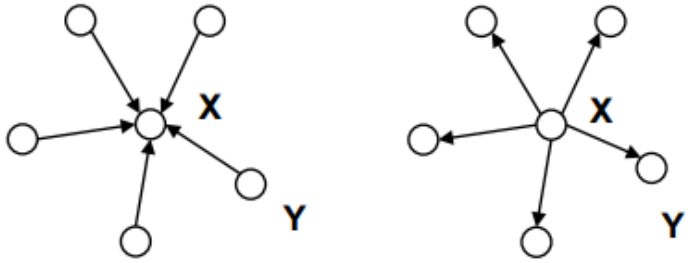
- For *directed network,*

$$C(v_i) = \frac{\#\ edges\ in\ G_i}{max\ \#\ edges\ in\ G_i} = \frac{m_i}{n_i(n_i - 1)}$$

- Global Clustering coefficient is computed on triads, instead of vertices

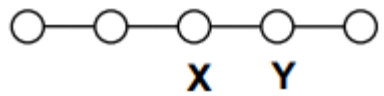
Centrality

- Centrality: How *central* a node is in the network
- Degree centrality: degree of a node (local measure)



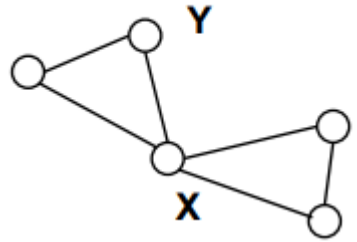
- Eccentricity centrality: the less eccentric, the more central (relative to rest of network)
  - $c(v_i) = 1/e(v_i)$
  - Central node: $e(v_i) = r(G)$ (if it equals the radius of G)
  - Periphery node: $e(v_i) = d(G)$ (if it equals the diameter of G)
  - Often used in facility location, e.g., emergency center

- Closeness centrality: indicates how close a node is to all other vertices in the network
  - $c(v_i) = 1/\sum_j d(v_i, v_j)$
  - Facility location, e.g., shopping center, minimize total distance



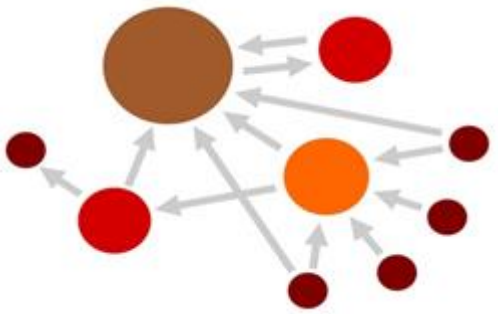
Centrality

- Betweenness centrality for a node v: # of shortest paths from all vertices to all others that pass through v
  - η_{jk}: # of shortest paths between vertices v_j and v_k
  - $\eta_{jk}(v_i)$: # of such paths that contain v_i
  - Betweenness centrality of a vertex v_i: $$c(v_i) = \sum_{j \neq i} \sum_{k \neq i, k > j} \frac{\eta_{jk}(v_i)}{\eta_{jk}}$$



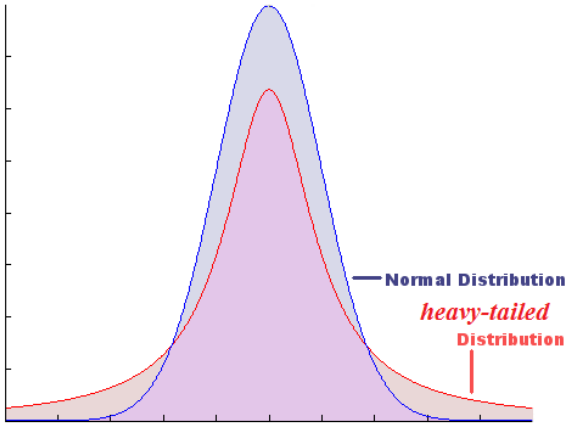
  - Indicating a central "monitoring role" played by v_i for various pairs of vertices

- Eigenvector centrality: Measure the influence of a node in a network, i.e., connections to high-scoring vertices contribute more to the score of the node in question than equal connections to low-scoring vertices



