Network Science: Graph Theory

Ozalp Babaoglu
Dipartimento di Informatica — Scienza e Ingegneria
Università di Bologna
www.cs.unibo.it/babaoglu/

Graph theory

- Branch of mathematics for the study of discrete structures called *graphs* for modeling pairwise relations between objects
- Invented by Swiss mathematician Leonhard Euler (15 April 1707 — 18 September 1783)

- Gives us the language and basic concepts to reason about networks

Graph theory
Terminology and notation

- Formally, a graph is a pair $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ is the set of nodes (vertices) and $\mathcal{E}$ is the set of edges (links, arcs)
- We let n denote the number of nodes and m denote the number of edges in the graph
- Example ($n = 4$, $m = 4$):

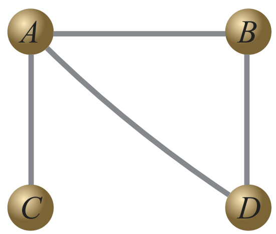
Use letters to label nodes, node pairs to label edges

$\mathcal{N}=\{A, B, C, D\}$

$\mathcal{E}=\{(A, B), (A, C), (A, D), (B, D)\}$

Graph theory
Graph visualization

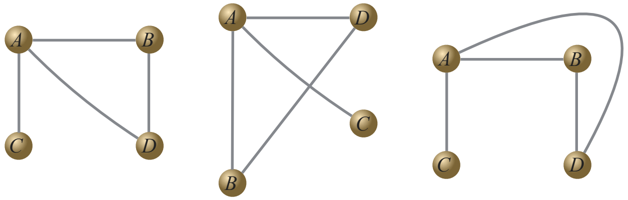
- It is customary to draw the nodes as circles and the edges as lines that join two nodes



- Is a visualization for the graph
$\mathcal{G} = (\{A, B, C, D\}, \{(A, B), (A, C), (A, D), (B, D)\})$

Graph theory
Graph visualization

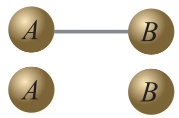
- The graph is defined by the list of nodes and edges, not by its particular visualization
- The same graph may have many different visualizations



- All represent the same graph but some visualizations can be better than others

Graph theory
Binary relations

- Graphs represent arbitrary *binary relations* among objects
- Nodes are the objects, the presence of an edge indicates that some relation $\mathscr{R}$ holds between the nodes, the absence indicates that relation $\mathscr{R}$ does not hold



$A \mathscr{R} B$ is true

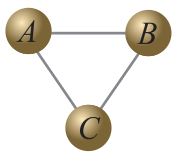
$A \mathscr{R} B$ is false

Examples of binary relation $\mathscr{R}$:

"greater than", "is a friend of", "trusts", "loans money to", "co-authored paper with", "sits on a board-of-directors with"

Graph theory
Binary relations

- Note that binary relations are limiting
- For example, co-authorship among *three* people cannot be expressed through binary relations
- If authors A, B and C publish a paper together, the co-authorship graph will represent this through three binary relations



- But loses the information that they actually co-authored a *common* paper

Graph theory
Directed graphs

- An edge as we have defined it, is undirected and corresponds to a *symmetric* binary relation

$A \mathscr{R} B$ is true and $B \mathscr{R} A$ is true

- An *asymmetric* binary relation holds in one direction only and is represented by a directed edge

$A \mathscr{R} B$ is true and $B \mathscr{R} A$ is false

Examples of *asymmetric* binary relations:

"follows (on Twitter)", "trusts", "connected by a direct flight", "loans money to", "has a URL to"

Graph theory
Directed graphs

- Directed graphs are more general than undirected graphs

is equivalent to

Edge (A, B)

Edges (A, B) and (B, A)

