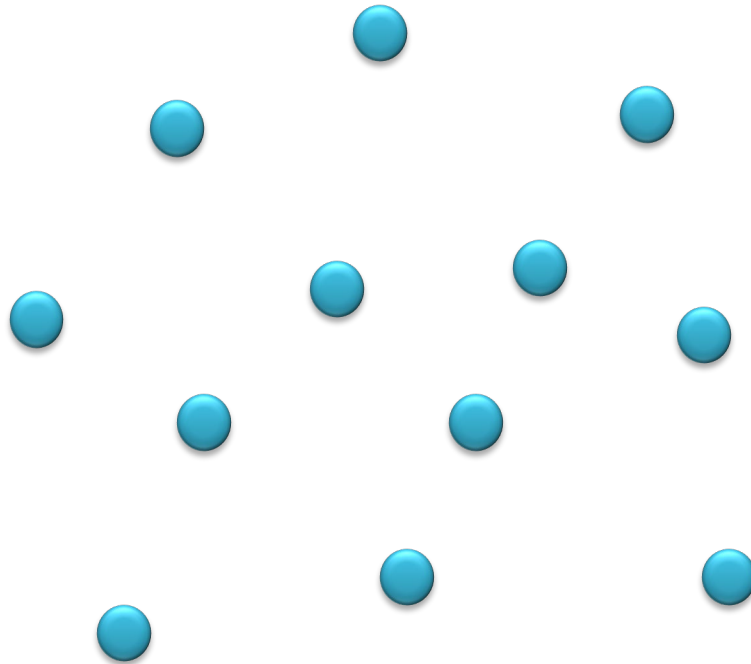


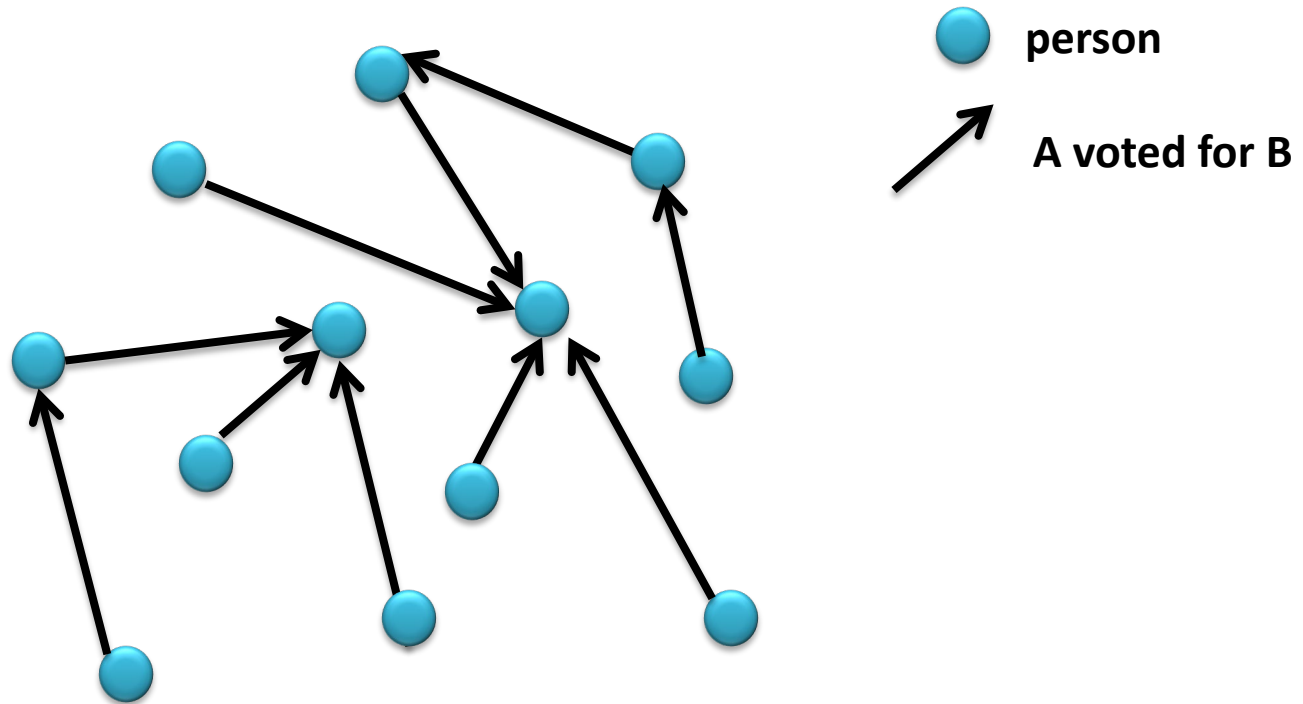
# Centrality indices

# Complex Social System, Elections



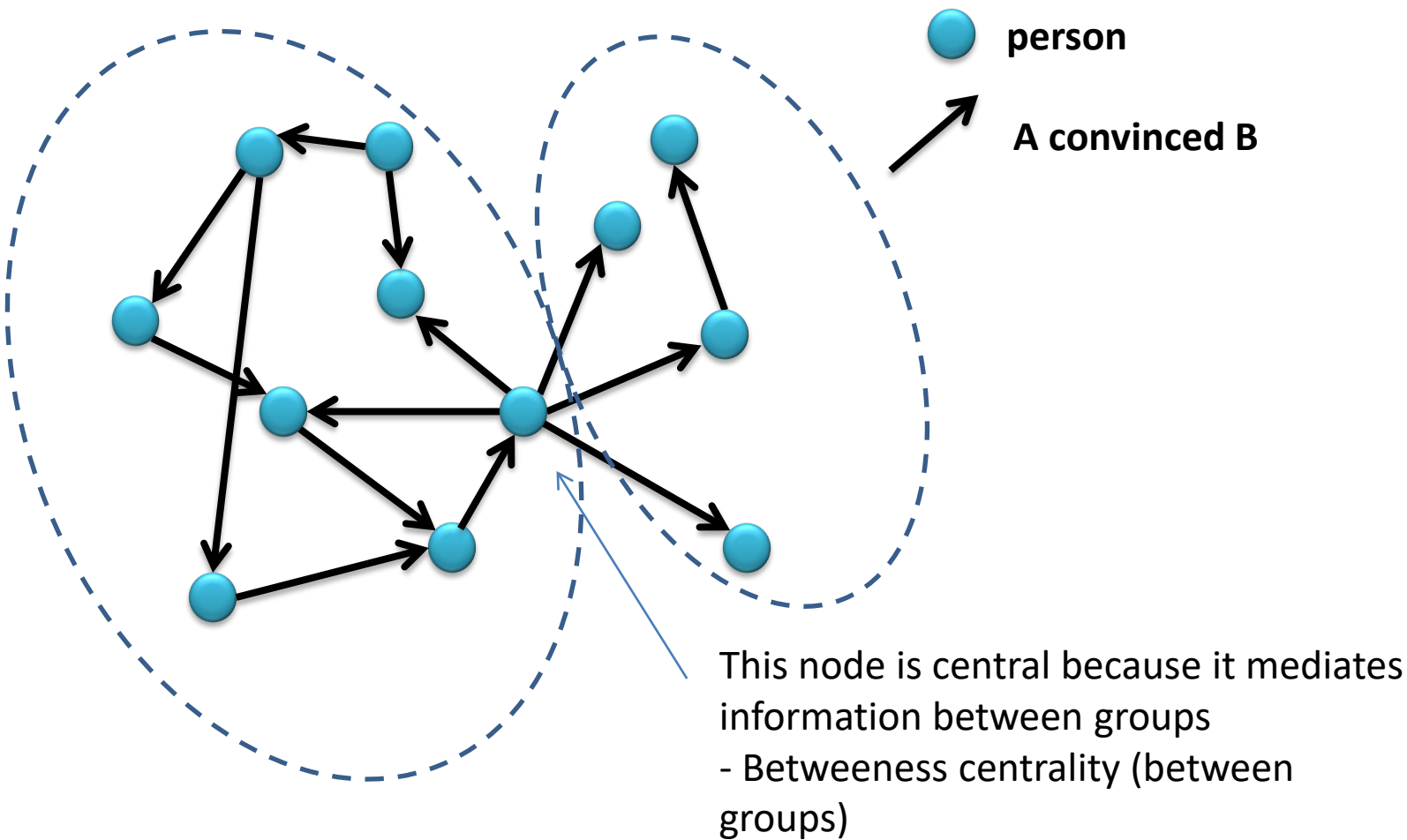
Centrality measures an importance of such network elements as nodes and edges.