Network Modeling

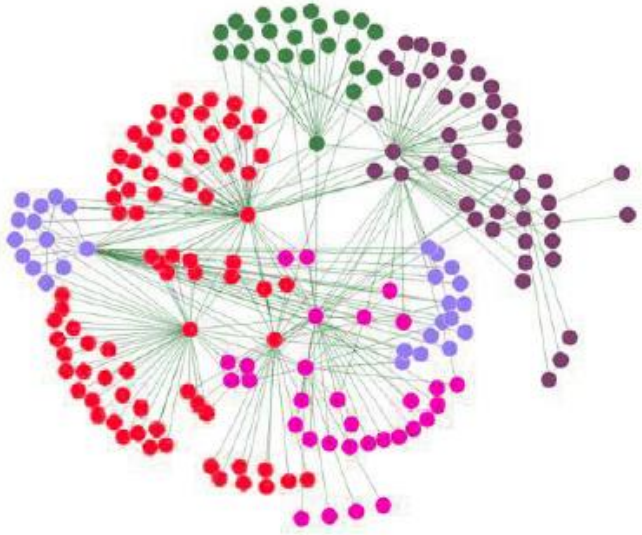
- A real-world network has the following common properties:
  - Few connected components:
    - often only 1 or a small number, independent of network size
  - Small diameter:
    - often a constant independent of network size
    - growing only logarithmically with network size or even shrink
    - typically exclude infinite distances
  - A high clustering coefficient
    - considerably more so than for a random network
  - A heavy-tailed degree distribution:
    - a small but reliable number of high-degree vertices
    - often of power law form



Network Modeling

In social networks:

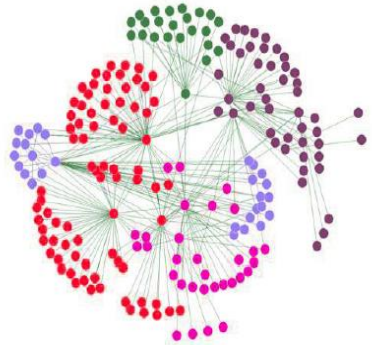
- Homophily: the tendency of individuals to associate and bond with others who are similar. Homophily shows that people's social networks are homogeneous with regard to many sociodemographic, behavioral, and intra-personal characteristics
- Selection: tendency of people to form friendships with others who are like them
- Socialization or Social Influence: the existing social connections in a network are influencing the characteristics of the individuals



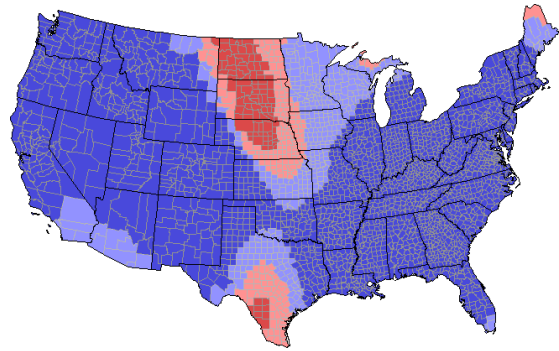
Network Modeling

In social networks:

- Homophily: the tendency of individuals to associate and bond with others who are similar. Homophily shows that people's social networks are homogeneous with regard to many sociodemographic, behavioral, and intra-personal characteristics
- Selection: tendency of people to form friendships with others who are like them
- Socialization or Social Influence: the existing social connections in a network are influencing the characteristics of the individuals

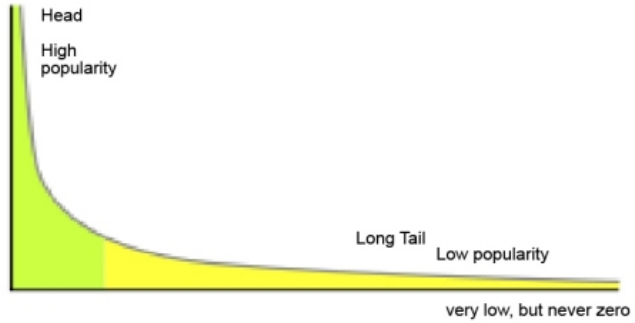


In spatial networks, this is equivalent to the spatial auto-correlation, where "everything is related to everything else, but near things are more related than distant things."



