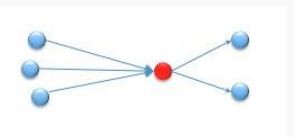
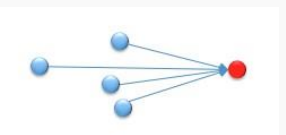
(Social) Network Analysis

Interesting phenomena in a network (examples)

- Bridges
  - are important to connect two parts of the network
  - are involved in many indirect connections



- "Prestigious" nodes
  - tend to be referred to by many other nodes



25

Social Network Metrics

- Social network graphs can be analyzed using several **metrics** including:
  - **Cohesion** of the network or sub-network

refers to the degree to which the nodes in the network are connected together

  - **Density** of the network or sub-network

how close a network is to being a complete graph

  - **Centrality** of the nodes

gives a rough indication of the social power of a node in the network

    - Degree Centrality, Betweenness, Closeness, Eigenvector (PageRank), Network centralization, …

26

Degree Centrality

- Conveys the number of links to / from other nodes in the network
- Higher degree of a node might indicate that the node is popular in the network or have more influence on the information flow
- **Degree Centrality** of nodes is derived from the immediate connections

- **In-degree**
  how many directed edges are incident on a node

$\sum_{i=1}^{n} A_{ij}$
in-degree=3

- **Out-degree**
  how many directed edges originate at a node

$\sum_{j=1}^{n} A_{ij}$
out-degree=2

- **(Total) degree**
  number of edges incident on a node (in or out)

degree=5

$$C_D(i) = deg(i)$$

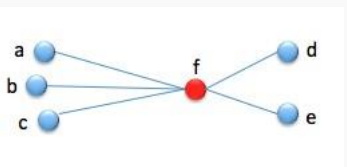
(using the total degree of a node)

27

Normalized Degree Centrality

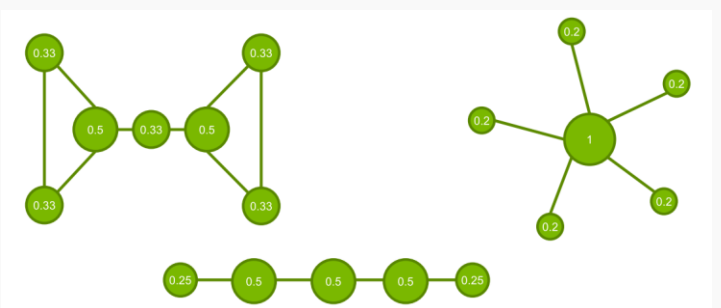
- Divide the degree of a node *deg*(*i*) by the maximum possible (i.e., **n-1** nodes)
- Range ∈ [0,1]

$$C'_D(i) = \frac{deg(i)}{n-1}$$



i	$C_D(i)$
a	1/5
b	1/5
c	1/5
d	1/5
e	1/5
f	5/5

There are also:
In-degree centrality
Out-degree centrality



28

Betweenness Centrality

- **Number of shortest paths** between each node pair in which a node is on
- Observation: nodes that bridge different groups in the network have a high betweenness
- High betweenness generally indicates a **powerful / influential** position in the network

$$C_B(i) = \sum_{j<k, j\neq i, k\neq i} \frac{P_{jk}(i)}{P_{jk}}, \text{ where}$$

  - $P_{jk}(i)$ is the number of shortest paths between j and k that go through i ($i \neq j, i \neq k$)
  - P_{jk} is the number of shortest paths between j and k

  - Range $[0, (n-1) \times (n-2)]$, with n as the number of nodes (for **non-directed** networks)
  - **$(n-1)(n-2)/2$** is the total number of pairs of nodes not including i (for **directed** networks).