# Complex Social System, Network I



A is more “central” than B if more people voted for A  
In-degree centrality index = number of in-edges

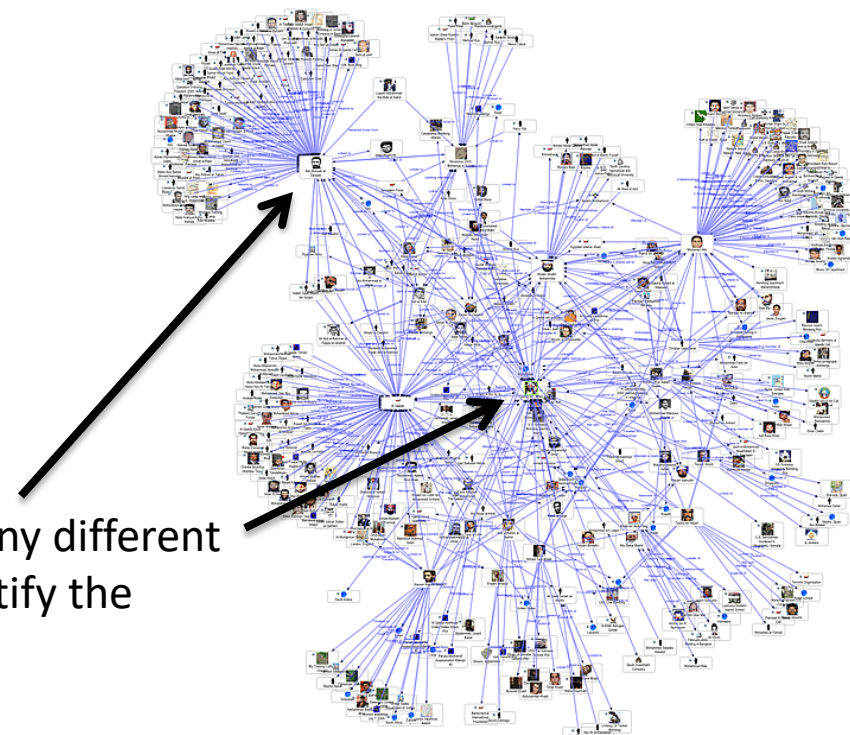
# Complex Social System, Network II



Another example: friendship network, degree-based centrality

# Measures and Metrics

Q: What are the most important nodes and edges?



There are many different ways to quantify the importance

- Degree centrality:  $\forall i \in V \ d(i)$
- Closeness centrality:  $\forall i \in V \ l_i = \frac{1}{n} \sum_{j \in V} \delta_{ij}, \delta_{ij} = \text{distance from } i \text{ to } j$
- Edge removal centrality:  $\forall e \in E \ r_e = \frac{\sum_{i,j \in V(G)} \delta_{ij}}{\sum_{i,j \in V(G|_e)} \delta_{ij}}$

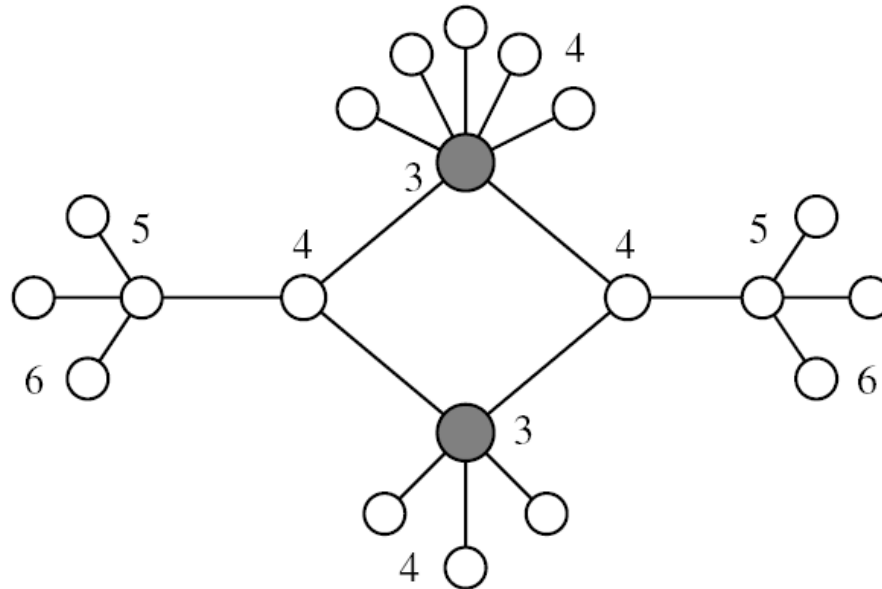


Can you find a new interesting measure of importance? Does it work for many different types of networks? Can you compute it efficiently?

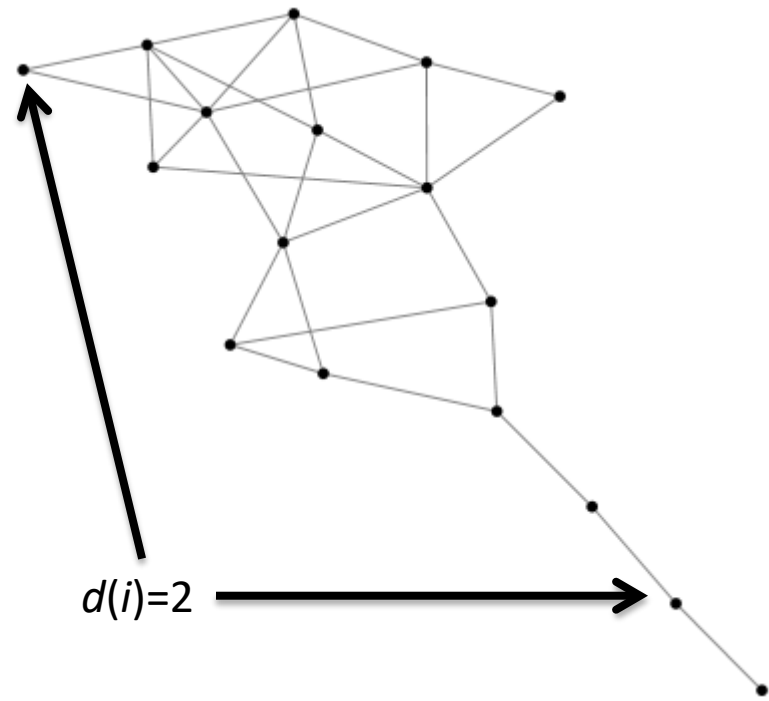
# Eccentricity

$\forall i \in V$  we define  $e(i) = \max\{\delta_{ij} | j \in V\}$

Eccentricity centrality  $c_E(i) = 1/e(i)$



# Eigenvector Centrality



# Eigenvector Centrality

- $\forall i \in V \ x_i^{(0)} = 1$
- $\forall i \in V \ x_i^{(1)} = \sum_j A_{ij} x_j^{(0)}$
- $\forall i \in V \ x_i^{(2)} = \sum_j A_{ij} x_j^{(1)}$

- - - - degree centrality - - - -

Repeating last step  $t$  times

$$\mathbf{x}^{(t)} = A^t \mathbf{x}^{(0)}$$

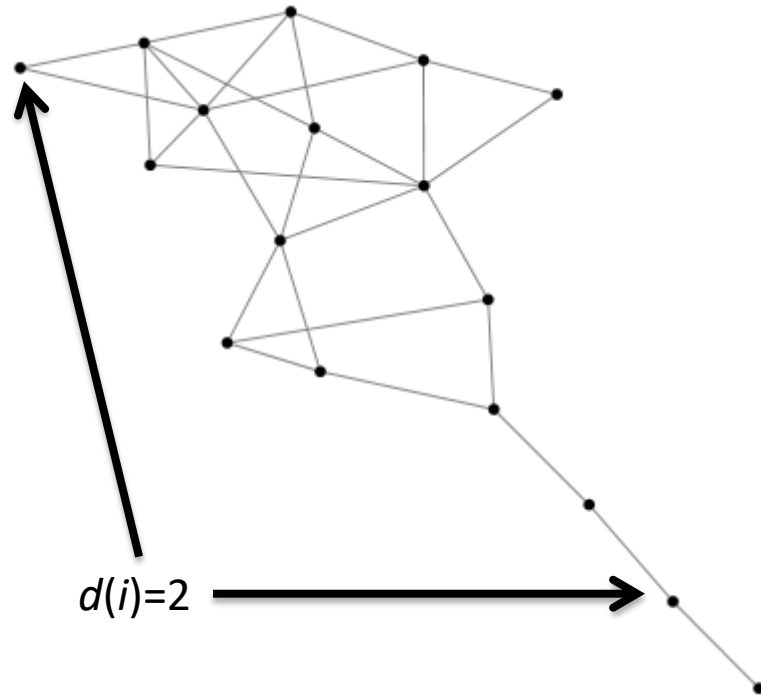
Rewrite  $\mathbf{x}^{(0)}$  with linear combination of eigenvectors of  $A$

$$\mathbf{x}^{(t)} = A^t \sum_i c_i \mathbf{v}_i = \sum_i c_i \lambda_i^t \mathbf{v}_i = \lambda_1^t \sum_i c_i \left(\frac{\lambda_i}{\lambda_1}\right)^t \mathbf{v}_i, \text{ where } \lambda_1 \geq \dots \geq \lambda_n, \lambda_i \in \Lambda(A)$$