Graph theory
Weighted graphs

- Both directed and undirected graphs can have a *weight* associated with edges to represent the strength of the relation
- Examples of weighted graphs:
  - "co-authorship" (how many joint publications)
  - "actors" (number of joint films)
  - "citations" (number of times one author cites another)
  - "flight routes" (number of daily non-stop flights)
  - "interstate highway" (distance between cities)
  - "Internet" (transmission capacity of a link)

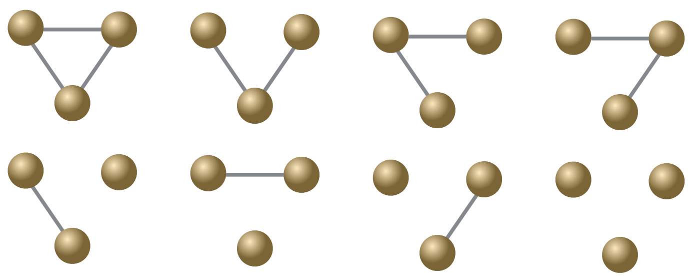
Graph theory
Some basic facts

- What is the maximum number of edges that an undirected graph with n nodes can have?
  - Every node has an edge to every other node
  - Excluding self edges, each node will have $n-1$ edges, for a total of $n(n-1)/2$ edges (corrected for double counting)
  - Thus, for any undirected graph, $m \leq n(n-1)/2$
- How many different undirected graphs with n nodes can there be?
  - There can be at most $n(n-1)/2$ edges
  - Each edge can be present or absent
  - Resulting in a total of $2^{n(n-1)/2}$ combinations

Graph theory
Some basic facts

- How many different undirected graphs with 3 nodes can there be?

$$2^{3(3-1)/2} = 2^3 = 8$$



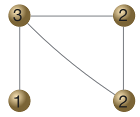
Graph theory
Some basic facts

- How does $2^{n(n-1)/2}$ grow with the number of nodes?

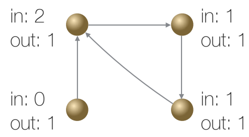
n	$2^{n(n-1)/2}$
5	1,024
6	32,768
7	2,097,152
8	268,435,456
9	68,719,476,736
10	35,184,372,088,832
15	40,564,819,207,303,340,847,894,502,572,032
20	1.569×10^{57}
24	1.214×10^{83}
30	8.872×10^{130}

Node degree

- *Degree* of a node counts the number of edges that are incident on it — its neighbors

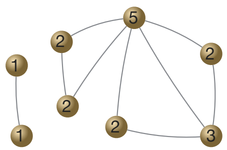


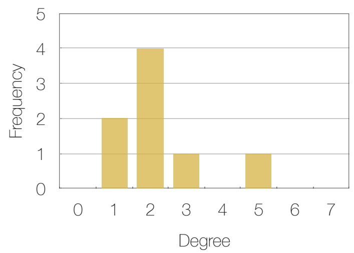
- For a directed graph, we distinguish between the *in-degree* and the *out-degree* of a node



Node degree distribution

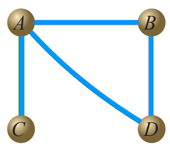
- In a graph with n nodes, the node degrees are in the range between 0 and $n-1$ (excluding self loops)
- How are node degrees *distributed* in this interval?
- Are all degrees equally likely or are some degrees more common than others?





Paths, cycles

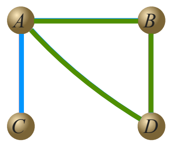
- A *path* in a graph is an alternating sequence of nodes and edges of the graph



CABD
CAD
ADBAC

- If the graph is directed, the path must respect the direction of edges
- A *simple path* is a path where the nodes do not repeat
- A *cycle* is a path where the first and last nodes are the same, but otherwise all nodes are distinct

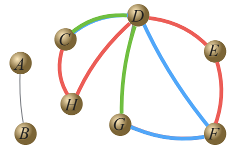
Paths, cycles



- *CABD*: simple path
- *ADBAC*: path but not a simple path
- *BDAB*: cycle

Distance

- The *length* of a path in a graph is the number of steps it contains from beginning to end — the number of edges