29

Normalized Betweenness Centrality

- **Betweenness centrality** of a node

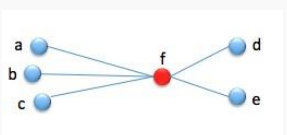
$$C_B(i) = \sum_{j<k, j\neq i, k\neq i} \frac{P_{jk}(i)}{P_{jk}}$$

- **Normalized Betweenness centrality** of a node

$$C'_B(i) = \frac{C_B(i)}{\frac{(n-1)(n-2)}{2}}$$

Number of pairs of vertices excluding the vertex itself **not considering directions**

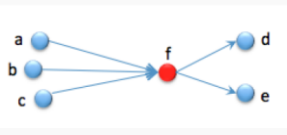
non-directed graphs



i	$C'_B(i)$
a	0/10
b	0/10
c	0/10
d	0/10
e	0/10
f	10/10

$$C'_B(i) = \frac{C_B(i)}{(n-1)(n-2)}$$

Number of pairs of vertices excluding the vertex itself **considering directions**

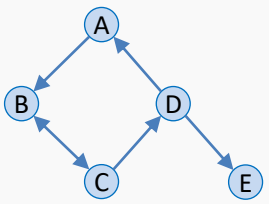


i	$C'_B(i)$
a	0/20
b	0/20
c	0/20
d	0/20
e	0/20
f	6/20

30

Closeness Centrality

- Inverse of the **mean shortest path** between a node and all other nodes in the network reachable from it
- Reflects the ability of a node in accessing information through the network
- High closeness generally indicates better visibility of what's going on in the network



	A	B	C	D	E
A		1	2	3	4
B	3		1	2	3
C	2	1		1	2
D	1	2	3		1
E	-	-	-	-	

$$C_C(i) = \frac{1}{\sum_{i \neq j} d(i,j)}$$

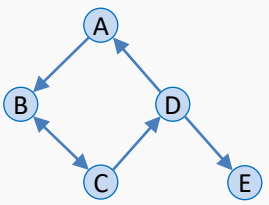
Examples:

$C_c(A) = \frac{1}{1+2+3+4} = 1/10 = 0.1$

$C_c(D) = \frac{1}{1+2+3+1} = 1/7 = 0.14$

31

Normalized Closeness Centrality



	A	B	C	D	E
A		1	2	3	4
B	3		1	2	3
C	2	1		1	2
D	1	2	3		1
E	-	-	-	-	

$$C'_c(i) = C_c(i) \times (n-1)$$

Last examples:

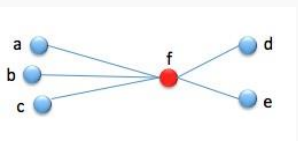
$C_c(A) = \frac{1}{1+2+3+4} = 1/10 = 0.1$

$C_c(D) = \frac{1}{1+2+3+1} = 1/7 = 0.14$

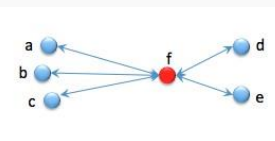
Normalized:

$C'_c(A) = \frac{(5-1)}{10} = 4/10 = 0.4$

$C'_c(D) = \frac{(5-1)}{7} = 4/7 = 0.6$



i	$C'_C(i)$
a	0.55
b	0.55
c	0.55
d	0.55
e	0.55
f	1.0



i	$C'_C(i)$
a	0.55
b	0.55
c	0.55
d	0.55
e	0.55
f	1.0

32

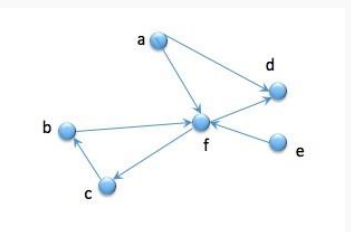
Network Analysis: centrality measures - Wrap-up

- **Degree centrality**:
  - measures the **involvement** of a node in the network;
  - it is an effective measure of the **importance, influence,** and **popularity** of a node in the network.
- **Betweenness centrality**:
  - measures if a node has a **critical role** in the network, i.e., if it acts as a connection between different regions of the network;
  - identifies ***gatekeepers or bridges***, nodes that tend to control the flow of information between tightly knit groups.
- **Closeness centrality**:
  - measures the **overall position** of a node in the network;
  - a measure of **reachability and distance** that measures how fast a given node can reach every other nodes in the network.

