Data Mining: Principles and Algorithms

Introduction to Networks

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign

www.cs.uiuc.edu/~hanj

©2013 Jiawei Han. All rights reserved.
Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks
- Modeling of Network Formation
- Primitives of Social Networks
- Summary
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 \( G \)).
  - Multi-edge: if more than one edge between the same pair of vertices
  - Loop: if an edge connects vertex to itself (i.e., \((v_i, v_i)\))
- Simple network: if a network has neither self-edges nor multi-edges

Adjacency matrix:
- \( A_{ij} = 1 \) if there is an edge between vertices \( i \) and \( j \); 0 otherwise
- Directed graph (digraph): if each edge has a direction (tail \( \rightarrow \) head)
  - \( A_{ij} = 1 \) if there is an edge from \( j \) to \( i \); 0 otherwise
- Weighted graph: If a weight \( w_{ij} \) (usually a real number) is associated with each edge \( v_{ij} \)
Basic Network Structures and Properties

- **Subgraph**: A subset of the nodes and edges in a graph/network
- Given 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 $V'$
- **Clique** (complete graph): Every node is connected to every other
- **Singleton vs. dyad** (two nodes and their relationship) vs. **triad**:

```
  A -- B -- C
  ^   ^   ^
  |   |   |
  E   D   F
```

- **Ego-centric network**: A network pull out by selecting a node and all of its connections
  - The 1-degree egocentric network of A
  - The 1.5-degree egocentric network of A
  - The 2-degree egocentric network of A
Vertex Degree for Undirected & Directed Networks

Let a network $G = (V, E)$

Undirected Network

- Degree (or degree centrality) of a vertex: $d(v_i)$
  
  \[
  d(v_i) = |v_j| \quad s.t. \quad e_{ij} \in E \land e_{ij} = e_{ji}
  \]

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

Directed network

- In-degree of a vertex $d_{in}(v_i)$:
  
  \[
  d_{in}(v_i) = |v_j| \quad s.t. \quad e_{ij} \in E
  \]

  # of edges pointing to $v_i$

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

- Out-degree of a vertex $d_{out}(v_i)$:
  
  \[
  d_{out}(v_i) = |v_j| \quad s.t. \quad e_{ji} \in E
  \]

  # of edges from $v_i$

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

- **Degree sequence** of a graph: The list of degrees of the nodes 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 freq. distrib.: $(0, 1, 2, 1, 1)$

- **Degree distribution** of a graph:
  - Probability mass function $f$ for random variable $X$
  - $(f(0), f(1), ..., f(t), \text{ where } f(k) = P(X = k) = \frac{N_k}{n}$
  - E.g., in graph $G_1$, degree distrib.: $(0, 0.2, 0.4, 0.2, 0.2)$

- **Walk** in a graph $G$ between nodes $X$ and $Y$: ordered 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 vertices
  - **Distance**: the length of the shortest path
**Radius and Diameter of a Network**

- **Eccentricity**: The eccentricity of a node $v_i$ is the maximum distance from $v_i$ to any other nodes in the graph
  - $e(v_i) = \max_j \{d(v_i, v_j)\}$
  - E.g., $e(A) = 1$, $e(F) = e(B) = e(D) = e(H) = 2$

- **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., $d(G_1) = 2$

- Diameter is sensitive to outliers. Effective diameter: min # of hops for which a large fraction, typically 90%, of all connected pairs of nodes can reach each other.
Paths

- Path: A sequence of vertices such 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
- Total # of paths of length 2 from j to i, via any vertex, $N_{ij}^{(2)}$, is
  - $N_{ij}^{(2)} = \sum_{k=1}^{n} A_{ik} A_{kj} = [A^2]_{ij}$
  - where $[...]_{ij}$ denotes the $ij$th element of a matrix
- Generalizing to paths of arbitrary length $r$, we have
  - $N_{ij}^{(r)} = [A^r]_{ij}$
- When starting and ending at the same vertex $i$, we have
  - $L_r = \sum_{i=1}^{n} [A^r]_{ii} = Tr A^r$
- # of loops can be expressed in terms of the eigenvalues of the adjacency matrix
  - Matrix $A$ written in the form of $A = UKU^T$ where $U$ is the orthogonal matrix of eigenvalue and $K$ is the diagonal matrix of eigenvalues
  - $L_r = Tr (UKU^T)^r = Tr (UK^rU^T) = Tr (UU^TK^r) = Tr (K^r) = \sum_i k_i^r$
  - where $k_i$ is the $i$th eigenvalue of the adjacency matrix
Other Paths

- Geodesic path: shortest path
  - 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
  - Diameter of a graph: the length of the longest geodesic path between any pair of vertices in the network for which a path actually exists
- Eulerian path: a path that traverses each edge in a network exactly once

The Königsberg bridge problem

- Hamilton path: a path that visits each vertex in a network exactly once
Components in Directed & Undirected Network

Undirected network:
- Component: A subset of the vertices of a network such that there exists at least one path from each vertex to each other vertex.
- Adjacency matrix of a network with more than one component can be written in block diagonal form.
  \[
  A = \begin{pmatrix}
  [ & 0 & \cdots \\
  0 & [ & \cdots \\
  \vdots & \cdots & \ddots
  \end{pmatrix}
  \]

Directed network:
- Weakly vs. strongly connected component
  - Weakly connected if the vertices are connected by 1 or more paths where one can go either way along any edge.
- Out-component vs. in-component of a vertex (A) or a strongly connected component (A-C-F)
  - Out-component: Those reachable from vertex A.

The graph contains 2 weakly connected and 5 strongly connected components.
Independent Paths, Connectivity, and Cut Sets

- Two paths connecting a pair of vertices (A, B) are **edge-independent** if they share no edges.
- Two paths 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 disconnect a specified pair of vertices.

An **edge cut set** is a set of edges whose removal will disconnect a specified pair of vertices.

A **minimum cut set**: the smallest cut set that will disconnect a specified pair of vertices.

Menger’s theorem $\Rightarrow$ maxflow/min-cut theorem: For a pair of vertices, size of min-cut set $=$ vertex connectivity $=$ maximum flow.