$$\lim_{t \rightarrow T} \mathbf{x}^{(t)} = c_1 \lambda_1^T \mathbf{v}_1$$

it is proportional to the first eigenvector

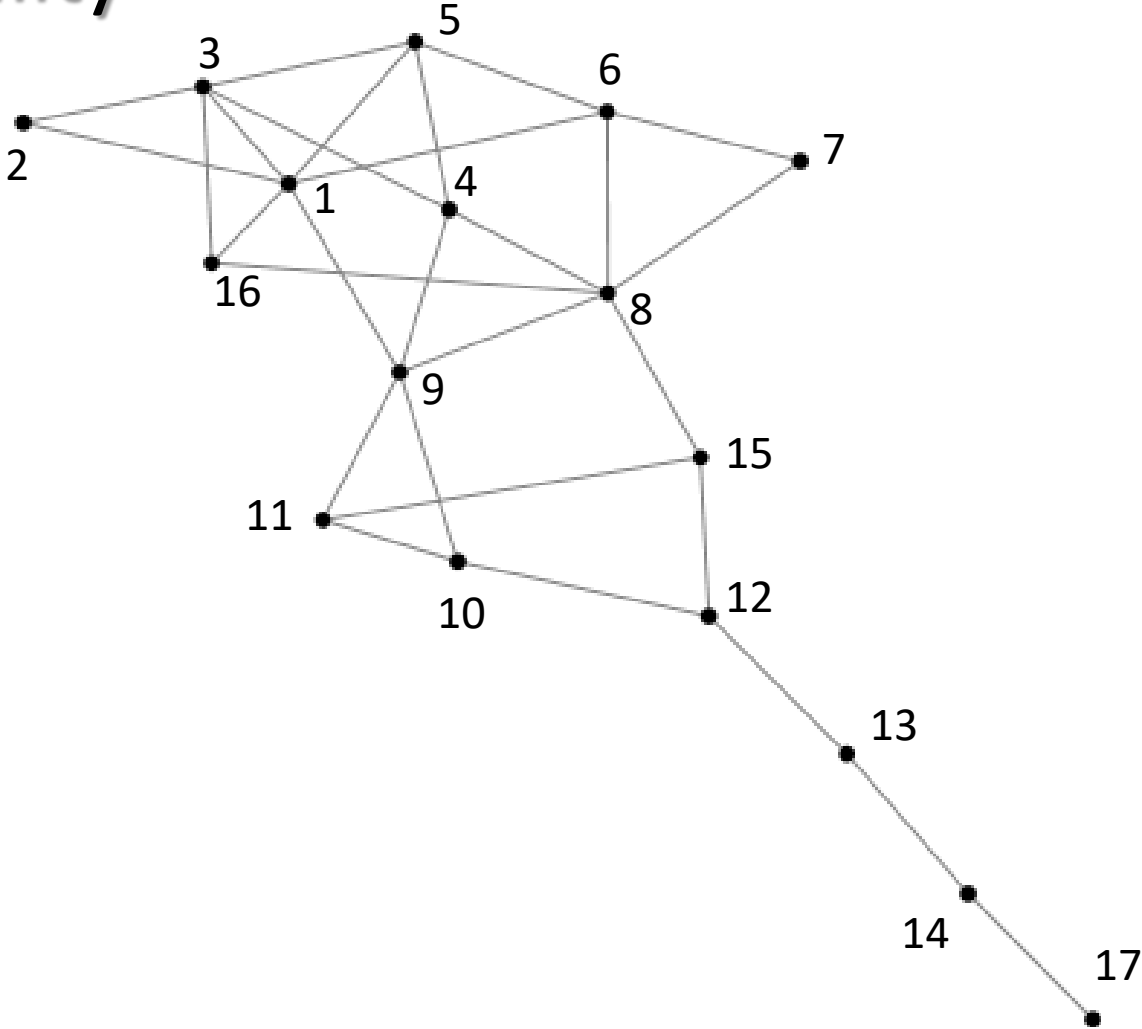
The eigenequation is  $Ax = \lambda_1 x$ , and the eigenvector centrality is defined by values

$$x_i = \lambda_1^{-1} \sum_j A_{ij} x_j$$


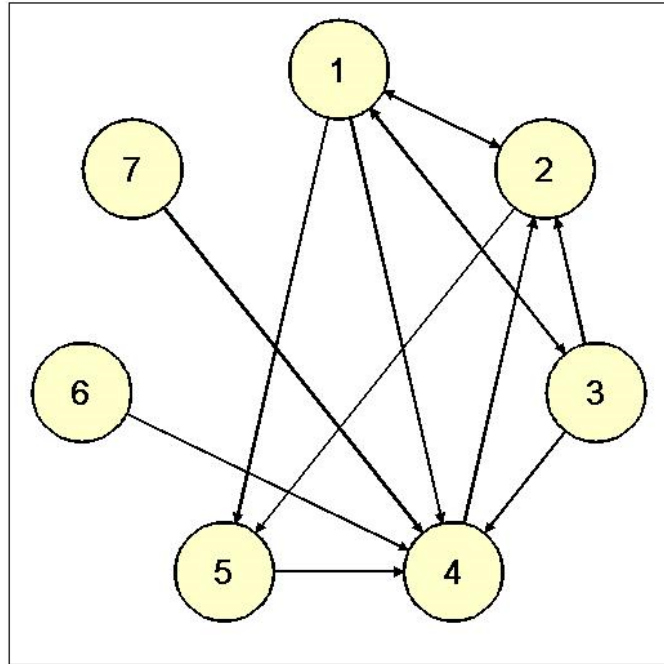


# Eigenvector Centrality

1	0.120
3	0.103
8	0.101
5	0.095
4	0.093
9	0.092
6	0.085
16	0.077
2	0.053
7	0.044
11	0.039
15	0.038
10	0.035
12	0.018
13	0.005
14	0.001
17	0.000

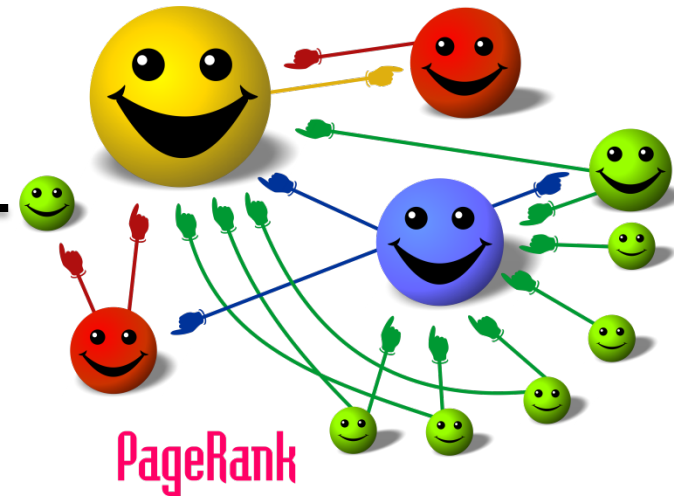


# PageRank



$$P = \begin{bmatrix} 0 & 0.50 & 0.33 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.25 & 0.00 & 0.33 & 0.50 & 0.00 & 0.00 & 0.00 \\ 0.25 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.25 & 0.00 & 0.33 & 0.50 & 1.00 & 1.00 & 1.00 \\ 0.25 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$

transition matrix



$$x = \left(\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}\right)$$

$P^{20}x$  - see pagerank-example.m

Markov Chain: probabilities of visiting the pages after  $k$  steps is  $P^kx$ .

Problem: dangling nodes ( $d^+(i)=0$ )

Solution: damping factor  $\alpha$  (usually small)

$$P' = (1 - \alpha)P + \alpha T, \text{ where } T = \frac{1}{n} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$$

↑ regular random walk

← teleportation

Note:  $P'$  is still a Markov matrix with no 0s

# Eigenproblem-based Centralities: Computational Problems

- These problems are equivalent to solving eigenproblems, namely,  $P'v = v$ .
- Gaussian elimination is very expensive.

$$P' = (1 - \alpha)P + \alpha T, \text{ where } T = \frac{1}{n} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

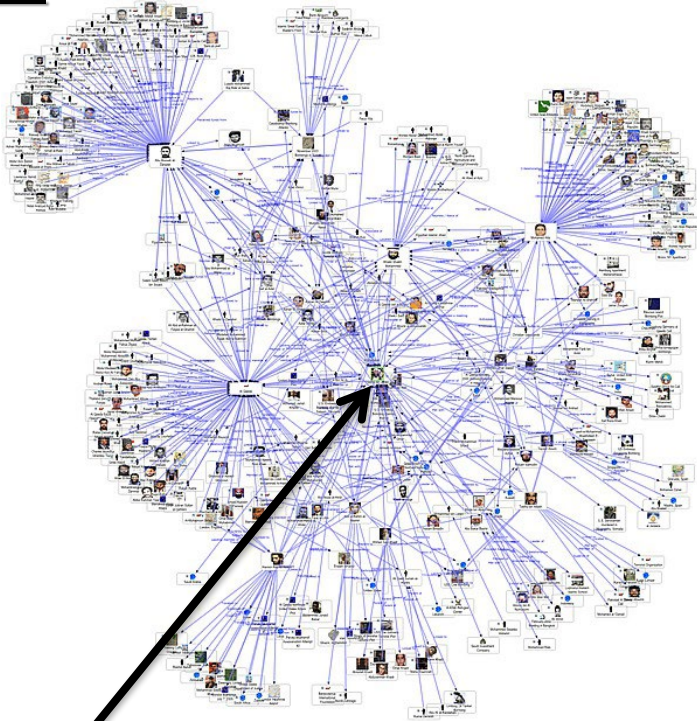
- Iterative computation of solving the system with  $P'$  is less expensive.
- The principal eigenvector of  $P'$  is a PageRank. It can be solved by Power Method (iterative), Algebraic Multigrid, etc.
- Parallel computation
- Sublinear computation to compute individual entries
- Additional problems include updating PageRank vector with evolving networks