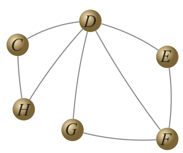


CHDEFG	length 5
CDFG	length 3
CDG	length 2

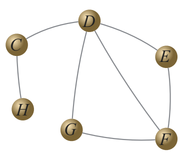
- The *distance* between two nodes in a graph is the length of the shortest path between them
  - Distance between *C* and *G* is 2
  - Distance between *A* and *B* is 1
  - Distance between *A* and *C* is infinite (or undefined)

Diameter

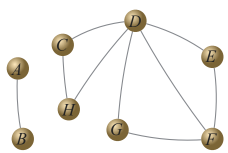
- *Diameter* of a graph is the longest of the distances between all pairs of nodes — the longest shortest path



Diameter 2



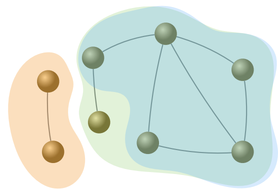
Diameter 3



Diameter ∞

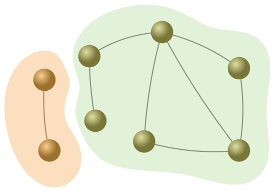
Connectivity, components

- A subgraph is *connected* if there is a path between every pair of nodes
- A *component* of a graph is a maximal connected subgraph

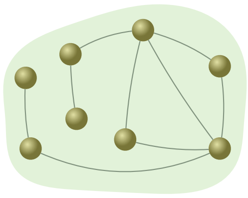


Connectivity, components

- A graph is *connected* if it contains a single component



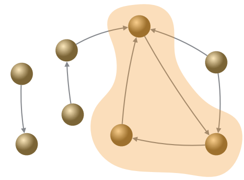
Not connected



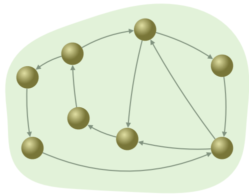
Connected

Connectivity, components

- For directed graphs, definitions extended to *strongly-connected components* and *strongly-connected graphs* taking into consideration the direction of edges



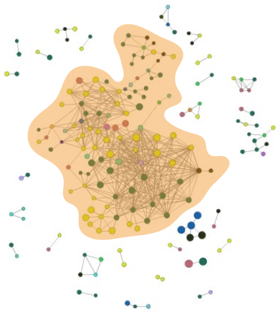
Strongly-connected component

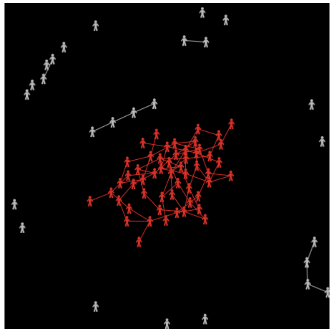


Strongly-connected graph

Giant components

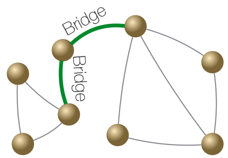
- If the largest component of a graph contains a significant proportion of all nodes, it is called the *giant component*





Bridge

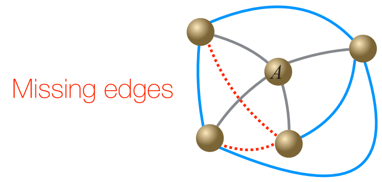
- An edge in a graph is a *bridge* if deleting it increases the number of components of the graph



Clustering coefficient of a node

- Clustering is a measure of how "bunched up" (unevenly distributed) the edges of a graph are
- Formally, the *clustering coefficient* of node A is defined as the probability that two randomly selected *friends* of A are friends themselves
- The fraction of all pairs of A's friends who are also friends
- Defined only if A has at least two friends (otherwise 0)
- The clustering coefficient is always between 0 and 1