This works also for weighted networks.

multiple size-2 edge/vertex cut set
Clustering Coefficient

- The clustering coefficient of a node $v_i$ is a measure of the density of edges in the neighborhood of $v_i$.
- 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 $v_i$).
- **Clustering coefficient of** $v_i$ **for undirected network** is
  \[
  C(v_i) = \frac{\text{# edges in } G_i}{\max \ # \text{ edges in } G_i} = \frac{m_i}{\binom{n_i}{2}} = \frac{2 \times m_i}{n_i(n_i - 1)}
  \]
  (corresp. to when $G_i$ is a complete graph)
- For directed network,
  \[
  C(v_i) = \frac{\text{# edges in } G_i}{\max \ # \text{ edges in } G_i} = \frac{m_i}{n_i(n_i - 1)}
  \]
- Clustering coefficient of a graph $G$: Averaging the local clustering coefficient of all the vertices (Watts & Strogatz):
  \[
  C(G) = \frac{1}{n} \sum_{i} C(v_i)
  \]
Co-citation and Bibliographic Coupling

- Co-citation of vertices i and j: # of vertices having outgoing edges pointing to both i and j
  \[ A_{ik} A_{jk} = 1 \text{ if } i \text{ and } j \text{ are both cited by } k \]
  - Co-citation of i and j:
    \[ C_{ij} = \sum_{k=1}^{n} A_{ik} A_{jk} = \sum_{k=1}^{n} A_{ik} A_{kj}^T \]
  - Co-citation matrix: It is a symmetric matrix
    \[ C = AA^T \]
  - Diagonal matrix (C_{ii}): total # papers citing i

- Bibliographic coupling of vertices i and j: # of other vertices to which both point
  \[ A_{ki} A_{kj} = 1 \text{ if } i \text{ and } j \text{ both cite } k \]
  - Bibliographic coupling of i and j:
    \[ B_{ij} = \sum_{k=1}^{n} A_{ki} A_{kj} = \sum_{k=1}^{n} A_{ik}^T A_{kj} \]
  - Co-citation matrix: \[ B = A^T A \]
  - Diagonal matrix (B_{ii}): total # papers cited by i
Cocitation & Bibliographic Coupling: Comparison

- Two measures are affected by the number of incoming and outgoing edges that vertices have.
- For strong co-citation: must have a lot of incoming edges
  - Must be well-cited (influential) papers, surveys, or books
  - Takes time to accumulate citations
- Strong bib-coupling if two papers have similar citations
  - A more uniform indicator of similarity between papers
  - Can be computed as soon as a paper is published
  - Not change over time
- Recent analysis algorithms
  - HITS explores both co-citation and bibliographic coupling
Bipartite Networks

- Bipartite Network: two kinds of vertices, and edges linking only vertices of unlike types
- Incidence matrix:
  - $B_{ij} = 1$ if vertex $j$ links to group $i$
  - 0 otherwise
- One can create a one-mode project from the two-mode partite form (but with info loss)
- The projection to one-mode can be written in terms of the incidence matrix $B$ as follows
  - $P_{ij} = \sum_{k=1}^{g} B_{ki} B_{kj} = \sum_{k=1}^{g} B_{ik}^T B_{kj}$
- The product of $B_{ki} B_{kj}$ will be 1 if $i$ and $j$ both belong to the sample group $k$ in the bi-partite network
The Small World Phenomenon & Erdös number

- Breadth-first search
- Erdös number: Distance from him/her to Erdös in the coauthor graph
  - Paul Erdös (a mathematician who published about 1500 papers)
  - Similarly, Kevin Bacon number (co-appearance in a movie)
- Small world phenomenon
  - Stanley Milgram’s experiments (1960s)
  - Microsoft Instant Messaging (IM) experiment (240 M active user accounts)
    - Jure Leskovec and Eric Horvitz (WWW 2008)
    - Est. avg. distance 6.6 & est. mean median 7
Network Data Sets

- Collaboration graphs
  - Co-authorships among authors
  - co-appearance in movies by actors/actresses
- Who-Talks-to-Whom graphs
  - Microsoft IM (Instant-Messaging)-graphs
- Information Linkage graphs
  - Web, citation graphs
- Technological graphs
  - Interconnections among computers
  - Physical, economic networks
- Networks in the Natural World
  - Food Web: who eats whom
  - Neural connections within an organism’s brain
  - Cells metabolism
Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks
- Modeling of Network Formation
- Primitives of Social Networks
- Summary
Centrality: Basic Measure in a Network

- **Centrality**: How “central” a node is in the network

- **Degree centrality**: degree of a node (the higher degree, more important the node)

- **Eccentricity centrality**: the less eccentric, the more central
  - \( c(v_i) = 1/e(v_i) \)
  - 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**: the average of the shortest path length from the node to every other node in the network, indicating how close a node is to all other nodes in the network
  - \( c(v_i) = 1/\sum_j d(v_i, v_j) \)
  - Median node \( v_m \) if \( v_m \) has the smallest total distance \( \sum_j d(v_m, v_j) \)
  - Facility location, e.g., shopping center, minimize total distance
Centrality Measures (II)

- **Betweenness centrality** for a node \( v \): # of shortest paths from all vertices to all others that pass through \( v \)
  - \( \eta_{jk} \): # 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 nodes

- **Eigenvector centrality**: Measure the influence of a node in a network, i.e., connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes
Centrality Measures on the Web (I): Eigenvector Centrality

- Web: a directed graph, PageRank and HITS are typical algorithms
- **Eigenvector centrality**, or **prestige**, importance, or rank of a node \( v \)
  - The more nodes point to \( v \), the higher \( v \)’s prestige
  - The more prestige of a node pointing to \( v \), the higher \( v \)’s prestige
- Let \( p(u) \) be prestige score for node \( u \). Then
  \[
  p(v) = \sum_u A(u, v) \cdot p(u) = \sum_u A^T(v, u) \cdot p(u)
  \]
- Written in vector form: \( p' = A^T p \)
- At \( k \)-th iteration, we have
  \[
  p_k = (A^T)^k p_0
  \]
  - Vector \( p_k \) converges to the dominant eigenvector of \( A^T \) with increasing \( k \)
Centrality Measures on the Web (II): PageRank

- **Random surfing assumption**: A web surfer randomly chooses one of the outgoing links from the current page or with some very small probability randomly jumps to any other page in the web graph.

- **Pagerank** of a page $v$: the probability of a random web surfer landing at $v$.

- **Normalized prestige**:
  - The prob. of visiting a page pointed by $v$ is $1/d_{\text{out}}(v)$, $d_{\text{out}}$ is outdegree of $v$.
  - Compute updated pagerank vector for $v$,
    \[
    p(v) = \sum_u \frac{A(u, v)}{d_{\text{out}}(u)} \cdot p(u) = \sum_u N(u, v) \cdot p(u) = \sum_u N^T(v, u) \cdot p(u), \text{ or } p = N^T \cdot p
    \]
    where $N(u, v)$ is the normalized adjacency matrix of the graph, and $N(u, v) = 1/d_{\text{out}}(u)$ if $(u, v)$ in $E$ or $0$ o.w.