Further reading:    Berkhin “A Survey on PageRank computing”  
                          Langville, Meyer “Deeper Inside PageRank”

# Hubs and Authorities

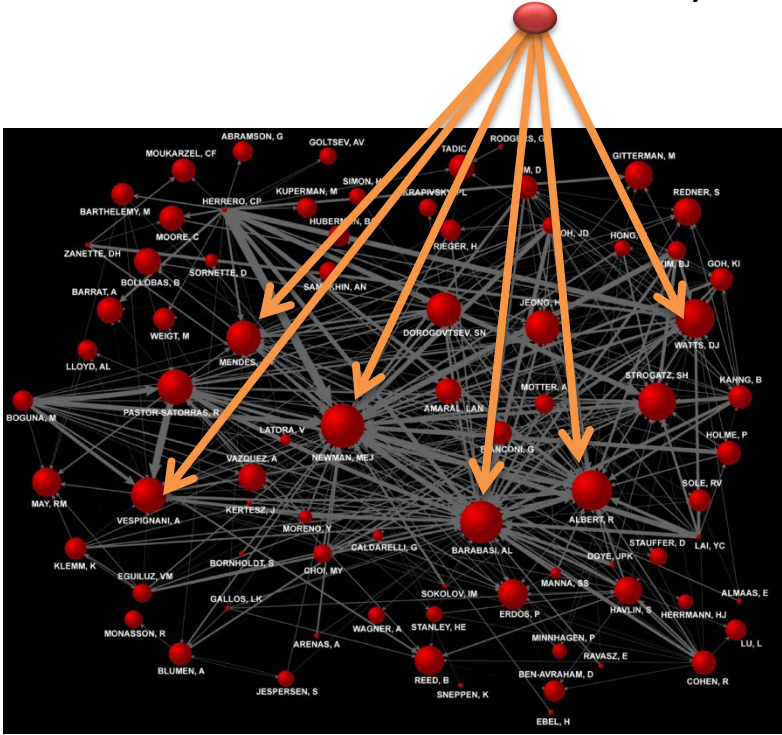
Observation: In some cases a vertex may be important if it points to others with high centrality.

Examples:



X sends an information to many leaders who don't know X, i.e., X is influential.

X, Y, Z “Cool Survey”



## Scientific paper citations

Radicchi et al. “Diffusion of scientific credits and the ranking of scientists”

A review article (hub) may cite other articles (authorities) that are authoritative sources for information

# Hubs and Authorities, HITS Algorithm

**Authorities** are nodes that contain useful information on a topic of interest.

**Hubs** are nodes that point to best authorities.

**HITS Algorithm** computes for each node  $i$  its authority and hub centralities  $x_i$  and  $y_i$ , respectively

$$x_i = \alpha \sum_j A_{ij} y_j \text{ and } y_i = \beta \sum_j A_{ji} x_j, \text{ where } \alpha, \beta \text{ are constants}$$

$$\implies AA^T x = \lambda x \text{ and } A^T A y = \lambda y, \text{ where } \lambda = (\alpha\beta)^{-1}$$

same leading (see eigenvector centrality) eigenvalue in both cases!

# Hubs and Authorities, HITS Algorithm

**Authorities** are nodes that contain useful information on a topic of interest.

**Hubs** are nodes that point to best authorities.

**HITS Algorithm** computes for each node  $i$  its authority and hub centralities  $x_i$  and  $y_i$ , respectively

$$x_i = \alpha \sum_j A_{ij} y_j \text{ and } y_i = \beta \sum_j A_{ji} x_j, \text{ where } \alpha, \beta \text{ are constants}$$
$$\implies AA^T x = \lambda x \text{ and } A^T A y = \lambda y, \text{ where } \lambda = (\alpha\beta)^{-1}$$

same leading (see eigenvector centrality) eigenvalue in both cases!

All eigenvalues of  $AA^T$  and  $A^T A$  are the same (check and prove!).

Multiplying both sides of the first equation by  $A^T$  gives

$$A^T \cdot AA^T x = A^T \cdot \lambda x \implies A^T A(A^T x) = \lambda(A^T x) \implies y = A^T x$$

i.e., once we have a vector of authorities, the hub centrality can be calculated faster

Note: Authority centrality of  $A$  = Eigenvector centrality of co-citation matrix  $AA^T$

Current applications: Teoma.com and Ask.com

**Further reading:** Kleinberg, “Authoritative sources in a hyper-linked environment”

# Homework

## Paper review 1

Jon Kleinberg “Authoritative Sources in a Hyperlinked Environment”

Submit your review in Canvas by 9/12.

Note: Your review must be **detailed**, i.e., include short summary, and detailed constructive criticism. Don't copy-paste the paper!

# Closeness Centrality

$\delta_{ij}$  = length of  $i - j$  shortest path  
 $\forall i \in V \ C_i = 1/l_i$ , where  $l_i = \frac{1}{n} \sum_{j \in V} \delta_{ij}$

large for peripheral vertices

## Problems:

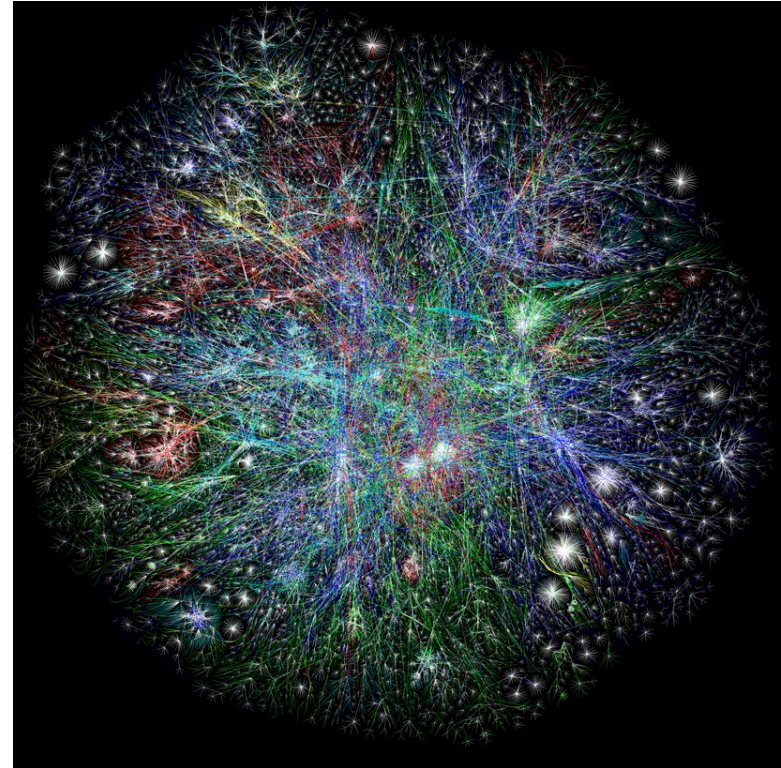
- $i-j$  path is infinity for different connected components

**Solution:** Harmonic Mean Distance Centrality  $\forall i \in V \ C'_i = \frac{1}{n-1} \cdot \sum_{j \neq i} 1/\delta_{ij}$