Network Modeling

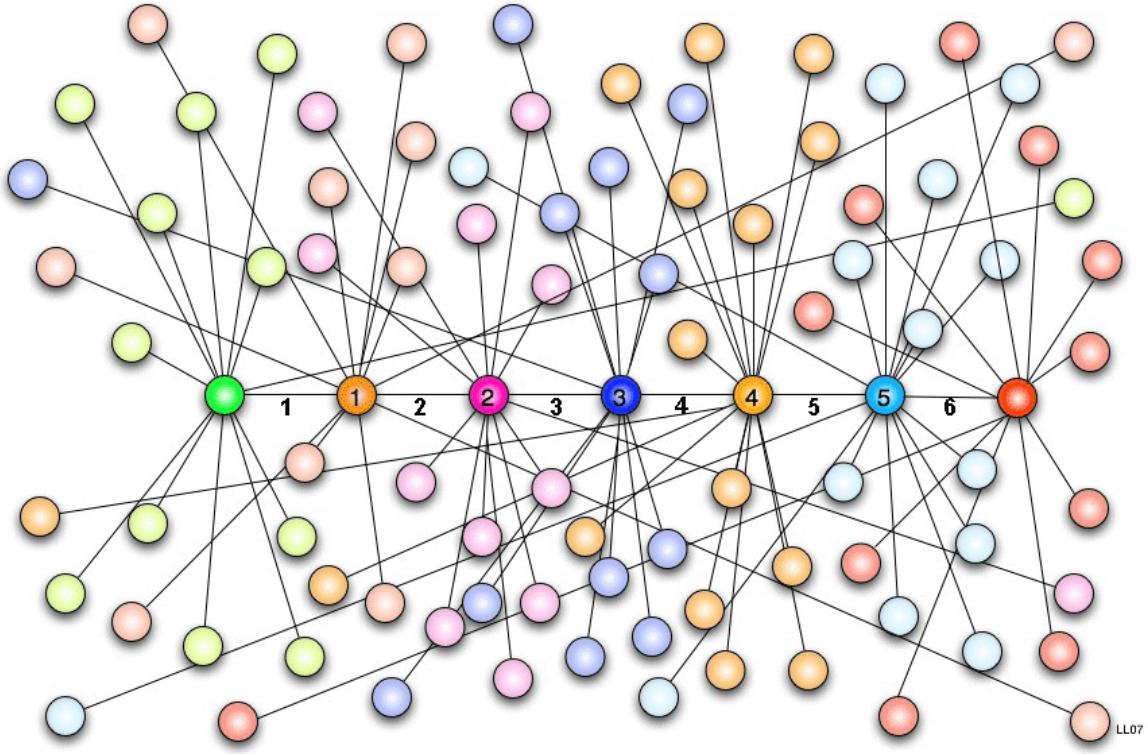
- Many real-world networks exhibit certain common characteristics, even though they come from different domains, e.g., communication, social, and biological networks
- Small-world networks
  - Small diameter
  - high clustering coefficient
- Scale-free networks
  - power law degree distribution
  - power law clustering coefficient distribution



small occurrences are extremely common, whereas large occurrences are extremely rare ("There are a few mega-cities, but many small towns")

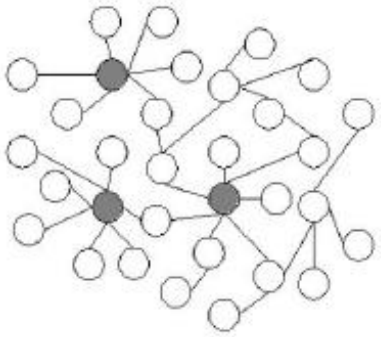
Small-world networks

- Six degrees of separation (Milgram experiment)
- 7-degrees of separation (Imdb actors)



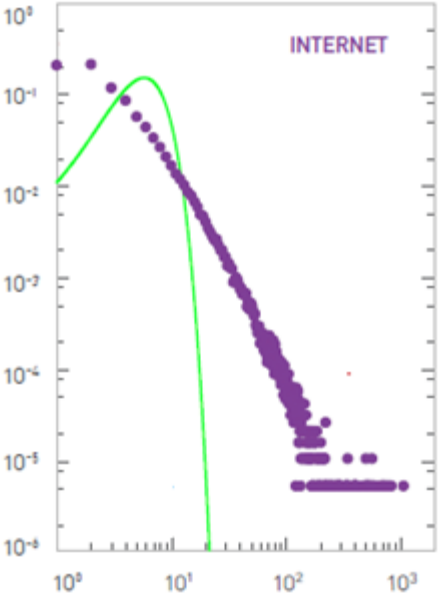
Scale-free networks

- hubs and communities
- hub refers to a vertex that connects to a lot of other vertices and communities
- low-degree vertices are members of dense groups (communities), which are connected to each other through hubs
- very few users are popular (hubs)
- the clustering coefficient decreases as the vertix degree increases



Some Models of Network Generation

- Erdös-Rényi Random graph model:
  - a random graph is obtained by starting with a set of *N* vertices and adding edges between them at random. Each node pair is connected with probability of *p*
  - usually, N is large and p ~ 1/N
  - gives few components and small diameter
  - does give neither high clustering and nor heavy-tailed degree distributions
  - is the mathematically most well-studied and understood model
  - real-worlds networks are not randomly generated:
    - it significantly underestimates the number of high degree vertices
    - for instance, if Internet was random, we expect a portion of high degree vertices of 2.57, but it is 14.14



Some Models of Network Generation

- Watts-Strogatz small world graph model:
  - gives few components
  - does not give heavy-tailed degree distributions
  - extension of the random models which incorporates two properties:
    - in real networks the average distance between two nodes depends logarithmically on N (small diameter)
    - the average clustering coefficient of real networks is much higher