- **Random Jumps**: a small prob. jumping to any other node (viewing web as a fully connected graph, i.e., adjacency matrix $A_r = 1_{n \times n}$).
  \[
  p(v) = \sum_u \frac{A_r(u, v)}{d_{\text{out}}(u)} \cdot p(u) = \sum_u N_r(u, v) \cdot p(u) = \sum_u (\frac{1}{n}) \cdot p(u), \text{ or } p = N_r^T \cdot p
  \]

- Pagerank score computation:
  \[
  p' = (1 - \alpha)N^T p + \alpha N_r^T p = ((1 - \alpha)N^T + \alpha N_r^T)p = M^T p
  \]
PageRank: Capturing Page Popularity (Brin & Page’98)

- Intuitions
  - Links are like citations in literature
  - A page that is cited often can be expected to be more useful in general
  - PageRank is essentially “citation counting”, but improves over simple counting
    - Consider “indirect citations” (being cited by a highly cited paper counts a lot…)
    - Smoothing of citations (every page is assumed to have a non-zero citation count)
  - PageRank can also be interpreted as a random surfing model (thus capturing popularity)
    - At any page,
      - With prob. \( \alpha \), randomly jumping to a page
      - With prob. \( (1 - \alpha) \), randomly picking a link to follow
Centrality Measures on the Web (III): HITS (Computing Hub & Authority Scores)

- For a specific query, a page of high Pagerank score may not be that relevant.
- HITS (Hyperlink Induced Topic Search) computes two values for a page:
  - Authority score: analogous to pagerank/prestige scores.
  - Hub score: based on how many “good” pages it points to.
- How is HITS query-based?
  - First uses standard search engines to retrieve the set of relevant pages.
  - Then expands the set to include any page that points to or is pointed to by some pages in the set.
  - Any pages originating from the same host are eliminated.
  - HITS is only applied on this expanded query-specific graph G.
- Computation:
  \[
  a(v) = \sum_u A^T(v, u) \cdot h(u) \quad h(v) = \sum_u A(v, u) \cdot a(u)
  \]
- In matrix computation (essentially two eigenvector computation):
  \[
  a_k = A^T h_{k-1} = A^T (A a_{k-2}) = (A^T A) a_{k-2}
  \]
  \[
  h_k = A a_{k-1} = A (A^T h_{k-2}) = (A A^T) h_{k-2}
  \]
HITS: Capturing Authorities & Hubs (Kleinberg’98)

- Intuitions of HITS (Hyperlink Induced Topic Search)
  - Pages that are widely cited are good authorities
  - Pages that cite many other pages are good hubs
- The key idea of HITS
  - Good authorities are cited by good hubs
  - Good hubs point to good authorities
  - Iterative reinforcement ...
- $AA^T$ is the co-citation matrix and $A^T A$ is the bibliographic coupling matrix. Authority centrality is eigenvector centrality for the co-citation network
Metrics (Measures) in Social Network Analysis (I)

- **Betweenness**: The extent to which a node lies between other nodes in the network. This measure takes into account the connectivity of the node's neighbors, giving a higher value for nodes which bridge clusters. The measure reflects the number of people who a person is connecting indirectly through their direct links.

- **Bridge**: An edge is a bridge if deleting it would cause its endpoints to lie in different components of a graph.

- **Centrality**: This measure gives a rough indication of the social power of a node based on how well they "connect" the network. "Betweenness", "Closeness", and "Degree" are all measures of centrality.

- **Centralization**: The difference between the number of links for each node divided by maximum possible sum of differences. A centralized network will have many of its links dispersed around one or a few nodes, while a decentralized network is one in which there is little variation between the number of links each node possesses.
**Metrics (Measures) in Social Network Analysis (II)**

- **Closeness**: The degree an individual is near all other individuals in a network (directly or indirectly). It reflects the ability to access information through the "grapevine" of network members. Thus, closeness is the inverse of the sum of the shortest distances between each individual and every other person in the network.

- **Clustering coefficient**: A measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'.

- **Cohesion**: The degree to which actors are connected directly to each other by cohesive bonds. Groups are identified as ‘cliques’ if every individual is directly tied to every other individual, ‘social circles’ if there is less stringency of direct contact, which is imprecise, or as structurally cohesive blocks if precision is wanted.

- **Degree (or geodesic distance)**: The count of the number of ties to other actors in the network.
(Individual-level) **Density**: The degree a respondent's ties know one another/proportion of ties among an individual's nominees. Network or global-level density is the proportion of ties in a network relative to the total number possible (sparse versus dense networks).

**Flow betweenness centrality**: The degree that a node contributes to sum of maximum flow between all pairs of nodes (not that node).

**Eigenvector centrality**: A measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to nodes having a high score contribute more to the score of the node in question.

**Local Bridge**: An edge is a local bridge if its endpoints share no common neighbors. Unlike a bridge, a local bridge is contained in a cycle.

**Path Length**: The distances between pairs of nodes in the network. Average path-length is the average of these distances between all pairs of nodes.
Metrics (Measures) in Social Network Analysis (IV)

- **Prestige**: In a directed graph, prestige is the term used to describe a node's centrality. "Degree Prestige", "Proximity Prestige", and "Status Prestige" are all measures of Prestige.

- **Radiality Degree**: an individual’s network reaches out into the network and provides novel information and influence.

- **Reach**: The degree any member of a network can reach other members of the network.

- **Structural cohesion**: The minimum number of members who, if removed from a group, would disconnect the group.

- **Structural equivalence**: Refers to the extent to which nodes have a common set of linkages to other nodes in the system. The nodes don’t need to have any ties to each other to be structurally equivalent.

- **Structural hole**: Static holes that can be strategically filled by connecting one or more links to link together other points. Linked to ideas of *social capital*: if you link to two people who are not linked you can control their communication.
Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks
- Modeling of Network Formation
- Primitives of Social Networks
- Summary
Why Network Modeling?

Many real-world networks exhibit certain common characteristics, even though they come from very different domains, e.g., communication, social, and biological networks.