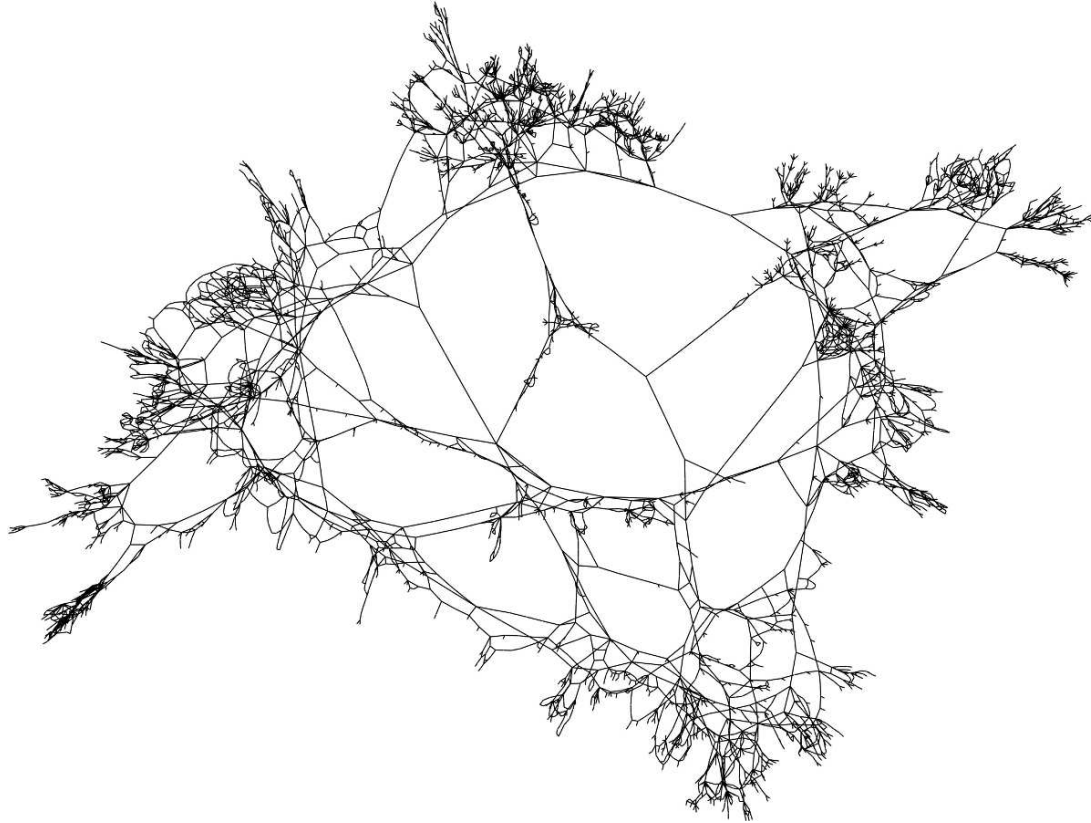
- In practice  $C$  spans relatively small range of values, i.e., leaders are not well separated
- Highly sensitive (one edge removal and all distances are increased)
- Geodesic distances are integers, **often small because distance increases logarithmically with the size of the network**
- Example: Internet Movie Database  
highest centrality = 0.4143, lowest centrality = 0.1154, ratio 3.6 for 500K actors



# Closeness Centrality



# Closeness and Degree-based Centralities



Example in Gephi: degrees vs eigenvector centrality  
Some correlation is expected for certain types of networks

# Enumeration of Shortest Paths-based Centrality

$\sigma_{st}(i)$  is a number of  $s$ - $t$  shortest paths containing  $i$

$\sigma_{st}$  is a number of all  $s$ - $t$  shortest-paths

**Observation:** In practice, communication or transport of goods in networks follow different kinds of paths that tend to be shortest.

**Intuitive Question:** How much work can be done by a node?

Disadvantage?

## ● Stress Centrality

$$c_S(i) = \sum_{s \neq i} \sum_{t \neq i} \sigma_{st}(i)$$

for nodes

$$c_S(ij) = \sum_{s \in V} \sum_{t \in V} \sigma_{st}(ij)$$

for edges

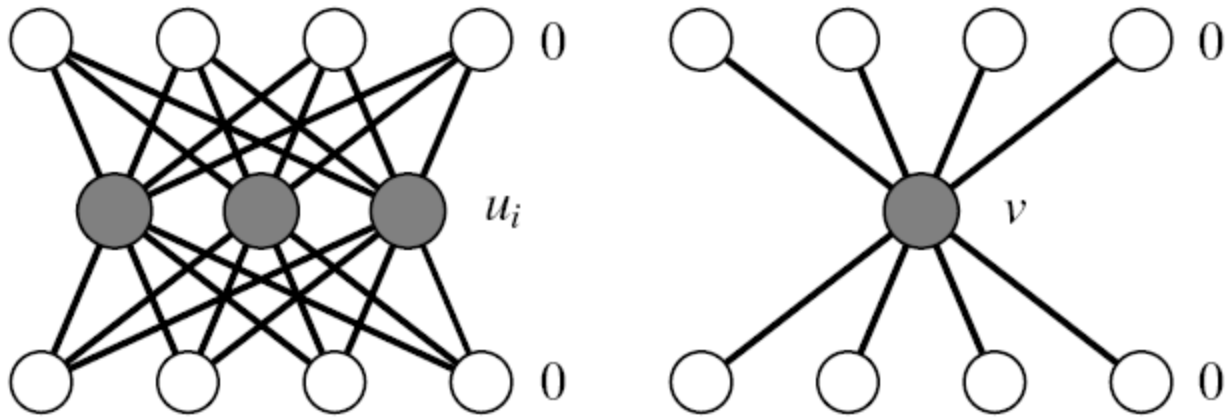
Relation between stress centralities  $c_S(i) = \frac{1}{2} \sum_{ij \in \Gamma(i)} c_S(ij) - \sum_{i \neq s \in V} \sigma_{si} - \sum_{i \neq t \in V} \sigma_{it}$

## ● Betweenness Centrality

$$c_B(i) = \sum_{s \neq i} \sum_{t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Interpretation: BC is a quantification of communication control which has a vertex over all pairs of nodes.

# Communication Control Quantification: Example



$c_S(u_i) = 16$  and  $c_B(u_i) = \frac{1}{3}$ ,  $i = 1, 2, 3$  and  $c_S(v) = 16$  but  $c_B(v) = 1$   
 [BE] "Network Analysis"

## BC for edges

$$c_B(ij) = \sum_{s \in V} \sum_{t \in V} \frac{\sigma_{st}(ij)}{\sigma_{st}}$$

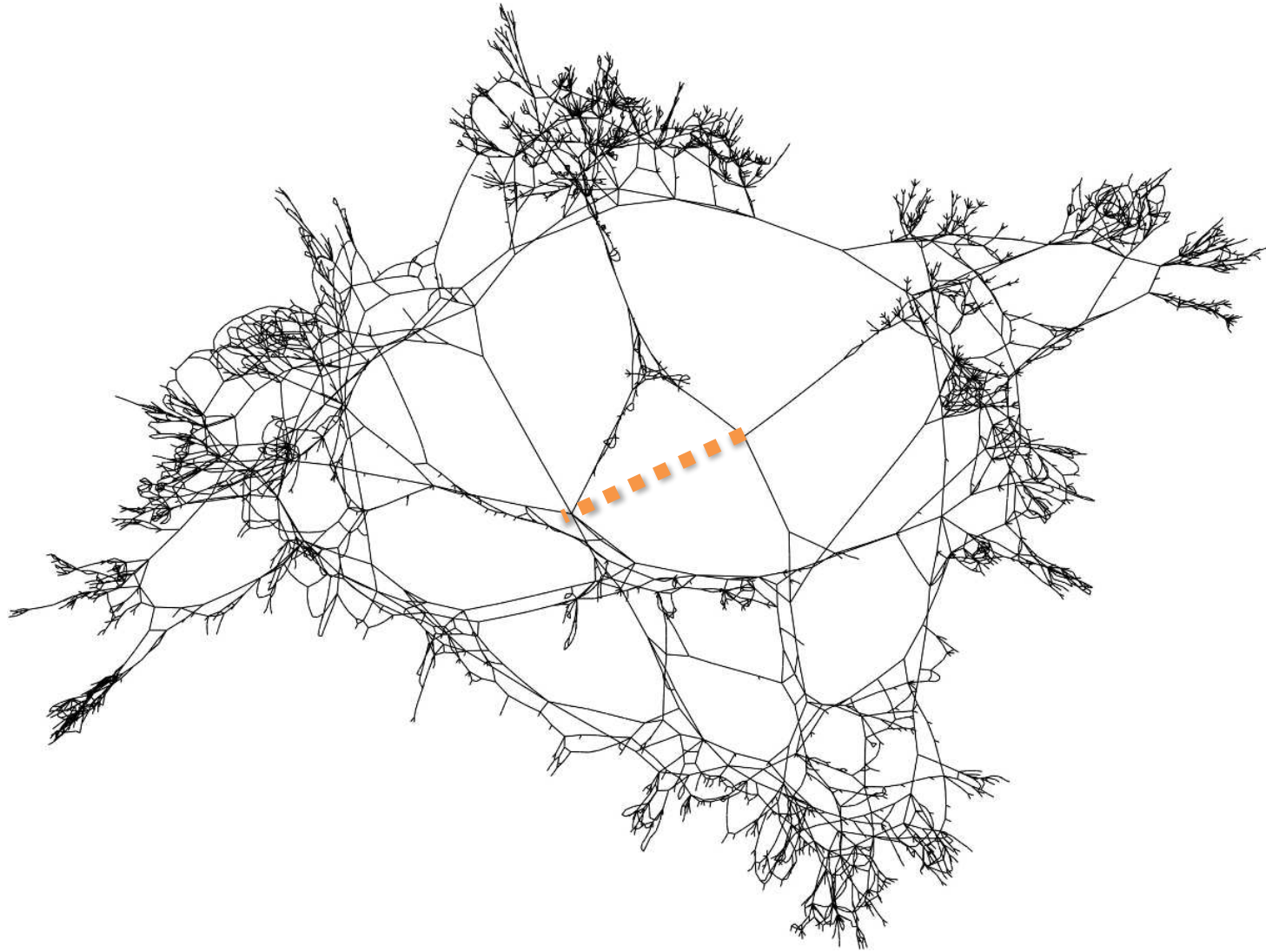
## Problem

BC is very sensitive to network dynamics (edge/node removal/addition)

## Solution?

$\epsilon$ -BC: in BC replace all shortest s-t paths with all shortest paths that are not longer than  $(1+\epsilon)\delta_{ij}$