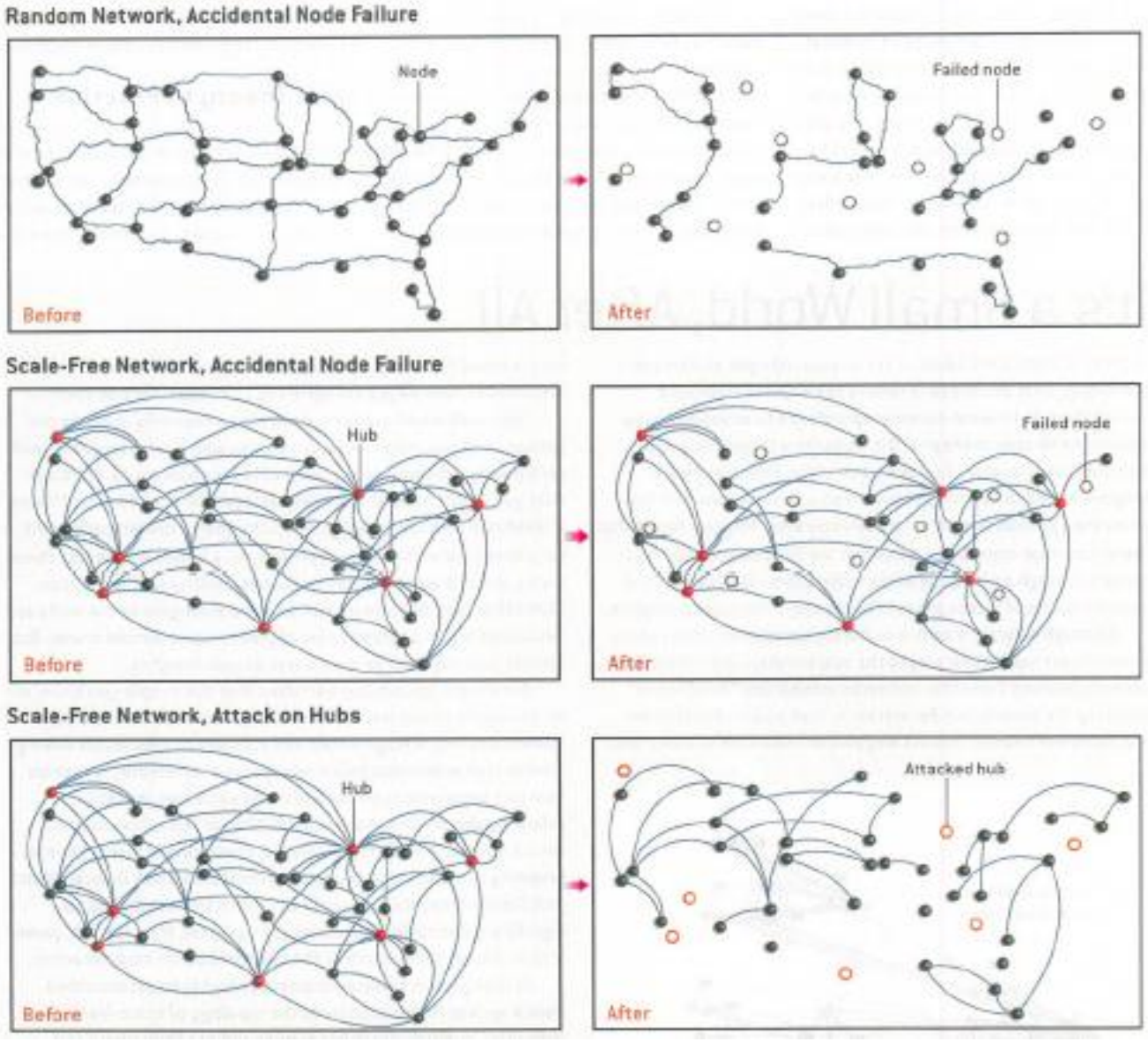
Some Models of Network Generation

- Barabási-Albert Scale-free model:
  - gives few components, small diameter and heavy-tailed distribution
  - does not give high clustering
  - the number of nodes (N) is not fixed
  - revises Watts model, in that the networs are not homogeneous in degree
  - the probability of connecting to a node is proportional to the current degree of that node: new edges are more likely to link to nodes with higher degrees (preferential-attachment)
  - scale-free network, a network whose degree distribution follows a power law

Some Models of Network Generation

- Barabási-Albert Scale-free model
  - networks continuously expand by additional new nodes
    - WWW: addition of new nodes
    - citation: publication of new papers
  - the preferential-attachment is not uniform
  - a node is linked with higher probability to a node that already has a large number of links
    - WWW: new documents link to well known sites (CNN, Yahoo, Google)
    - Citation: Well cited papers are more likely to be cited again