A typical 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 (like 6)
  - growing only logarithmically with network size or even shrink?
  - typically exclude infinite distances

- **A high** degree of clustering:
  - 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
Common Properties of Real Networks

- Real network: **large, sparse** (# of edges \( |E| = O(n) \), \( n \): # of nodes)
- **Small-world property**: Avg. path length \( \mu_L \) scales logarithmically with \( n \) (# of nodes in the graph): \( \mu_L \propto \log n \)
  - Ultra-small-world property: \( \mu_L \ll \log n \)
- **Scale-free property** (**power law distribution**): most nodes have very small degree, but a few hub nodes have high degrees
  - The probability that a node has degree \( k \): \( f(k) \propto k^{-\gamma} \)
  - log-log plot shows a straight line:
    \[
    \log f(k) = \log(\alpha k^{-\gamma}) = -\gamma \log k + \log \alpha
    \]
- **Clustering effect**: Two nodes are more likely to be connected if they share a common neighbor
  - Clustering effect: a high clustering coefficient for graph \( G \)
  - \( C(k) \): avg clustering coefficient for nodes with degree \( k \)
  - Power law relationship between \( C(k) \) and \( k \): \( C(k) \propto k^{-\gamma} \)
All of the network generation models we will study are *probabilistic* or *statistical* in nature.

They can generate networks of any size.

They often have various *parameters* that can be set:
- size of network generated
- average degree of a vertex
- fraction of long-distance connections

The models generate a *distribution* over networks.

Statements are always *statistical* in nature:
- *with high probability*, diameter is small
- *on average*, degree distribution has heavy tail

Thus, we’re going to need some basic statistics and probability theory.
A random variable $X$ is simply a variable that probabilistically assumes values in some set

- set of possible values sometimes called the sample space $S$ of $X$
- sample space may be small and simple or large and complex
  - $S = \{\text{Heads, Tails}\}$, $X$ is outcome of a coin flip
  - $S = \{0,1,\ldots,\text{U.S. population size}\}$, $X$ is number voting democratic
  - $S = \text{all networks of size N}$, $X$ is generated by preferential attachment

Behavior of $X$ determined by its distribution (or density)

- for each value $x$ in $S$, specify $p(X = x)$
- these probabilities sum to exactly 1 (mutually exclusive outcomes)
- complex sample spaces (such as large networks):
  - distribution often defined implicitly by simpler components
  - might specify the probability that each edge appears independently
  - this induces a probability distribution over networks
  - may be difficult to compute induced distribution
Independence, Expectation & Variance

- **Independence:**
  - Let X and Y be random variables.
  - Unconditional independence: for any x & y, p(X = x, Y = y) = p(X=x) × p(Y=y)
  - Intuition: value of X does not influence value of Y, vice-versa
  - Conditional independence: p(X, Y | Z) = p(X|Z) p(Y|Z)

- **Expected (mean) value** of X: µ
  - Only makes sense for *numeric* random variables
  - “Average” value of X according to its distribution
  - Formally, E[X] = ∑_{x ∈ X} x p(x), i.e., sum over all x in X
  - True only for *independent* random variables: E[XY] = E[X]E[Y]

- **Variance** of X:
  - Var[X] = E[(X − µ)^2]; often denoted by σ^2
  - *Standard deviation* σ = sqrt(Var[X])

- **Union bound**:
  - For any X, Y, p(X=x, Y=y) ≤ p(X=x) + p(Y=y)
Convergence to Expectations

- Let $X_1, X_2, \ldots, X_n$ be:
  - *independent* random variables
  - with the *same* distribution $p(X=x)$
  - expectation $\mu = E[X]$ and variance $\sigma^2$
  - independent and identically distributed (i.i.d.)
  - essentially $n$ repeated “trials” of the same experiment
  - natural to examine random variable $Z = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - example: number of heads in a sequence of coin flips
  - example: degree of a vertex in the random graph model
  - $E[Z] = E[X]$; what can we say about the *distribution* of $Z$?

- *Central Limit Theorem*:
  - as $n$ becomes large, $Z$ becomes *normally distributed*
    - with expectation $\mu$ and variance $\sigma^2/n$
The normal or Gaussian density applies to continuous, real-valued random variables characterized by mean $\mu$ and standard deviation $\sigma$.

Density at $x$ is defined as:
\[
\frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\]

A special case is $\mu = 0, \sigma = 1$: $\alpha \exp(-x^2/\beta)$ for some constants $\alpha, \beta > 0$.

Peaks at $x = \mu$, then dies off exponentially rapidly.

The classic “bell-shaped curve”, e.g., exam scores, human body temperature.

Remarks:
- Can control mean and standard deviation independently.
- Can make as “broad” as we like, but always have finite variance.
The Binomial Distribution

- Coin with $p(\text{heads}) = p$, flip $n$ times, probability of getting exactly $k$ heads:
  - choose $(n, k) = p^k(1-p)^{n-k}$
- For large $n$ and $p$ fixed:
  - approximated well by a normal distribution with
    $\mu = np$, $\sigma = \sqrt{np(1-p)}$
  - $\sigma/\mu \to 0$ as $n$ grows
  - leads to strong, large deviation bounds

www.professionalgambler.com/binomial.html
The Poisson Distribution

- Like binomial, applies to variables taken on integer values $> 0$
- Often used to model *counts* of events
  - number of phone calls placed in a given time period
  - number of times a neuron fires in a given time period
- Single free parameter $\lambda$, probability of exactly $x$ events:
  - $\exp(-\lambda) \frac{\lambda^x}{x!}$
  - mean and variance are both $\lambda$
- Binomial distribution with $n$ large, $p = \lambda/n$ ($\lambda$ fixed)
  - converges to Poisson with mean $\lambda$
Power Law (or Pareto) Distributions

- Pareto distribution (heavy-tailed, or power law):
  - Pareto\( (x | k,m) = k m^k x^{-(k+1)} 1(x \geq m) \)
  - \( x \) must be greater than some constant \( m \) but not too much greater
  - Distributions: mode = \( m \)
    - mean = \( km/(k-1) \) if \( k > 1 \)
    - Variance = \( m^2k/((k-1)^2(k-2)) \) if \( k > 2 \)

- For variables assuming integer values > 0, probability of value \( x \sim 1/x^\alpha \)
  - This is why it is called power law distribution, also referred to as scale-free
  - If we plot the distribution on a log-log scale, it forms a straight line
  - Typically \( 0 < \alpha < 2 \); smaller \( \alpha \) gives heavier tail

- Why long tails or heavy tails? For binomial, normal, and Poisson distributions, the tail probabilities approach 0 exponentially fast