33

Exercise

1. Consider the following data on **friendship requests** in a social network:



(a) Who would you pick for collecting information?
(b) Who would you pick as a marketing mate?
(c) Who would you pick for distribution of goods?

34

Network Analysis: "prestige" measures

Prestige is a metric similar to Centrality, but mainly-used for in-degree cases.

However, in-degree centrality may not provide a complete picture of a node's prestige. Other metrics can be useful:

- Betweenness centrality: (not directly related, however) a node with high betweenness centrality serves as a critical intermediary, potentially enhancing its influence
- Closeness centrality: (NDRH) nodes with high closeness centrality can quickly interact with other nodes making them more influential
- Eigenvector centrality: takes into account the quality/importance of the connections

Let's simplify this a bit….

35

Network Analysis: "prestige" measures

Prestige is a metric similar to Centrality, but mainly-used for in-degree cases.

- Assume that:
  - node A is referred by *n* ordinary nodes;
  - node B is referred by *n* nodes, *k* of which are "prestigious".
- Which node has higher "prestige"?
- We must take the prestige of the pointing nodes into account.
- **HITS** (1998) and **PageRank** (1998) do just that!
  - Find the most **valuable**, **authoritative** or **influential** node (e.g. web page).

36

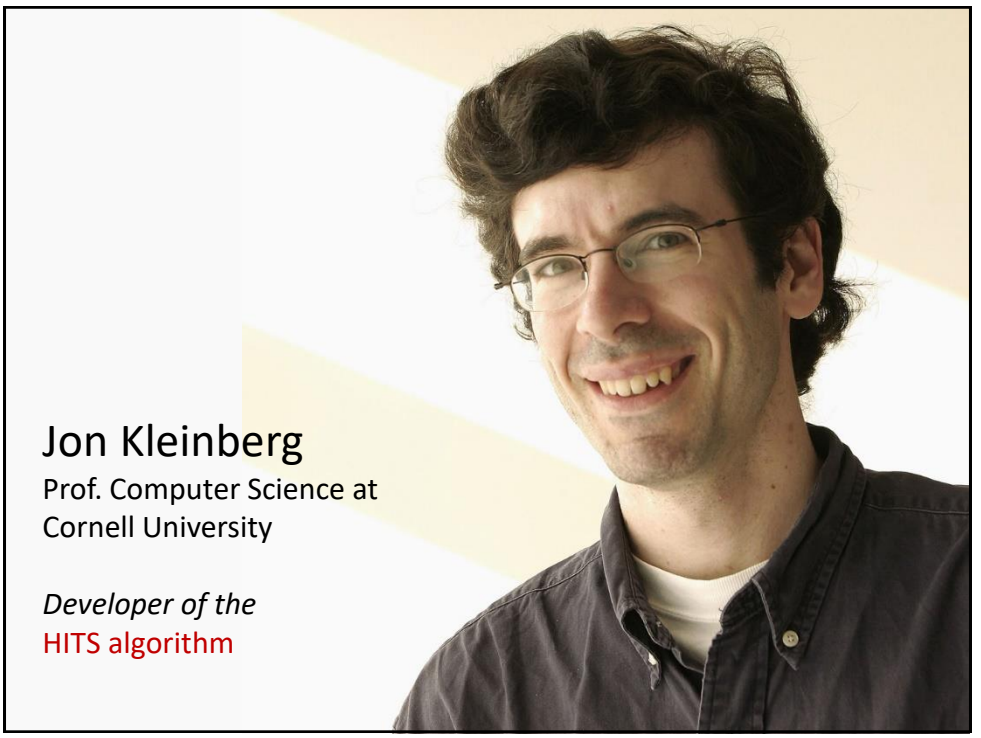
Important Pages Discovery

37

Using Web Structure for Information Retrieval

- **Search**
  - Search a page about topic X
  - Then, each resulting page Y is relevant according to:
    - similarity between the contents of X and Y
- **Link analysis**
  - Each page Y is relevant according to
    - number of references to page Y
    - content of pages which refer to Y
- Pages linked to pages with interesting content are also potentialy interesting.