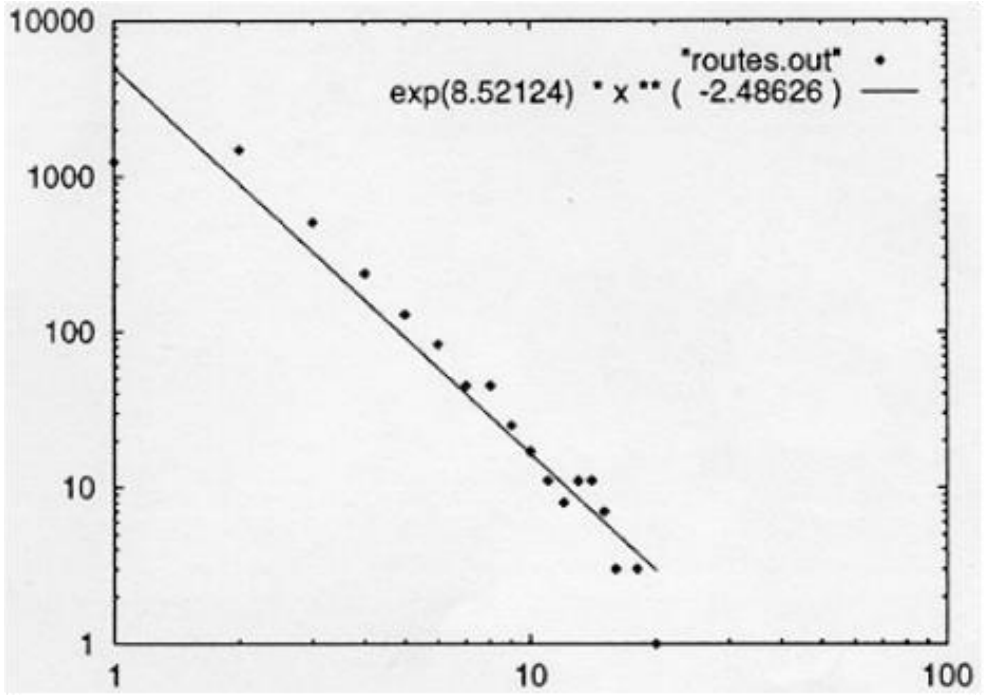
Random vs. Scale-Free Networks



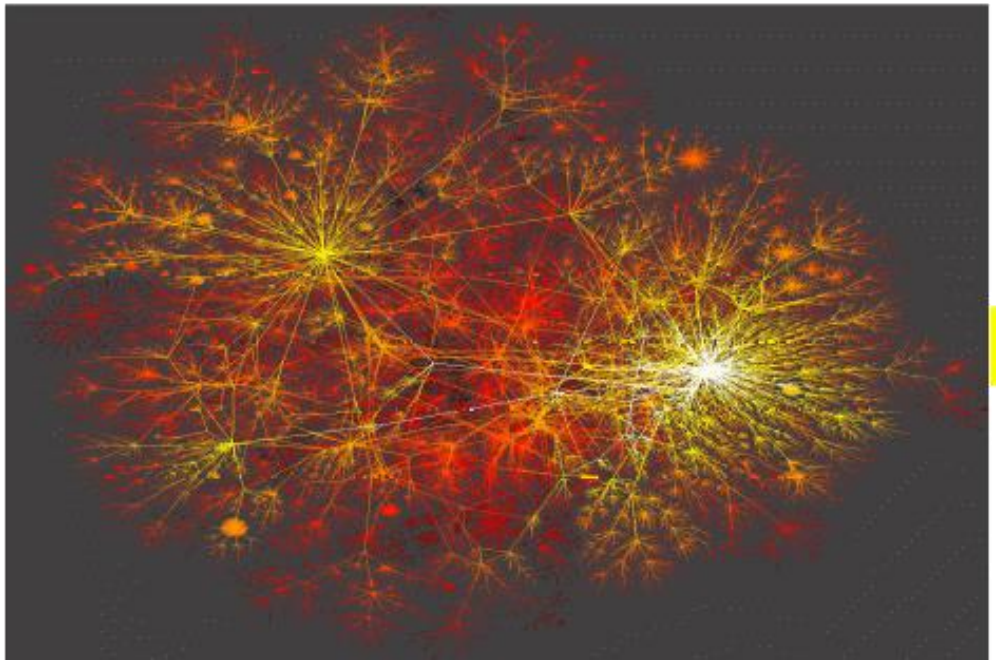
- The accidental failure of a number of nodes in a random network can fracture the system into non-communicating islands
- Scale-free networks are more robust in the face of such failures
- Scale-free networks are highly vulnerable to a coordinated attack against their hubs