- What kind of phenomena does this distribution model?
  - Word frequency vs their rank (the, of, ...); wealth distribution; ...
Distinguishing Distributions in Data

- All these distributions are *idealized models*
- In practice, we do not see distributions, but *data*
- Typical procedure to distinguish between Poisson, power law, ...
  - might restrict our attention to a *range* of values of interest
  - accumulate *counts* of observed data into equal-sized bins
  - look at counts on a *log-log plot*
- Power law:
  - \( \log(P(X = x)) = \log(1/x^\alpha) = -\alpha \log(x) \)
  - linear, slope \( -\alpha \)
- Normal:
  - \( \log(P(X = x)) = \log(\alpha \exp(-x^2/b)) = \log(\alpha) - x^2/b \)
  - non-linear, concave near mean
- Poisson:
  - \( \log(P(X = x)) = \log(\exp(-l) \ l^x/x!) \)
  - also non-linear
Zipf’s Law

- Pareto distribution vs. Zipf’s Law
  - Pareto distributions are continuous probability distributions
  - Zipf's law: a discrete counterpart of the Pareto distribution
- Zipf's law:
  - Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table
  - Thus the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc.
- General theme:
  - *rank* events by their *frequency of occurrence*
  - resulting distribution often is a power law!
- Other examples:
  - North America city sizes
  - personal income
  - file sizes
  - genus sizes (number of species)
Zipf Distribution

The same data plotted on linear and logarithmic scales. Both plots show a Zipf distribution with 300 data points.

Linear scales on both axes  Logarithmic scales on both axes
Some Models of Network Generation

- **Erdös-Rényi Random graph model:**
  - Gives few components and small diameter
  - does not give high clustering and heavy-tailed degree distributions
  - is the mathematically most well-studied and understood model

- **Watts-Strogatz small world graph model:**
  - gives few components, small diameter and high clustering
  - does not give heavy-tailed degree distributions

- **Barabási-Albert Scale-free model:**
  - gives few components, small diameter and heavy-tailed distribution
  - does not give high clustering

- **Hierarchical network:**
  - few components, small diameter, high clustering, heavy-tailed

- **Affiliation network:**
  - models group-actor formation
Erdös-Rényi (ER) Random Graph Model

- A random graph is obtained by starting with a set of $N$ vertices and adding edges between them at random

- Different *random graph models* produce different *probability distributions* on graphs

- Most commonly studied is the *Erdős–Rényi model*, denoted $G(N, p)$, in which *every possible edge occurs independently with probability $p$*

- Random graphs were first defined by Paul Erdős and Alfréd Rényi in their 1959 paper "On Random Graphs"

- The usual *regime of interest* is when $p \sim 1/N$, $N$ is large
  - e.g., $p = 1/2N$, $p = 1/N$, $p = 2/N$, $p = 10/N$, $p = \log(N)/N$, etc.
  - in expectation, each vertex will have a “small” number of neighbors
  - will then examine what happens when $N \to \infty$
  - can thus study properties of *large networks* with *bounded degree*
  - Sharply concentrated; *not* heavy-tailed
Erdös-Rényi Model (1959)

Connect with probability $p$

$p = 1/6$

$N = 10$

$\langle k \rangle \sim 1.5$

Poisson distribution

- Democratic
- Random

Pál Erdös (1913-1996)
The Watts and Strogatz Model

- Proposed by Duncan J. Watts and Steven Strogatz in their joint 1998 Nature paper
- A random graph generation model that produces graphs with **small-world properties**, including **short average path lengths** and **high clustering**
- The model also became known as the (Watts) **beta** model after Watts used $\beta$ to formulate it in his popular science book *Six Degrees*
- The **ER** graphs fail to explain two important properties observed in real-world networks:
  - By assuming a constant and independent probability of two nodes being connected, they do not account for local clustering, i.e., having a low *clustering coefficient*
  - Do not account for the formation of hubs. Formally, the degree distribution of ER graphs converges to a Poisson distribution, rather than a *power law* observed in most real-world, *scale-free networks*
The Watts-Strogatz Model: Characteristics

- $C(p)$ : clustering coeff.
- $L(p)$ : average path length

(Watts and Strogatz, Nature 393, 440 (1998))
Small Worlds and Occam’s Razor

- For small $\alpha$, should generate large clustering coefficients
  - we “programmed” the model to do so
  - Watts claims that proving precise statements is hard...
- But we do not want a new model for every little property
  - Erdos-Renyi $\rightarrow$ small diameter
  - $\alpha$-model $\rightarrow$ high clustering coefficient
- In the interests of Occam’s Razor, we would like to find
  - a single, simple model of network generation...
  - ... that simultaneously captures many properties
- Watt’s small world: small diameter and high clustering
Watts examines three real networks as case studies:
- the Kevin Bacon graph
- the Western states power grid
- the C. elegans nervous system

For each of these networks, he:
- computes its size, diameter, and clustering coefficient
- compares diameter and clustering to best Erdos-Renyi approx.
- shows that the best $\alpha$-model approximation is better
- important to be “fair” to each model by finding best fit

Overall,
- if we care only about diameter and clustering, $\alpha$ is better than $p$
Case 1: Kevin Bacon Graph

- Vertices: actors and actresses
- Edge between u and v if they appeared in a film together

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Average distance</th>
<th># of movies</th>
<th># of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rod Steiger</td>
<td>2.537527</td>
<td>112</td>
<td>2562</td>
</tr>
<tr>
<td>2</td>
<td>Donald Pleasence</td>
<td>2.542376</td>
<td>180</td>
<td>2874</td>
</tr>
<tr>
<td>3</td>
<td>Martin Sheen</td>
<td>2.551210</td>
<td>136</td>
<td>3501</td>
</tr>
<tr>
<td>4</td>
<td>Christopher Lee</td>
<td>2.552497</td>
<td>201</td>
<td>2993</td>
</tr>
<tr>
<td>5</td>
<td>Robert Mitchum</td>
<td>2.557181</td>
<td>136</td>
<td>2905</td>
</tr>
<tr>
<td>6</td>
<td>Charlton Heston</td>
<td>2.566284</td>
<td>104</td>
<td>2552</td>
</tr>
<tr>
<td>7</td>
<td>Eddie Albert</td>
<td>2.567036</td>
<td>112</td>
<td>3333</td>
</tr>
<tr>
<td>8</td>
<td>Robert Vaughn</td>
<td>2.570193</td>
<td>126</td>
<td>2761</td>
</tr>
<tr>
<td>9</td>
<td>Donald Sutherland</td>
<td>2.577880</td>
<td>107</td>
<td>2865</td>
</tr>
<tr>
<td>10</td>
<td>John Gielgud</td>
<td>2.578980</td>
<td>122</td>
<td>2942</td>
</tr>
<tr>
<td>11</td>
<td>Anthony Quinn</td>
<td>2.579750</td>
<td>146</td>
<td>2978</td>
</tr>
<tr>
<td>12</td>
<td>James Earl Jones</td>
<td>2.584440</td>
<td>112</td>
<td>3787</td>
</tr>
</tbody>
</table>

Kevin Bacon

No. of movies : 46
No. of actors : 1811
Average separation: 2.79

Is Kevin Bacon the most connected actor?