# We have to be careful with shortest path-based centralities, example



**Homework (due 9/12):** Prove that in directed graphs the relation between centralities holds

$$c_B(i) = \sum_{ij \in \Gamma^+(i)} c_B(ij) - (n - 1) = \sum_{ji \in \Gamma^-(i)} c_B(ji) - (n - 1)$$

# Traversal Sets

$\forall ij \in E$  we define the edge's traversal set

$$T_{ij} = \{(s, t) \in V \times V \mid \text{some shortest path } s - t \text{ contains } ij\}$$

and traversal set induced graph

$$G[T_{ij}] = (V', E'), \text{ where } V' = \{k \in V \mid (k, t) \in T_{ij} \text{ or } (s, k) \in T_{ij}\}, \text{ and} \\ E' = \{(s, t) \in T_{ij}\}.$$

**Homework (due 9/17):** Prove that  $G[T_{ij}]$  is bipartite.

$|T_{ij}|$  is an illuminating measure

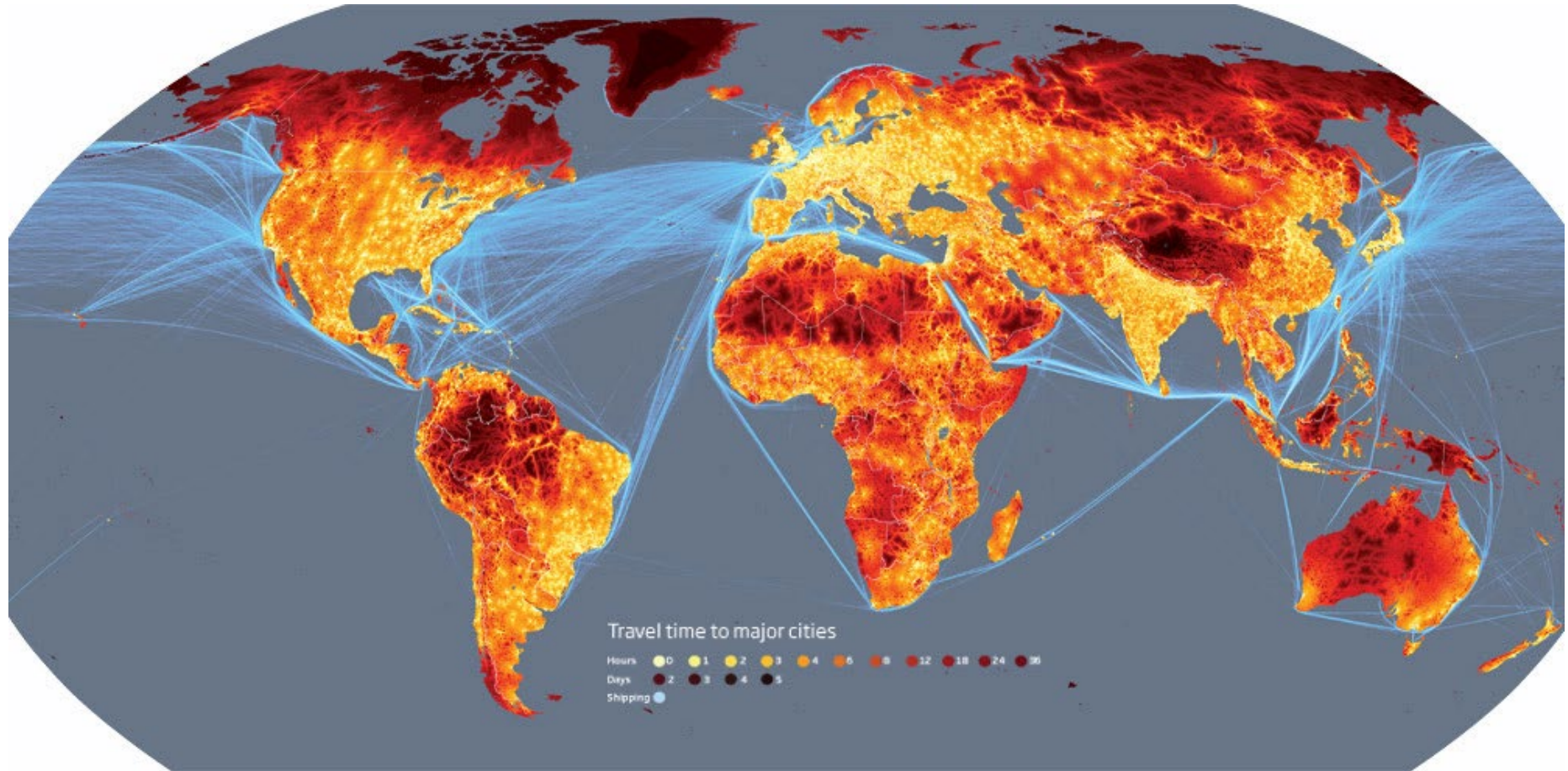
Another measure based on  $T_{ij}$  is the edge centrality index

$$c_{ts}(ij) = |H|, \text{ where } H \text{ is a minimum vertex cover in } G[T_{ij}]$$

that can be used in characterization of networks with hierarchical organization.

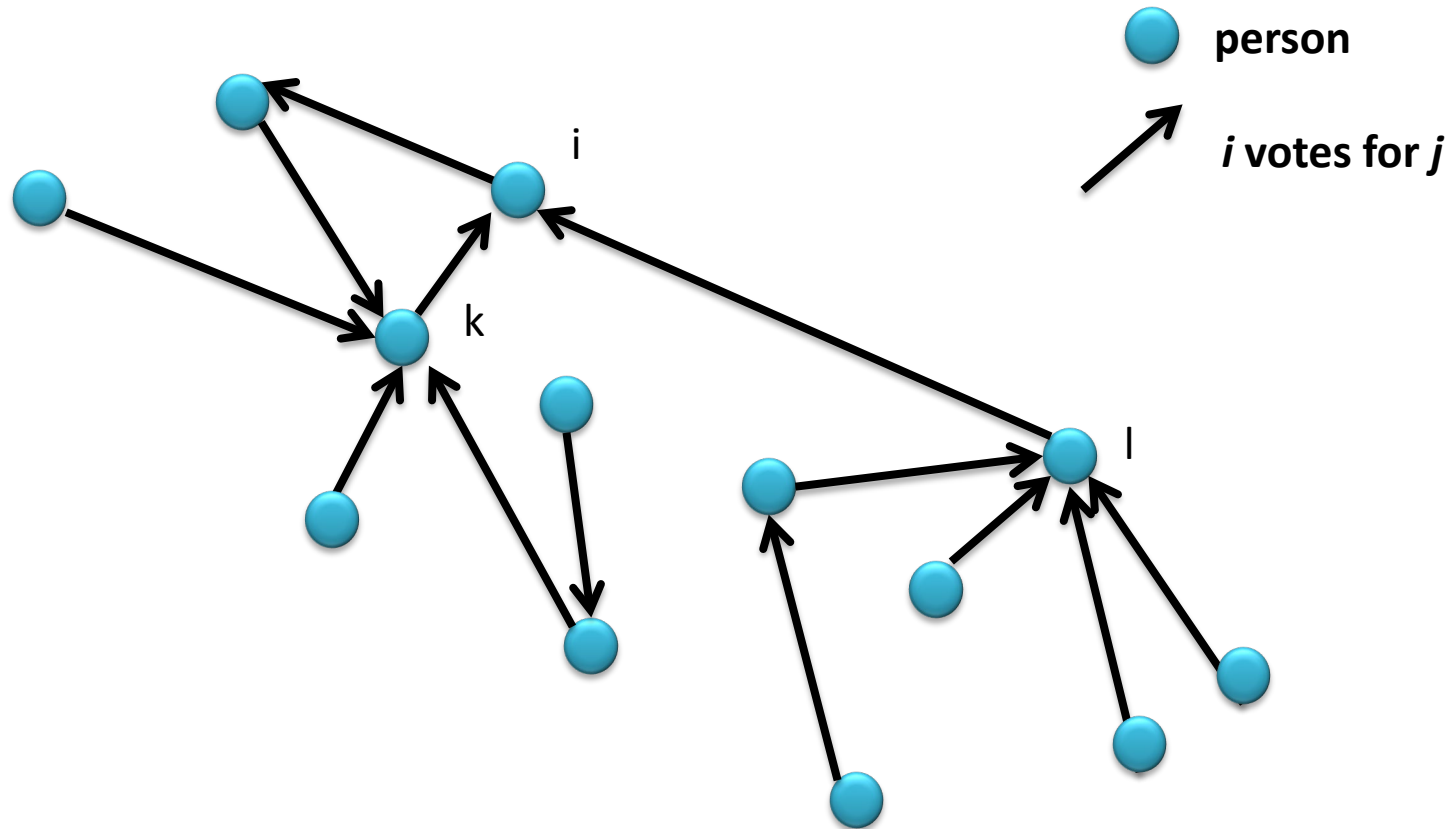
**Note:** there is a theorem that says the following: in bipartite graphs the minimum size of a vertex cover = the size of a maximum matching