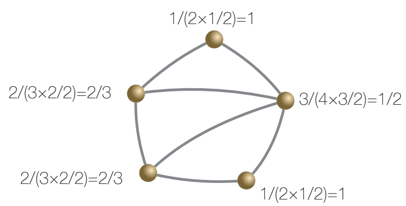
Clustering coefficient of a node



- A has four friends
- Among the four friends, there are (4×3)/2=6 possible friendships
- But only four of them are actually present
- Two are missing
- Thus, the clustering coefficient of node A is 4/6=0.6666

Clustering coefficient of a graph

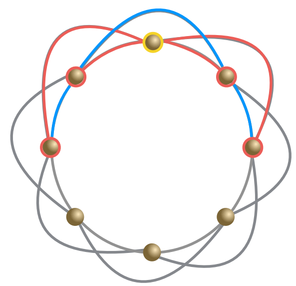
- The clustering coefficient CC of graph $\mathcal{G}$ is the average of the clustering coefficients of all nodes in $\mathcal{G}$



CC = (1+2/3+2/3+1+1/2)/5 = 0.7666

Clustering coefficient of a graph

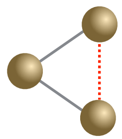
- All nodes are identical and have 4 neighbors



- Possible edges between pairs of neighbors is 4×3/2 = 6
- How many pairs of neighbors are actually connected? 3
- Clustering coefficient of any node: 3/6 = 0.5
- Clustering coefficient of the entire graph: CC = 0.5

Clustering coefficient of a graph

- Clustering quantifies the likelihood that nodes that share a common neighbor are neighbors themselves

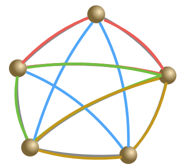


Pick two neighbors
Are they neighbors?

- In social networks, it is very likely that triangles will indeed close over time — *triadic closure*

Clustering coefficient of a graph

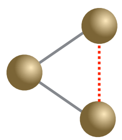
- Alternative definition of clustering coefficient of a graph:
  - Proportion of all possible triangles that are actually closed



- Number of possible triangles is 10 (5 choose 3 = 5!/3!2!)
- Number of closed triangles is 3
- Clustering coefficient is 3/10=0.3 (compare to 0.7666)

Highly clustered

- Recall that clustering quantifies the likelihood that nodes that share a common neighbor are neighbors themselves



Is the “triangle” closed?

- Clustering coefficient of the entire graph, CC, is the proportion of all possible triangles that are actually closed

Highly clustered

- Is CC alone sufficient to conclude that a graph is “highly clustered”?
- CC close to 1 ⇒ highly clustered?
- CC close to 0 ⇒ not highly clustered?
- Not necessarily true!
- Some number of triangles in a graph could be closed simply by chance
- A graph is highly clustered only if the actual likelihood of a triangle being closed is substantially greater than what we would expect due to pure chance

Edge density

- *Edge density* of a graph is the actual number of edges in proportion to the maximum possible number of edges

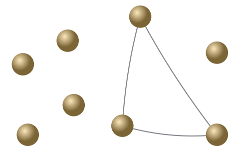
$$\rho = \frac{m}{n(n-1)/2} = \frac{2m}{n(n-1)}$$

- Clearly, the edge density of any graph is between 0 and 1
- Suppose we pick two nodes of a graph at random without regard to the graph structure (e.g., whether the two nodes share a common neighbor or not)
- What is the probability p that the two nodes are connected?
- It is given exactly by the edge density of the graph

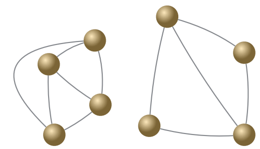
$$p = \rho$$

Sparse and dense graphs

- If ρ is small, then graph is *sparse*
- If ρ is large, then the graph is *dense*