NO!

876 Kevin Bacon 2.786981 46 1811
#1 Rod Steiger

Donald Pleasence

#3 Martin Sheen

#2

#876 Kevin Bacon

Bacon-map
Case 2: New York State Power Grid

- Vertices: generators and substations
- Edges: high-voltage power transmission lines and transformers
- Line thickness and color indicate the voltage level
  - Red 765 kV, 500 kV; brown 345 kV; green 230 kV; grey 138 kV
Case 3: C. Elegans Nervous System

- Vertices: neurons in the C. elegans worm
- Edges: axons/synapses between neurons
Two More Examples

- M. Newman on **scientific collaboration networks**
  - coauthorship networks in several distinct communities
  - differences in degrees (papers per author)
  - empirical verification of
    - giant components
    - small diameter (mean distance)
    - high clustering coefficient

- Alberich et al. on the **Marvel Universe**
  - purely fictional social network
  - two characters linked if they appeared together in an issue
  - “empirical” verification of
    - heavy-tailed distribution of degrees (issues and characters)
    - giant component
    - rather **small** clustering coefficient
Barabási–Albert Scale-Free Model

- Major limitation of the Watts-Strogatz model
  - It produces graphs that are homogeneous in degree
  - Real networks are often inhomogeneous in degree, having hubs and a scale-free degree distribution (*scale-free networks*)
- Scale-free networks are better described by the *preferential attachment* family of models, e.g., the Barabási–Albert (BA) model
  - Edges from the new vertex are more likely to link to nodes with higher degrees
  - The *rich-get-richer* approach
- This leads to the proposal of a new model: *scale-free network*, a network whose degree distribution follows a *power law*, at least asymptotically
World Wide Web: A Scale-free Network

**Nodes**: WWW documents

**Links**: URL links

800 million documents
(S. Lawrence, 1999)

**ROBOT**: collects all URL’s found in a document and follows them recursively

**Expected Result**

- $\langle k \rangle \sim 6$
- $P(k=500) \sim 10^{-99}$
- $N_{WWW} \sim 10^9$
- $\Rightarrow N(k=500) \sim 10^{-90}$

**Real Result**

- $P_{\text{out}}(k) \sim k^{-\gamma_{\text{out}}}$
- $P_{\text{in}}(k) \sim k^{-\gamma_{\text{in}}}$

- $\gamma_{\text{out}} = 2.45$
- $\gamma_{\text{in}} = 2.1$

- $P(k=500) \sim 10^{-6}$
- $N_{WWW} \sim 10^9$
- $\Rightarrow N(k=500) \sim 10^3$

Length of Paths and Number of Nodes

Finite size scaling: create a network with $N$ nodes with $P_{\text{in}}(k)$ and $P_{\text{out}}(k)$

$$< l > = 0.35 + 2.06 \log(N)$$

$1_{15} = 2 \ [1 \rightarrow 2 \rightarrow 5]$

$1_{17} = 4 \ [1 \rightarrow 3 \rightarrow 4 \rightarrow 6 \rightarrow 7]$

$\ldots \ < l > = ??$

19 degrees of separation

R. Albert et al, Nature (99)

based on 800 million webpages

[S. Lawrence et al Nature (99)]

A. Broder et al WWW9 (00)

nd.edu

IBM

$\langle l \rangle$ vs. $N$
What Does that Mean?

Poisson distribution

\[ P(k) \]

\[ \langle k \rangle \]

\[ k \]

Power-law distribution

\[ P(k) \]

\[ 1 \]

\[ 0.1 \]

\[ 0.01 \]

\[ 0.001 \]

\[ 0.0001 \]

Exponential Network

Scale-free Network
**Scale-Free Networks: Major Ideas**

- The number of nodes \( (N) \) is not fixed
  - Networks continuously expand by additional new nodes
    - WWW: addition of new nodes
    - Citation: publication of new papers
- The 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
Generation of Scale-Free Network

- Start with (say) two vertices connected by an edge
- For $i = 3$ to $N$:
  - for each $1 \leq j < i$, $d(j) =$ degree of vertex $j$ so far
  - let $Z = \sum d(j)$ (sum of all degrees so far)
  - add new vertex $i$ with $k$ edges back to $\{1, \ldots, i-1\}$:
    - $i$ is connected back to $j$ with probability $d(j)/Z$
- Vertices $j$ with high degree are likely to get more links! — "Rich get richer"
- Natural model for many processes:
  - hyperlinks on the web
  - new business and social contacts
  - transportation networks
- Generates a power law distribution of degrees
  - exponent depends on value of $k$
- Preferential attachment explains
  - heavy-tailed degree distributions
  - small diameter ($\sim \log(N)$, via “hubs”)
- Will not generate high clustering coefficient
  - no bias towards local connectivity, but towards hubs
Robustness of 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.
Case 1: Internet Backbone

**Nodes**: computers, routers

**Links**: physical lines

(Faloutsos, Faloutsos and Faloutsos, 1999)
Case 2: Actor Connectivity

Nodes: actors

Links: cast jointly

N = 212,250 actors
⟨k⟩ = 28.78

P(k) \sim k^{-\gamma}

\gamma = 2.3

Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)
Case 3: Science Citation Index

Nodes: papers
Links: citations

1736 PRL papers (1988)

Witten-Sander
PRL 1981

$P(k) \sim k^{-\gamma}$
$(\gamma = 3)$
(S. Redner, 1998)

* citation total may be skewed because of multiple authors with the same name
Case 4: Science Co-authorship

**Nodes**: scientist (authors)

**Links**: write paper together

(Newman, 2000, H. Jeong et al 2001)
Case 5: Food Web

Nodes: trophic species
Links: trophic interactions

Bio-Map

GENOME
protein-gene interactions

PROTEOME
protein-protein interactions

METABOLISM
Bio-chemical reactions
Biochemical Pathways
Prot Interaction Map: Yeast Protein Network

**Nodes**: proteins

**Links**: physical interactions (binding)

**Finding Proteins That Interact**

One technique, called the yeast two-hybrid system, relies on bringing into close proximity two halves (a and b) of a protein that activates a gene that causes a yeast cell to turn blue. It is used to determine which of a pool of unknown "prey" proteins binds to a known "bait" protein.