# Example of Hierarchical Organization: Transportation Roads Density





# Some ideas behind the feedback centralities

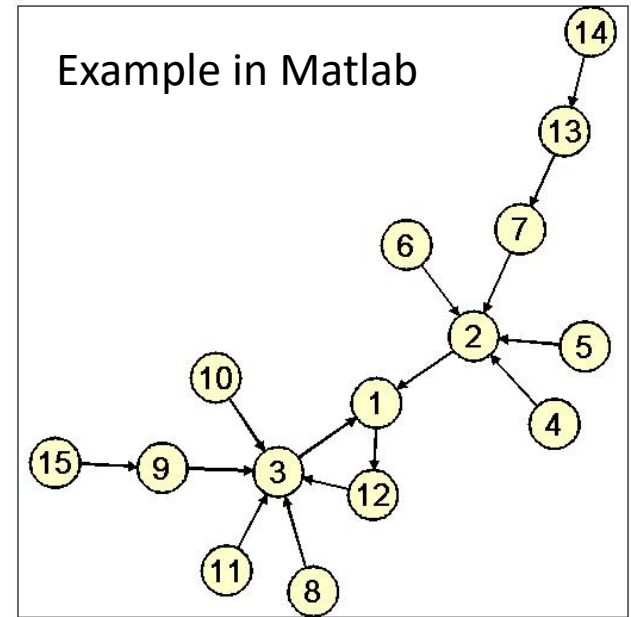


# Counting All Paths

Simple, directed  $G = (V, E)$ , no loops

$$\forall i \in V \quad c_K(i) = \sum_{k=1}^{\infty} \sum_{j=1}^n \alpha_k (A^k)_{ji}$$

The sum converges with restricted  $\alpha_k$ .



**Theorem.** If  $A$  is the adjacency matrix of  $G$ ,  $\alpha > 0$ , and  $\lambda_1$  the largest eigenvalue of  $A$ , then

$$\lambda_1 < 1/\alpha \iff \sum_{k=1}^{\infty} \alpha^k A^k \text{ converges}$$

and  $c_K = (I - \alpha A)^{-1} \cdot \mathbf{1}_n$ .

Leo Katz “A new status index derived from sociometric analysis”

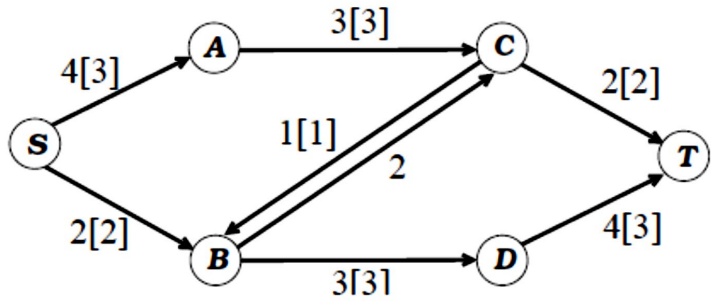
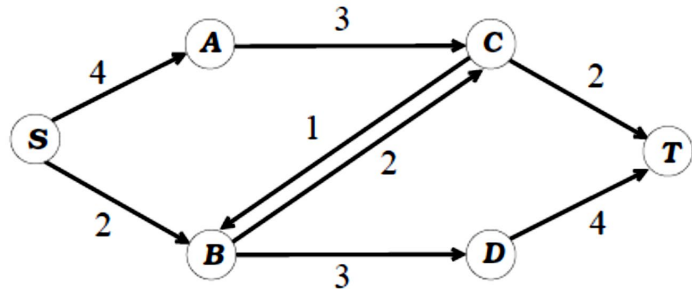
# Network Flow

A flow network is given by a directed graph  $G = (V, E)$ , edge capacity function  $u : E \rightarrow \mathbb{R}_{\geq 0}$ , and two distinct nodes  $s, t \in V$ . A flow from  $s$  to  $t$  is a function  $f : E \rightarrow \mathbb{R}$  satisfying the following constraints

- Capacity:  $\forall e \in E : 0 \leq f(e) \leq u(e)$
- Balance:  $\forall v \in V \setminus \{s, t\}$ :

$$\sum_{e \in \Gamma^-(v)} f(e) = \sum_{e \in \Gamma^+(v)} f(e)$$

The value of the flow  $f$  is defined as  $\sum_{e \in \Gamma^+(s)} f(e)$ .



Goldberg and Tarjan solved it in  $O(nm \log(n^2/m))$ . Ford-Fulkerson theorem says that the value of a maximum s-t-flow = the capacity of minimum s-t-cut. 27

# Vitality (robustness)

Let  $\mathcal{G}$  be the set of all simple, undirected and unweighted graphs  $G = (V, E)$  and  $f : \mathcal{G} \rightarrow \mathbb{R}$  be any real-valued function on  $G \in \mathcal{G}$ . A vitality index  $\mathcal{V}(G, x)$  is the difference of the values of  $f$  on  $G$  and on  $G$  without element  $x$ , i.e.,  $\mathcal{V}(G, x) = f(G) - f(G \setminus x)$ .

## Max-flow Betweenness Vitality

**Q:** How much flow must go over a vertex  $i$  in order to obtain the maximum flow value?  
How does the objective function value change if we remove  $i$  from the network?

$$c_{mf}(i) = \sum_{\substack{s,t \in V \\ i \neq s, i \neq t \\ f_{st} > 0}} \frac{f_{st}(i)}{f_{st}}, \text{ where } f_{st}(i) = f_{st} - \text{max s-t-flow in } G \setminus i$$

.

Examples of vitality: power grids with removed connections, social networks with no leader, collaboration networks

# Closeness Vitality

Wiener index of a network

$$I_W(G) = \sum_{i,j \in V} \delta_{ij}$$

*i-j* shortest path

or in terms of closeness centrality

$$I_W(G) = n \cdot \sum_{i \in V} \frac{1}{C_i}$$

Closeness vitality is defined on both vertices and edges

$$c_{CV}(x) = I_W(G) - I_W(G \setminus \{x\})$$

Computational problem with this vitality index?

# Stress Centrality as a Vitality Index

$\sigma_{st}(i)$  is a number of  $s$ - $t$  shortest paths containing  $i$

$\sigma_{st}$  is a number of all  $s$ - $t$  shortest-paths

## Stress Centrality

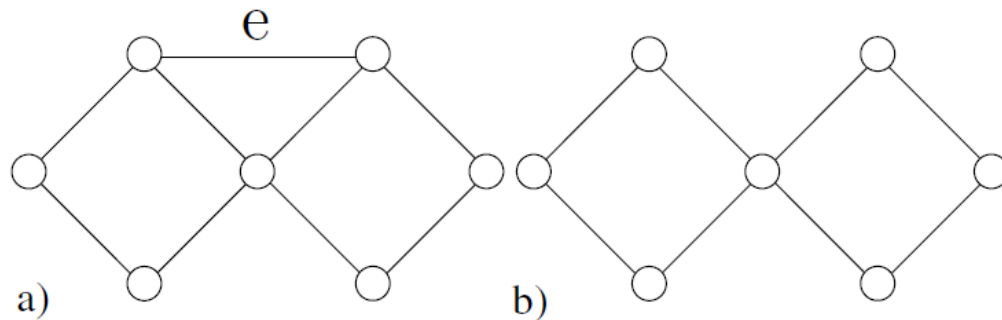
$$c_S(i) = \sum_{s \neq i} \sum_{t \neq i} \sigma_{st}(i)$$

for nodes

$$c_S(ij) = \sum_{s \in V} \sum_{t \in V} \sigma_{st}(ij)$$

for edges

**Can be interpreted as the number of shortest paths that are lost if the vertex or edge is removed from the graph. However, ... (do you see any problem with this method?)**



**Fig. 3.9.** The figure shows that the removal of an edge can actually increase the number of shortest paths in a graph

Homework 1) check at home; 2) when will the removal of an edge lead to an increase in the edge number? 3) any solution to this problem?

# Current Flow

- Electrical network is an undirected, simple, connected graph  $G = (V, E)$
- Conductance function  $c: E \rightarrow \mathbb{R}$
- Supply function  $b: V \rightarrow \mathbb{R}$  (external electrical current enters and leaves network)
- Positive  $b =$  entering current
- Negative  $b =$  leaving current

$$\sum_{i \in V} b(i) = 0$$

- Direction of the current: each edge  $ij$  in  $E$  is oriented arbitrarily

Function  $x : \overline{E} \rightarrow \mathbb{R}$  is called current if

$$\sum_{ij \in \overline{E}} x_{ij} - \sum_{ji \in \overline{E}} x_{ji} = b(i) \text{ and } \sum_{ij \in C} x_{ij} = 0$$

for every cycle  $C \subset E$ .  undirected

A function  $p : V \rightarrow \mathbb{R}$  is a potential if  $p(i) - p(j) = x_{ij}/c_{ij}$  for all  $ij \in \overline{E}$ . As an electrical network  $N = (G, c)$  has a unique current  $x$  for any supply  $b$ , it also has a potential  $p$  that is unique up to an additive factor.

Given edge weights  $c(i)$ , we define electrical network Laplacian  $L$ .

We can find  $p$  and  $b$  by solving  $Lp=b$ .