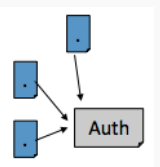
38



39

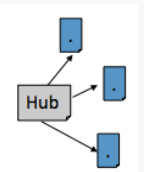
Authorities and Hubs

- Discovery of two kinds of pages
  - **Authorities**: pages referred to by many in a specific topic.



Many in-links

  - **Hubs**: pages that refer to many others.



Many out-links

40

HITS (Hypertext Induced Topic Search)

- In a first stage, HITS uses text similarity and then uses link structure.
- **Relevance of an Authority**
  - If a page is referred to by many others, then it must be relevant
    - this enables a more robust search to variation in terms
      - example: "data mining" and "machine learning"
- **Quality of a Hub**
  - If a hub refers to many important authorities, then it is a good hub

- The relevance of an Authority and the quality of a Hub are interdependent
- The definition is circular: *good authorities are pointed by good hubs and good hubs point to good authorities.*

41

HITS: The idea

1. Given a query (or topic) Q, collect a set of *seed* pages $S = \{s_1, s_2, \ldots, s_n\}$ (this is the root set)
2. S is then expanded to $T = S \cup \{d \mid s \rightarrow d \vee d \rightarrow s, s \in S\}$
3. Initialize each page $r \in T$ with
   authority weight $a(r) = 1$ and hub weight $h(r) = 1$
4. For each page, update the values of a and h

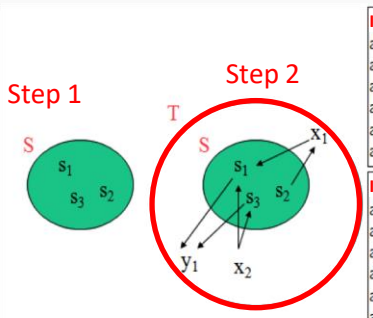
$$a(r) = \sum_{d \rightarrow r} h(d) \qquad h(r) = \sum_{r \rightarrow d} a(d)$$

5. Normalize a and h
6. Repeat step 4, until convergence (typically <10 iterations)
7. Select the top ranked pages with **highest a and h**.

- The pages with the **highest Authority scores** are the most relevant pages for the **query topic**.
- The pages with the **highest Hub scores** are the most useful pages for **finding additional information on the query topic**.

42

HITS: Example (Iterative solution)



Step 3

x1→s1	a(s1)=1, h(s1)=1
x2→s1	a(s2)=1, h(s2)=1
s1→y1	a(s3)=1, h(s3)=1
s2→x1	a(x1)=1, h(x1)=1
x2→s3	a(x2)=1, h(x2)=1
s3→y1	a(y1)=1, h(y1)=1

Iteration 1	Iteration 1 (norm.)
a1(s1)=2, h1(s1)=1	a2(s1)=1, h2(s1)=0,5
a1(s2)=0, h1(s2)=1	a2(s2)=0, h2(s2)=0,5
a1(s3)=1, h1(s3)=1	a2(s3)=0,5, h2(s3)=0,5
a1(x1)=1, h1(x1)=1	a2(x1)=0,5, h2(x1)=0,5
a1(x2)=0, h1(x2)=2	a2(x2)=0, h2(x2)=1
a1(y1)=2, h1(y1)=0	a2(y1)=1, h2(y1)=0
Iteration 2	**Iteration 2 (norm.)**
a2(s1)=1,5 h2(s1)=1	a2(s1)=1 h2(s1)=0,66
a2(s2)=0 h2(s2)=0,5	a2(s2)=0 h2(s2)=0,33
a2(s3)=1 h2(s3)=1	a2(s3)=0,66 h2(s3)=0,66
a2(x1)=0,5, h2(x1)=1	a2(x1)=0,33 h2(x1)=0,66
a2(x2)=0, h2(x2)=1,5	a2(x2)=0, h2(x2)=1
a2(y1)=1 h2(y1)=0	a2(y1)=0,66 h2(y1)=0
Iteration 3	**Iteration 3 (norm.)**
a2(s1)=1,66 h2(s1)=0,66	a2(s1)=1 h2(s1)=0,4
a2(s2)=0 h2(s2)=0,33	a2(s2)=0 h2(s2)=0,2
a2(s3)=1 h2(s3)=0,66	a2(s3)=0,6 h2(s3)=0,4
a2(x1)=0,33 h2(x1)=1	a2(x1)=0,2 h2(x1)=0,6
a2(x2)=0 h2(x2)=1,66	a2(x2)=0 h2(x2)=1
a2(y1)=1,33 h2(y1)=0	a2(y1)=0,8 h2(y1)=0
Top3 Authority	**s1** (1), **y1** (0,8), **s3** (0,6)
Top3 Hub	**x2** (1) **x1**(0,6), **s1** (0,4)