1. Insert DNA encoding a known "bait" protein linked to DNA for half (a) of the activator protein.
2. Insert DNA for the other half (b) of the activator protein linked to DNA encoding random "prey" proteins.
3. Look for color change, which indicates "prey" protein binding to "bait".

Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks
- Modeling of Network Formation
- Primitives of Social Networks
- Summary
Social Networks

- Social network: A social structure made of nodes (individuals or organizations) that are related to each other by various interdependencies like friendship, kinship, like, ...

- Graphical representation
  - Nodes = members
  - Edges = relationships

- Examples of typical social networks on the Web
  - Social bookmarking (Del.icio.us)
  - Friendship networks (Facebook, Myspace, LinkedIn)
  - Blogosphere
  - Media Sharing (Flickr, Youtube)
  - Folksonomies
Web 2.0 Examples

- Blogs
  - Blogspot
  - Wordpress
- Wikis
  - Wikipedia
  - Wikiversity
- Social Networking Sites
  - Facebook
  - Myspace
  - Orkut
- Digital media sharing websites
  - Youtube
  - Flickr
- Social Tagging
  - Del.icio.us
- Others
  - Twitter
  - Yelp
### Friendship Networks vs. Blogosphere

<table>
<thead>
<tr>
<th>Friendship Networks</th>
<th>Blogosphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit Links/Edges</td>
<td>Implicit Links/Edges</td>
</tr>
<tr>
<td>Undirected Graph</td>
<td>Directed Graph</td>
</tr>
<tr>
<td>Network Centrality Measures</td>
<td>Blog Statistics</td>
</tr>
<tr>
<td>Quantifying Spread of Influence</td>
<td>Quantifying Influential Members</td>
</tr>
<tr>
<td>Nodes are members/actors</td>
<td>Nodes can be bloggers/blogs or blog sites</td>
</tr>
<tr>
<td>Strictly defined graph structure</td>
<td>Loosely defined graph structure</td>
</tr>
<tr>
<td>“Being in touch” or “Making Friends”</td>
<td>Sharing ideas and opinions</td>
</tr>
<tr>
<td>Person-to-person</td>
<td>Person-to-group</td>
</tr>
<tr>
<td>Friendship Oriented</td>
<td>Community Oriented</td>
</tr>
<tr>
<td>Member’s Reputation/Trust based on network connections</td>
<td>Member’s Reputation/Trust based on the response to</td>
</tr>
<tr>
<td>and/or location in the network</td>
<td>other member’s knowledge solicitations</td>
</tr>
</tbody>
</table>

Adapted from H. Liu & N. Agarwal, KDD’08 tutorial
Nodes: individuals

Links: social relationship (family/work/friendship/etc.)

S. Milgram (1967)  
John Guare

Six Degrees of Separation

Social networks: Many individuals with diverse social interactions between them
Communication Networks

The Earth is developing an electronic nervous system, a network with diverse **nodes** and **links** are:

- computers
- routers
- satellites
- phone lines
- TV cables
- EM waves

Communication networks: Many non-identical components with diverse connections between them
Humans have only about three times as many genes as the fly, so human complexity seems unlikely to come from a sheer quantity of genes. Rather, some scientists suggest, each human has a network with different parts like genes, proteins and groups.
Natural Networks and Universality

- Consider many kinds of networks:
  - social, technological, business, economic, content, ...
- These networks tend to share certain informal properties:
  - large scale; continual growth
  - distributed, organic growth: vertices “decide” who to link to
  - interaction restricted to links
  - mixture of local and long-distance connections
  - abstract notions of distance: geographical, content, social, ...
- Social network theory and link analysis
  - Do natural networks share more quantitative universals?
  - What would these “universals” be?
  - How can we make them precise and measure them?
  - How can we explain their universality?
Introduction to Networks

- Basic Measures of Networks
- Centrality Analysis in Networks
- Modeling of Network Formation
- Primitives of Social Networks
- Summary
Summary

- Primitives for networks
- Measure and metrics of networks
  - Degree, eigenvalue, Katz, PageRank, HITS
- Models of network formation
  - Erdös-Rényi, Watts and Strogatz, scale-free
- Social networks
Ref: Introduction to Networks

- D. Cai, X. He, J. Wen, and W. Ma, Block-level Link Analysis. SIGIR'2004.
- 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
Unused class slides

The following slides were covered in a previous course but not in 2014
**Eigen Vector Centrality**

- Not all neighbors are equal: A vertex’s importance is increased by having connections to other vertices that are themselves important.

- Eigen vector centrality gives each vertex a score proportional to the sum of the scores of its neighbors. (Note: $A_{ij}$ is an element of the adjacency matrix $A$)

$$x'_i = \sum_j A_{ij} x_j \quad \text{or in its matrix form } x' = Ax$$

- After $t$ steps, we have (where $k_i$: the eigenvalues of $A$, and $k_1$: the largest one)

$$x(t) = A^t X(0) = A^t \sum_i c_i v_i = \sum_i c_i k_i^t v_i = k_1^t \sum_i c_i \left[ \frac{k_i}{k_1} \right]^t v_i$$

- Since $\frac{k_i}{k_1} < 1$ for all $i \neq 1$, when $t \to \infty$, we get $x(t) \to c_1 k_1^t v_1$

- That is, the limiting vector of centrality is simply proportional to the leading eigenvector of the adjacency matrix. Thus we have,

$$Ax = k_1 x \quad \text{or } x_i = k_1^{-1} A_{ij} x_j$$

- Difficulty for directed networks: Only vertices that are in a strongly connected component of two or more vertices, or the out-component of such a component, can have non-zero eigenvector centrality.
Katz Centrality

To overcome the difficulty of eigenvector centrality, simply give each vertex a small amount of centrality “for free”:

\[ x_i = \alpha \sum_j A_{ij} x_j + \beta \]

where \( \alpha \) and \( \beta \) are positive constants, or written in matrix form:

\[ \mathbf{x} = \alpha \mathbf{Ax} + \beta \mathbf{1} \]

where \( \mathbf{1} \) is the vector \((1, 1, \ldots)\)

For convenience, set \( \beta = 1 \), we have

\[ \mathbf{x} = (\mathbf{I} - \alpha \mathbf{A})^{-1} \mathbf{1} \]

This centrality measure was proposed by Katz in 1953, called Katz centrality.