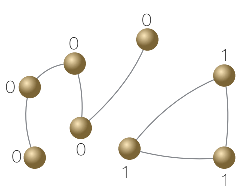
Sparse (ρ=3/(8×7/2)=3/28=0.1071)



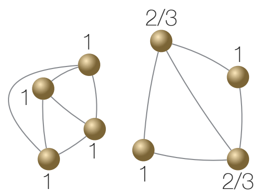
Denser (ρ=11/28=0.3928)

Highly clustered

- We will compare the clustering coefficient CC of a graph to its edge density ρ
- We consider a graph to be *highly clustered* if $CC \gg \rho$



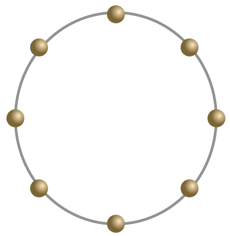
CC = 3/8 = 0.375
ρ = 0.2142
"Not highly clustered"



CC = (6+4/3)/8 = 0.9166
ρ = 0.3928
"Highly clustered"

Highly clustered

- Consider a ring with eight nodes



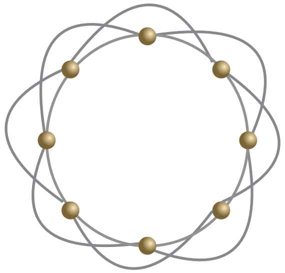
Clustering coefficient: CC=0
Edge density: ρ=2x8/56=0.2857

- What if there are one thousand nodes?

Clustering coefficient: CC=0
Edge density: ρ=2×1000/(1000×999)=0.002

Highly clustered

- Consider an *augmented* ring with eight nodes

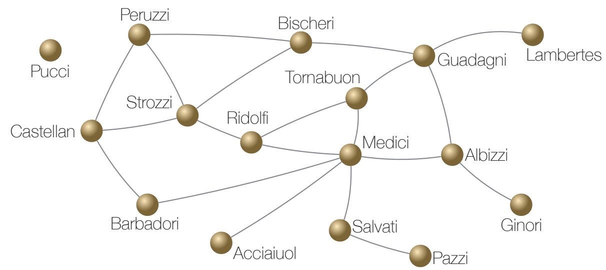


Clustering coefficient: CC=0.5
Edge density: ρ=2x16/56=0.5714

- What if there are one thousand nodes?
  Clustering coefficient: CC=0.5
  Edge density: ρ=2×2000/(1000×999)=0.004

Centrality metrics

- For nodes in a graph, *centrality metrics* try to formalize notions such as "important", "influential" or "popular"



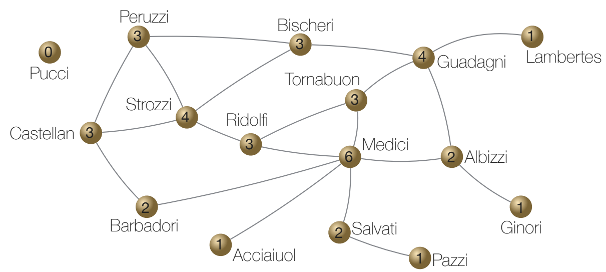
- Why was the Medici an important family in 15th century Florence?

Centrality metrics

- Different notions of centrality
  - Degree — well connectedness
  - Betweenness — criticality for connectedness
  - Closeness — short distances to the rest of the graph
  - Eigenvector — importance
- Centrality is a property of a single node but in the context of the entire graph
- We can also define a global notion of centrality that applies to the entire graph — *centralization*

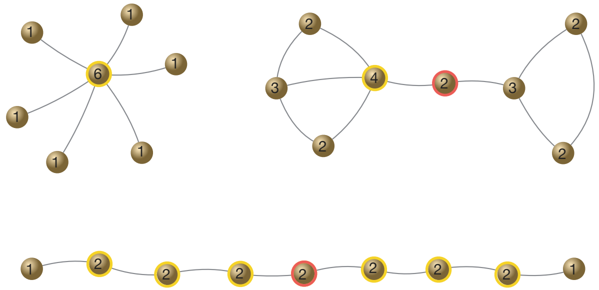
Centrality metrics

- Degree centrality — the greater the degree of a node, the more "important"
- Appropriate for some settings (social networks) since nodes with high degree are better connected and can serve as *introducers*