Real World Case : Internet Backbone

Nodes: computers, routers
Links: physical lines



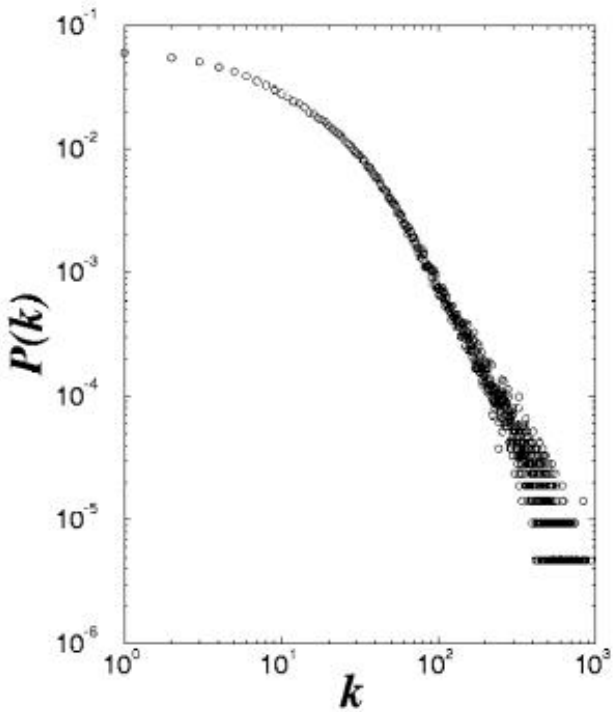
(Faloutsos, Faloutsos and Faloutsos, 1999)