What \( \alpha \) value one should take? Most researchers have employed values close to the maximum of \( \frac{1}{k_i} \), *which* places the max weight on the eigen vector term and the smallest amount on the constant term.

Efficient computation often use the 1st equation instead of inverting matrix.

An extension to Katz centrality: \( x_i = \alpha \sum_j A_{ij} x_j + \beta_i \)

where \( \beta_i \) is some intrinsic non-network contribution, e.g., age/income.
PageRank

- The centrality gained by receiving an edge from a prestige vertex (e.g., Yahoo!) should be diluted for being shared with many others
  \[ x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta \]
  where centrality derived by its network neighbors should be divide by their out-degrees \( k_j^{out} \). Note if its out-degree is 0, we can set \( k_j^{out} = 1 \)

- In matrix form, it has
  \[ \mathbf{x} = \alpha \mathbf{AD}^{-1}\mathbf{x} + \beta \mathbf{1} \]
  where \( \mathbf{1} \) is vector \((1, 1, 1, \ldots)\), and \( \mathbf{D} \): diagonal matrix with \( D_{ii} = \max(k_j^{out}, 1) \)

- Re-arranging the equation, we have
  \[ \mathbf{x} = (\mathbf{I} - \alpha \mathbf{AD}^{-1})^{-1} \mathbf{1} = \mathbf{D} (\mathbf{D} - \alpha \mathbf{A})^{-1} \mathbf{1} \]
  This centrality measure is known as PageRank

- Google search engine uses \( \alpha = 0.85 \) (likely a guess based on experiments)

- Considering different constants for different vertices (e.g., textual relevance to a search query)
  \[ x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta_i \]
  or in matrix form:
  \[ \mathbf{x} = \mathbf{D} (\mathbf{D} - \alpha \mathbf{A})^{-1} \mathbf{\beta} \]
## Relationship Among Four Centrality Measures

<table>
<thead>
<tr>
<th></th>
<th>with constant term</th>
<th>without constant term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divided by out-degree</td>
<td>$x = D (D - \alpha A)^{-1} 1$</td>
<td>$x = AD^{-1}x$ degree centrality</td>
</tr>
<tr>
<td>PageRank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No such division</td>
<td>$x = (I - \alpha A)^{-1} 1$</td>
<td>$x = k_1^{-1}Ax$ eigenvector centrality</td>
</tr>
<tr>
<td>Katz Centrality</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### A comparison of the four centrality measures
- Note that the diagonal matrix $D$, which normally has elements $D_{ii} = k_i$, must be defined slightly differently for PageRank as $D_{ii} = \max(k_{i, out}, 1)$
- Each of the measures can be applied to directed networks as well as undirected ones, although degree centrality takes more complicated values in the directed case and is not widely used
- Most commonly used measures: PageRank and eigenvector centrality
Hubs and Authorities

- HITS (hyperlink-induced topic search) algorithm by Kleinberg (1999)
- There exist two types of centrality for directed networks:
  - Authorities: nodes that contain useful info on a topic of interest
  - Hubs: nodes that tell us where the best authorities are to be found
- Ex: (1) review articles and authoritative research articles; (2) web pages
- HITS gives each vertex $i$ an authority centrality $x_i$ and a hub centrality $y_i$
  - A vertex has high authority centrality if it is pointed to by many hubs, i.e., many vertices with high hub centrality: $x_i = \alpha \sum_j A_{ij} y_j$
  - A vertex has high hub centrality if it is pointing to by many vertices with high authority centrality: $y_i = \beta \sum_j A_{ji} x_j$
- In matrix terms: $x = \alpha A y$  $y = \beta A^T x$
- Combining the two, $AA^T x = \lambda x$  $A^T A y = \lambda y$ where $\lambda = (\alpha \beta)^{-1}$
- Since $AA^T$ and $A^T A$ have the same leading eigenvalue $\lambda$, we have $y = A^T x$
  - This gives a fast way to computer hub centralities from authority ones
- $AA^T$ is the co-citation matrix, and $A^T A$ is the bibliographic coupling matrix
  Authority centrality is eigenvector centrality for the co-citation network
Random Network

- A random network is not a single graph but a statistical ensemble.

- The Gilbert model of a random graph (the $G_{N,p}$ model) for $N=3$ with prob. represented for all the configurations.

- The Edös-Renyi model (the $G_{N,L}$ model) assigns prob. on links.
Degree Distribution in a Random Network

- Degree distribution $P(q)$ is the probability that a randomly chosen node in a random network has degree $q$:
  
  $$P(q) = \frac{\langle N(q) \rangle}{N}$$

  where $\langle N(q) \rangle$ is the avg # of nodes of degree in the network

- In classical random graphs, degree distribution decay quite rapidly: $P(q) \sim \frac{1}{q!}$ for large $q$

  Mean degree $\langle q \rangle = \sum_q q P(q)$ is a typical scale for degrees

- Many real networks (e.g., Internet or cellular nets) have slowly decaying degree distributions (e.g., hub occur with noticeable probability)

  - A dependence w. power law asymptotics $P(q) \sim q^{-\gamma}$ at large $q$
  
  - A scale-free network: a rescale of $q$ by a constant $c \rightarrow cq$ only has the effect of multiplication by a const: $(cq)^{-\gamma} = c^{-\gamma} q^{-\gamma}$
The $\alpha$-model

- The $\alpha$-model has the following parameters or “knobs”:
  - N: size of the network to be generated
  - k: the average degree of a vertex in the network to be generated
  - p: the default probability two vertices are connected
  - $\alpha$: adjustable parameter dictating bias towards local connections

- For any vertices u and v:
  - define $m(u,v)$ to be the number of common neighbors (so far)

- Key quantity: the propensity $R(u,v)$ of u to connect to v
  - if $m(u,v) \geq k$, $R(u,v) = 1$ (share too many friends not to connect)
  - if $m(u,v) = 0$, $R(u,v) = p$ (no mutual friends $\rightarrow$ no bias to connect)
  - else, $R(u,v) = p + (m(u,v)/k)^\alpha (1 - p)$

- Generate new edges incrementally
  - using $R(u,v)$ as the edge probability; details omitted

- Note: $\alpha = \infty$ is “like” Erdos-Renyi (but not exactly)