Centrality metrics

- Problems with degree-based centrality



Betweenness

- Degree-based centrality is not able to capture the notion of *brokerage* — ability of a node in a graph to act as a bridge between different components
- Define *betweenness* of node u to be the fraction of all pairwise shortest paths that go through u

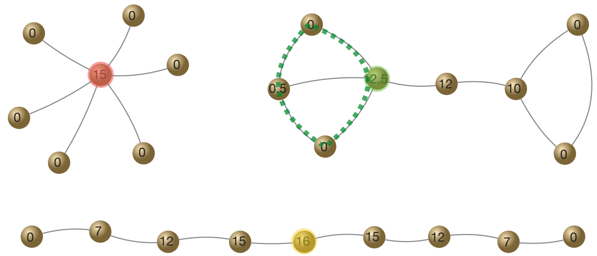
$$B(u) = \sum_{all\ pairs\ i,j} \frac{g_{ij}(u)}{g_{ij}}$$

where

g_{ij} = total number of shortest paths between i, j

$g_{ij}(u)$ = number of shortest paths between i, j that go through u

Betweenness



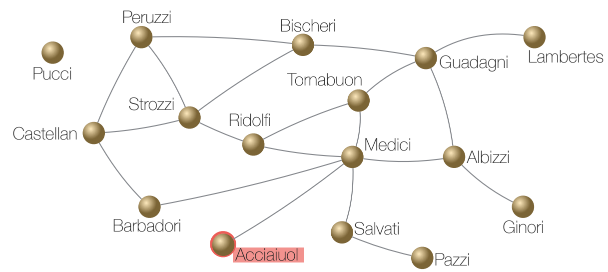
4×4=16 all shortest paths between the 4 nodes to the left and the 4 nodes to the right

6×(6−1)/2=30/2=15 possible pairs among the 6 neighbors of the central node and all shortest paths go through it

4×3+1/2=12.5 the node gets full credit for the 12 shortest paths that go through it but only half the credit for the two shortest paths between the top and bottom nodes

Closeness

- What if it is not important to have many friends
- Or be in a "broker" position?
- Important to be in a "central" position, close to the rest of the graph



- Acciaiuol have degree 1, betweenness 0 but are just one hop from the Medici

Closeness

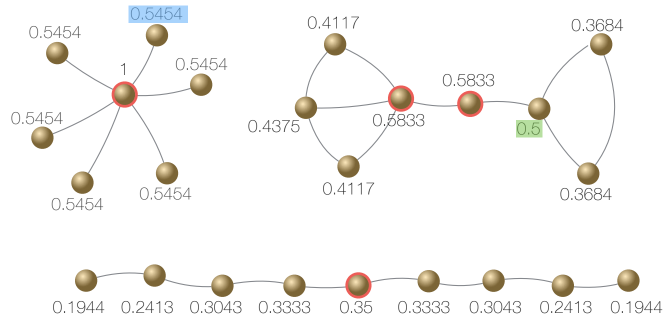
- Define *closeness* of node u based on the (inverse) average shortest path length between node u and every other node in the graph