Internet-Map

Real World Case : Actor Connectivity

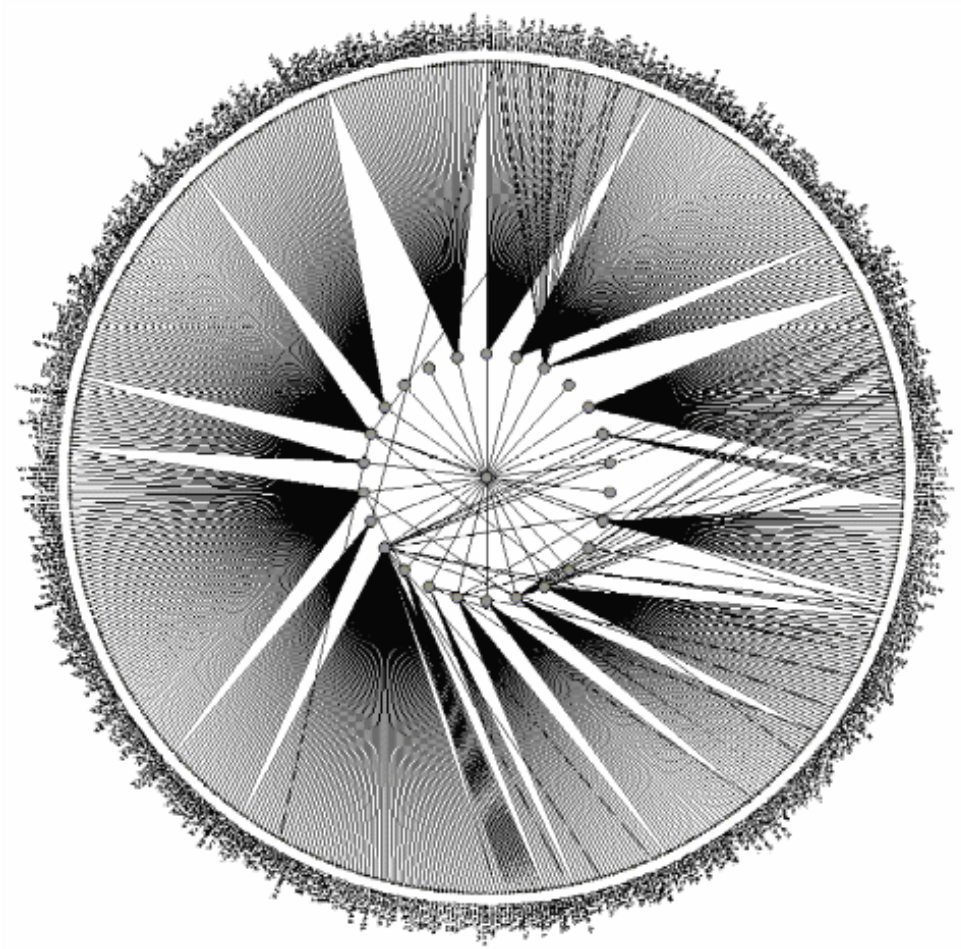
Nodes: actors
Links: cast jointly

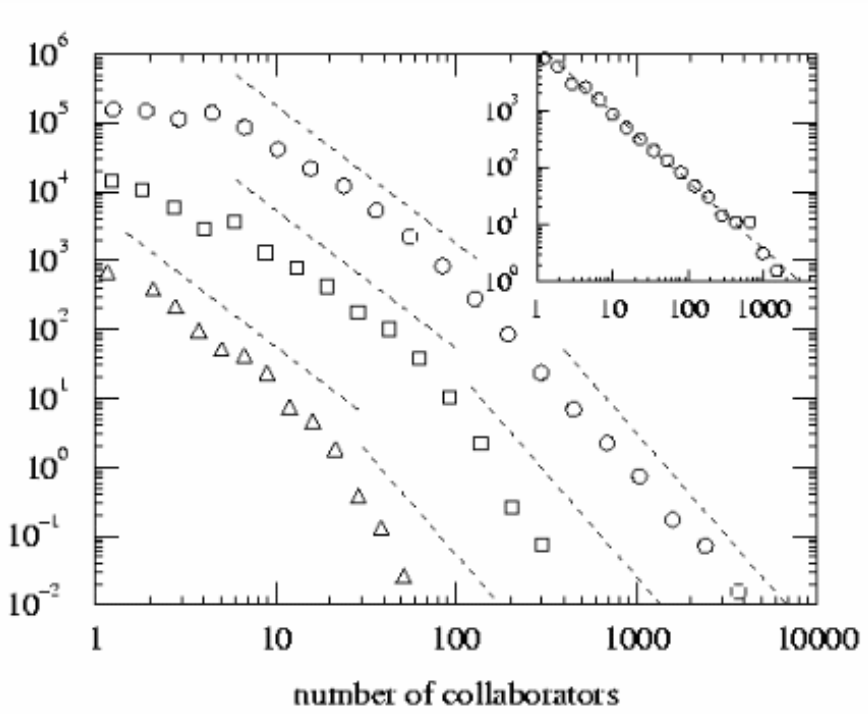


Real World Case : Co-authorship

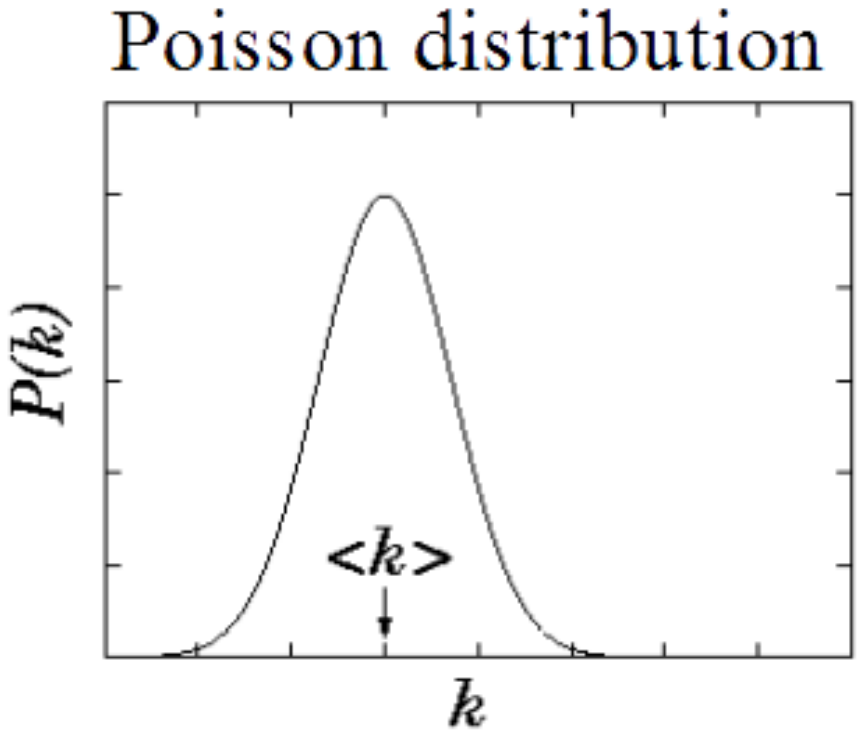
Nodes: scientist (authors)