43

HITS (mathematical perspective)

- We can use the **adjacency matrix** A, i.e.

$$A_{ij} = \begin{cases} 1 & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

A	S1	S2	S3	X1	X2	Y1
S1	0	0	0	0	0	1
S2	0	0	0	1	0	0
S3	0	0	0	0	0	1
X1	1	0	0	0	0	0
X2	1	0	1	0	0	0
Y1	0	0	0	0	0	0

- to define

$$a^{(k)} = A^T h^{(k-1)} \qquad h^{(k)} = A a^{(k-1)}$$

where a is the column vector with all **authority** scores

$$a = (a(1), a(2), \ldots, a(n))^T$$

and h is the column vector with all **hub** scores

$$h = (h(1), h(2), \ldots, h(n))^T$$

and k is the iteration number.

44

HITS (using iterative power method)

The computation of authority scores and hub scores uses the **iterative power method**:

1. Initialize $a^0 = h^0 = (1, 1, \ldots, 1)$
2. Until convergence, do:
    1. $a^{(k)} = A^T h^{(k-1)}$
    2. $h^{(k)} = A a^{(k-1)}$
    3. k=k+1
    4. Normalize $a^{(k)}$ and $h^{(k)}$

- a, the largest eigenvector[1] of $A^T A$
- h, the largest eigenvector[1] of AA^T

Note that:

$a^{(k)} = A^T h^{(k-1)}$

$h^{(k)} = A a^{(k-1)}$

can be simplified to:

$a^{(k)} = A^T A a^{(k-1)}$

$h^{(k)} = A A^T h^{(k-1)}$

[1] x is eigenvector of M, if $Mx = kx$ where k is a scalar

45

HITS: Some comments

- HITS ranks pages according to the query topic
    - search starts by content relevance (root set) and then the content is ignored, i.e. only the links are used.
- HITS finds the principal eigenvectors.
    - top ranked authorities and hubs represent **major communities**
    - In some cases, smaller communities can also be relevant.
      example: "jaguar" can be grouped in 2 clusters (animal and car)
- HITS scores each page with two attributes (a,h) and is dependent on the query.
- Overall, while the original formulation of HITS did focus on a two-level network structure, the algorithm can be applied to networks with any number of levels.

46

HITS: Some comments (cont.)

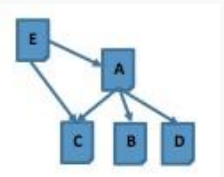
- It is easy to influence HITS by adding out-links from a page to point to many good authorities, increasing the hub score of the page.
- Topic drift is another problem: the expansion of the root set may include pages which have nothing to do with query topic.
- Small changes in the web graph topology can significantly change the results.
- Query time evaluation is a major drawback.

Note: a(i) and h(i) are sometimes written as x_i and y_i , respectively.

47

Exercise

2 Consider the following graph of web pages:



(a) Determine the most interesting hub.
(b) Determine the most important authority.
(c) Suppose we are looking for information about a car model X and page A contains that model, how would that change your previous results?

48