## Current-Flow Betweenness Centrality

Unit  $s - t$ -supply  $b_{st}$  is a supply of one unit that enters the network at  $s$  and leaves at  $t$ , that is,  $b_{st}(s) = 1$ ,  $b_{st}(t) = -1$ , and  $b_{st}(i) = 0$  for all  $i \in V \setminus \{s, t\}$ .

Throughput of  $i \in V$  with respect to a unit  $s - t$ -supply  $b_{st}$  is defined as

$$\tau_{st}(i) = \frac{1}{2} \left( -|b_{st}(i)| + \sum_{ij \ni i} |x(\overline{ij})| \right)$$

$$c_{CB}(i) = \frac{1}{(n-1)(n-2)} \sum_{s,t \in V} \tau_{st}(i)$$



## Homework

Paper review 2: Newman “A measure of betweenness centrality based on random walks”

Paper review 3: Freeman “A set of measures of centrality based on betweenness”

Submit by 9/19/2019

# How to compare different centrality concepts?

## Normalization in one network

$p$ -norm of the centrality vector for concept  $\mathbf{X}$

$$\|c_{\mathbf{X}}\|_p = \begin{cases} (\sum_{i=1}^n |c_{\mathbf{X}i}|^p)^{1/p} & 1 \leq p < \infty \\ \max_i \{|c_{\mathbf{X}i}|\} & p = \infty \end{cases} \implies \frac{c_{\mathbf{X}}}{\|c_{\mathbf{X}}\|_p} \implies c_{\mathbf{X}i} \leq 1$$

separation of positive and negative values of  $c_{\mathbf{X}}$

$$c'_{\mathbf{X}} = \begin{cases} c_{\mathbf{X}i} / (\sum_{j:c_{\mathbf{X}j} > 0} |c_{\mathbf{X}j}|^p)^{1/p} & c_{\mathbf{X}i} > 0 \\ 0 & c_{\mathbf{X}i} = 0 \\ c_{\mathbf{X}i} / (\sum_{j:c_{\mathbf{X}j} < 0} |c_{\mathbf{X}j}|^p)^{1/p} & c_{\mathbf{X}i} < 0 \end{cases}$$


**Exercise (do not submit):** Is  $c'_{\mathbf{X}}$  a norm? Prove or disprove.

Freeman "Centrality in social networks: Conceptual clarification"

# Normalization for different networks

Point-centrality

$$c''_{\mathbf{X}i} = c_{\mathbf{X}i} / \left( \max_{G \in \mathcal{G}_n} \max_{i \in V(G)} c_{\mathbf{X}i} \right)$$

 set of all graphs with  $n$  vertices

## Examples

- Degree centrality = normalization by factor  $(n-1)$
- Shortest paths betweenness centrality  $c_B(i) = \sum_{s \neq i} \sum_{t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$

What is the upper bound (or normalization factor)?

**Star graph,  $c_B(i) = (n-1)(n-2)/2$**

- Closeness centrality

$\forall i \in V C_i = 1/l_i$ , where  $l_i = \frac{1}{n} \sum_{j \in V} \delta_{ij}$ ,  $\delta_{ij}$  = length of  $i - j$  shortest path

What is the upper bound (or normalization factor)? It is  $1/(n-1)$

# Summary and How Does It Work in Practice

## Categories of centrality measures

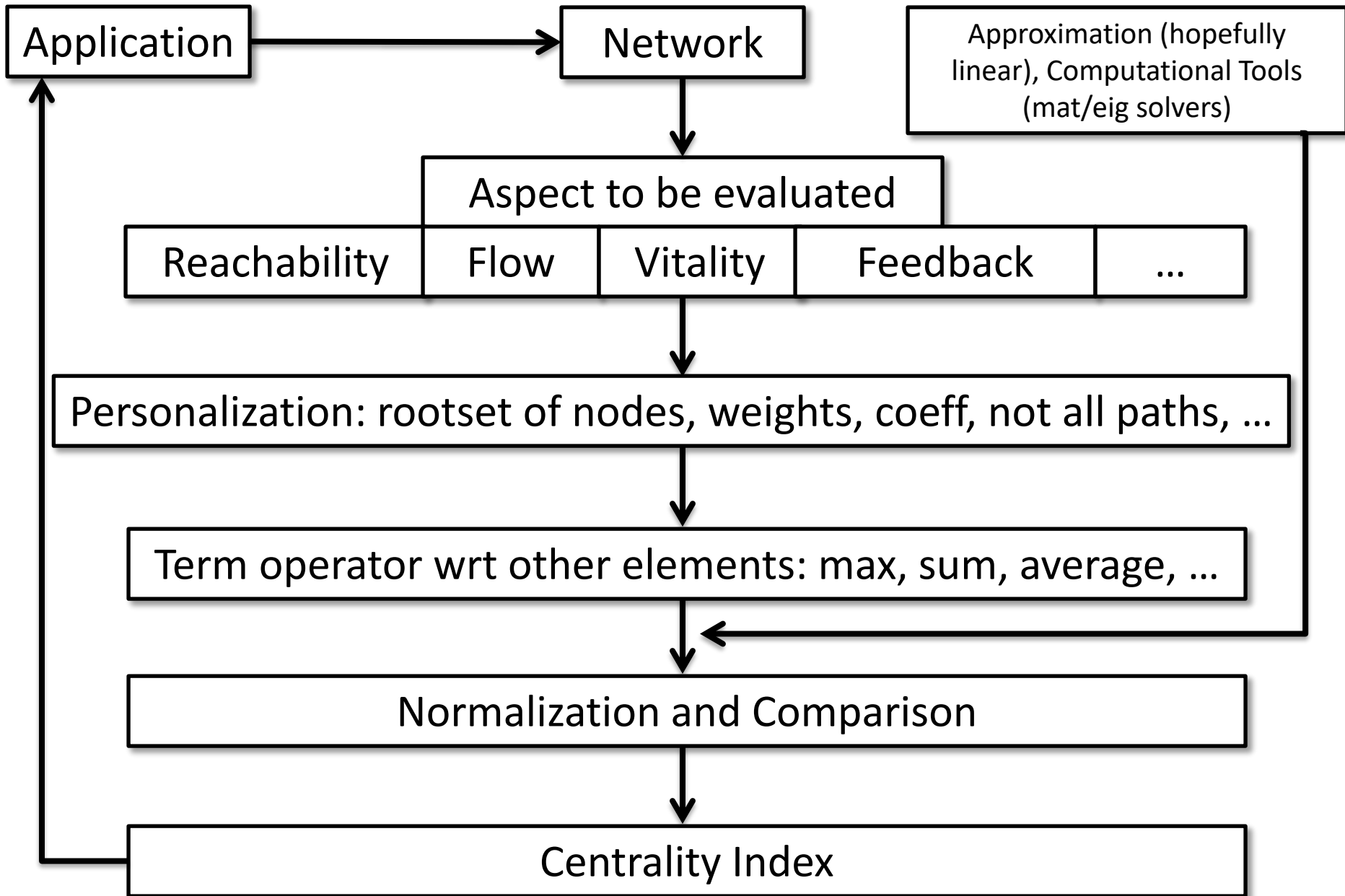
*Reachability.* A vertex is supposed to be central if it reaches many other vertices. Centrality measures of this category are the degree centrality, the centrality based on eccentricity and closeness, etc. All of these centralities rely on the distance concept between pairs of nodes.

*Amount of flow.* Based on the amount of flow  $f_{st}(i)$  from a vertex  $s$  to a vertex  $t$  that goes through a vertex or an edge  $i$ . Can be based on current flow and random walks (will see how it works in Spectral Methods). Also measures that are based on the enumeration of shortest paths, stress centrality; betweenness centralities measure the expected fraction of times a unit flow goes through the element if every vertex  $s$  sends one unit flow consecutively to every other vertex  $t$ .

*Vitality.* Based on the vitality, i.e., the centrality value of an element  $x$  is defined as the difference of a real-valued function  $f$  on  $G$  with and without the element. Recall, a general vitality measure is identified  $f(G) - f(G \setminus \{x\})$ . Such as the max-flow betweenness vitality.

*Feedback.* Centrality measures that are based on implicit definitions of a centrality given by the abstract formula  $c(i) = f(c(v_1), \dots, c(v_n))$ , where the centrality value of  $i$  depends on the centrality values of all vertices. Includes Katz and some of eigenvector-based centralities.

[BE] “Network Analysis”



## Homework for groups of 1-2 (\*) due 3/18/2021

Implement (don't use the existing implementation!) one of the following centrality indices:

- Katz
- Traversal sets (bonus +15 points if not in group of 3)
- Betweenness (approximation)
- Page rank
- HITS
- Flow network vitality
- Current flow
- Current flow betweenness (approximation)

Submit source code, documentation, and 2 examples of sparse networks of size  $\sim 10$ -20K nodes with results. You can use libraries with eigensolvers and solvers of systems of equations. Input format is always a list of (possibly equally) weighted edges.

(\*) A group of 3 students should implement either Traversal Sets or Betweenness or Current Flow or Current Flow Betweenness.

# Definitions and Axiomatization of Vertex Centrality

Sabidussi (1966) and Kishi (1980)  
(on the blackboard)