Links: write paper together





Real World Case : Highway Network



Nodes: cities

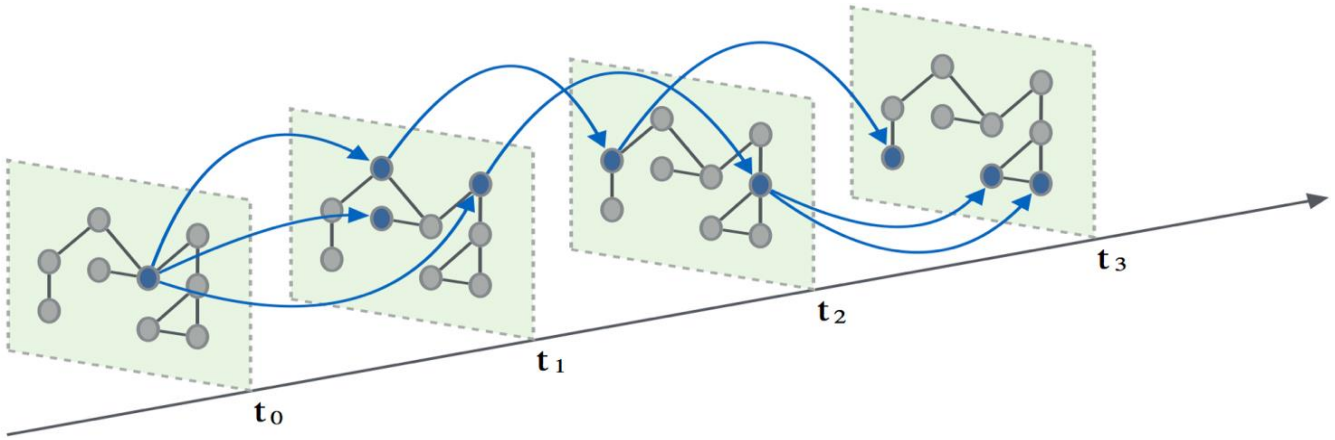
Links: highways, roads

Modeling Network Evolution

- Densification power law
  - The # of edges grows more than linearly to # of vertices, following a power law, with a positive *densification exponent*

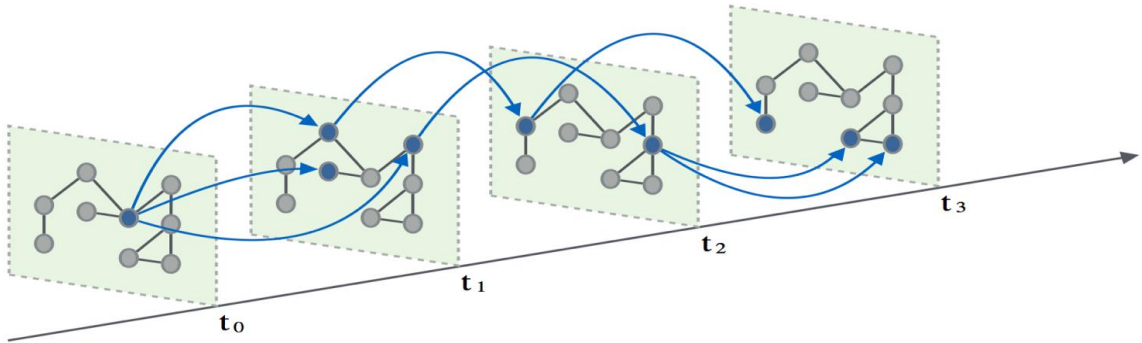
$$E(t) \propto N(t)^{\beta}$$ *where 2 > β > 1 in many real graphs*

- Shrinking diameter: The effective diameter of the graph *shrinks* as a graph grows over time



Modeling Network Evolution

- The Forest-Fire model: A preferential-attachment model that matches the densification power law and the shrinking diameter of graph evolution
  - The graph grows one node at a time. The new node *v* adds links to the existing node according to a "forest fire" process
    - Pick an ambassador node *w* uniformly at random and the links to *w*
    - Select some of ambassador's edges, and follow these edges and repeat
      - Similar to capture a "forest fire" at *w* and spread to other vertices
      - Example: a new computer science graduate student arrives at a university, meets some older CS students, who introduce him/her to their friends (CS or non-CS), and the introductions may continue recursively.



References

- S. Brin and L. Page, The anatomy of a large scale hypertextual Web search engine. WWW7.
- S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, Mining the link structure of the World Wide Web. IEEE Computer'99
- D. Cai, X. He, J. Wen, and W. Ma, Block-level Link Analysis. SIGIR'2004
- P. Domingos, Mining Social Networks for Viral Marketing. IEEE Intelligent Systems, 20(1), 80-82, 2005
- D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Cambridge Univ. Press, 2010
- L. Getoor: Lecture notes from Lise Getoor's website: www.cs.umd.edu/~getoor/
- D. Kempe, J. Kleinberg, and E. Tardos, Maximizing the Spread of Influence through a Social Network. KDD'03
- J. M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, J. ACM, 1999
- D. Liben-Nowell and J. Kleinberg. The Link Prediction Problem for Social Networks. CIKM'03
- M. Newman, Networks: An Introduction, Oxford Univ. Press, 2010
- D. Chakrabarti and C. Faloutsos, Graph Mining: Laws, Tools, and Case Studies, Morgan & Claypool, 2012