$$C(u) = \frac{n-1}{\sum_i d(u,i)}$$

where

$d(u,i)$ = length of shortest path between nodes u and i

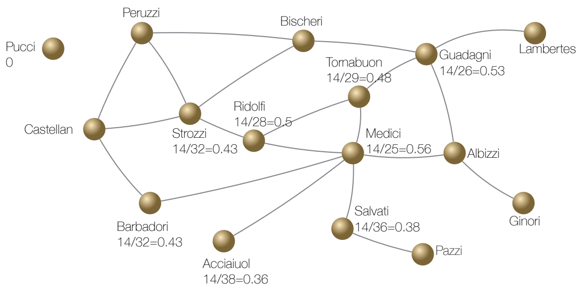
Closeness



6/(1+2+2+2+2+2)=6/11=0.5454

7/(1+1+1+2+3+3+3)=7/14=0.5

Closeness



Centrality metrics in directed graphs

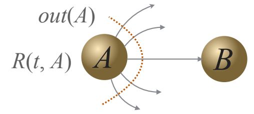
- *Degree*, *betweenness* and *closeness* centrality definitions extend naturally to directed graphs
- Out-degree centrality — based on out-degree
- In-degree centrality — based on in-degree
- Betweenness centrality of a node becomes the fraction of all pairwise shortest *directed* paths that go through it
- In-closeness — based on path lengths from all other nodes to the given node
- Out-closeness — based on path lengths from the given node to all other nodes

Eigenvector centrality

- Basic idea: the importance of a node in a graph is determined by the importance of its neighbors
- Recursive definition!
- Extremely relevant and important for the web graph
- Implemented for directed graphs by the PageRank algorithm that was the main technological innovation behind Google search
- On the web, what counts is not *how many* pages point to a given page but *which* pages point to that page
- The "slashdot effect"

Eigenvector Centrality
Page Rank

- Informally, an *important* node in a directed graph is pointed to by lots of other *important* nodes



$$R(t+1, B) = \sum_{\forall A:\ (A,B) \in E} \frac{R(t, A)}{out(A)}$$

- Let $R(t, A)$ be the rank of A at time t and let $out(A)$ be its out-degree
- A "distributes" its rank evenly over its out-edges so that each one receives $R(t, A)/out(A)$
- The rank of B at time t+1 is obtained by summing the ranks over all of its in-edges

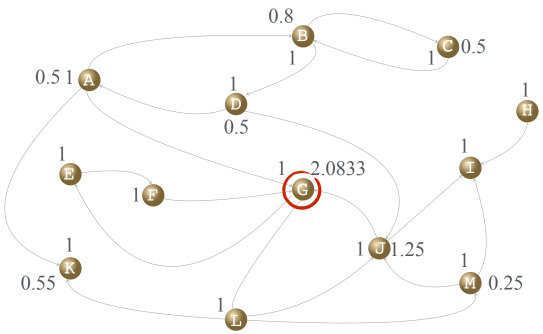
Eigenvector Centrality
Page Rank

- We have an equation like this for every node in the graph:

$$R(t+1, B) = \sum_{\forall A:\ (A,B) \in E} \frac{R(t, A)}{out(A)}$$

- How to assign ranks to all nodes such that the set of equations for the entire graph is consistent (stable)?
- Formally, the solution is equivalent to solving for the *eigenvector* of a matrix (describing the connectivity of the graph)
- Can be approximated algorithmically by iterating — contribution of Larry Page and Sergey Brin while at Stanford that lead to the Google search engine

Eigenvector Centrality
Page Rank



$R(1, G) = R(0, A)/3 + R(0, F) + R(0, J)/2 + R(0, L)/4$
$= 1/3 + 1 + 1/2 + 1/4$
$= 2.0833$

Recap
Classes of graph properties

- Global patterns — *macroscopic* aspects of graph structure
  - Degree distribution
  - Connectivity
  - Path lengths
  - Diameter
  - Edge density
- Local patterns — *microscopic* aspects of graph structure
  - Degree
  - Clustering coefficient
- Centrality — a single node in context (position) of graph
  - Betweenness
  - Closeness

Software tools

- Gephi: interactive visualization and exploration platform for networks
  - https://gephi.github.io/
- NetLogo: programmable multi-agent environment for modeling network dynamics
  - https://ccl.northwestern.edu/netlogo/