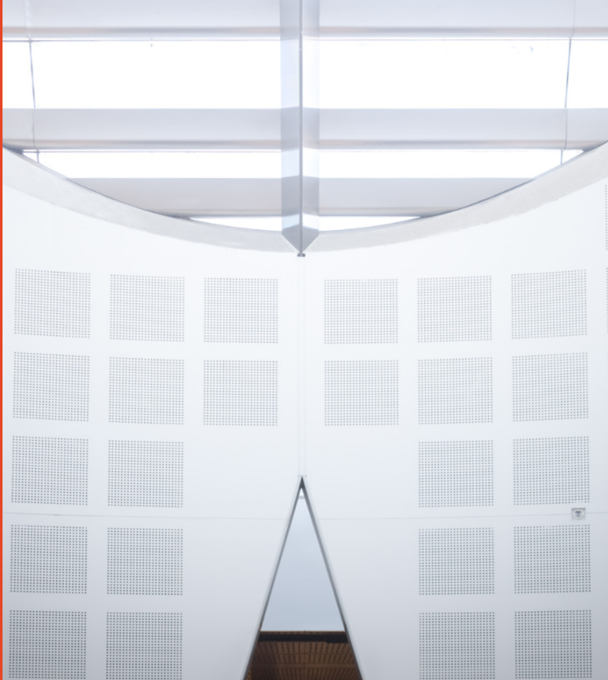
COMP5313/COMP4313 - Large Scale Networks

Week 5: The Structure of the Web, Hubs and Authorities

Lijun Chang

March 27, 2025

THE UNIVERSITY OF SYDNEY



Introduction

- We already looked at social networks
  - The basic units being connected are people or other social entities, like firms or organizations
  - The links connecting them generally correspond to opportunities for some kind of social or economic interaction
  - e.g., Facebook network, twitter network, Livejournal network

- We will now consider information networks
  - The basic units being connected are pieces of information
  - Links join pieces of information that are related to each other
  - e.g., the World Wide Web, citation networks, knowledge graphs

Outline

Information Networks

Bow-tie Structure of the Web

Link Analysis: Hubs and Authorities

Applications beyond the Web

Spectral Analysis of Hubs and Authorities

Advanced Material (Optional)

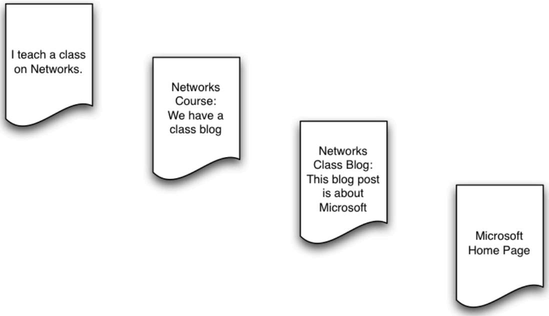
The World Wide Web

- The World Wide Web is probably the most prominent information network
- The Web is an application developed to let people share information over the Internet
- It was created by Tim Berners-Lee during the period of 1989-1991 [1]
- It features two components:
  - It makes a document available over the Internet through a Web page stored on a public folder of a computer
  - It provides a way for others to easily access Web pages through a browser

[1] T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret. The world-wide web. Commun. ACM, vol. 37, pp. 76–82, 1994.

The World Wide Web

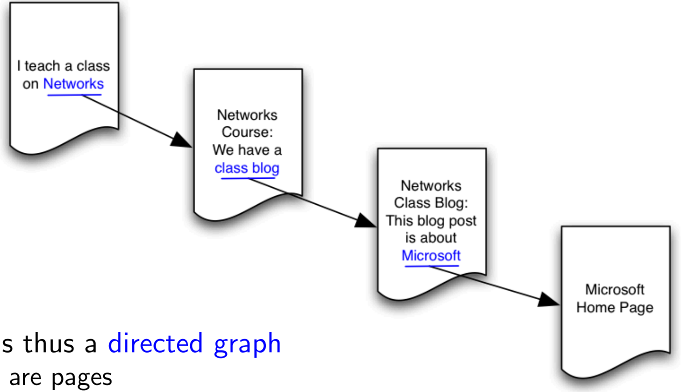
- A set of four Web pages
  - The home page of an instructor who teaches a class of network, the homepage of a network class she teaches, the blog for the class, with a post about Microsoft.



  - These pages are part of one system (the Web) but may be located on four different computers belonging to different institutions.

The World Wide Web

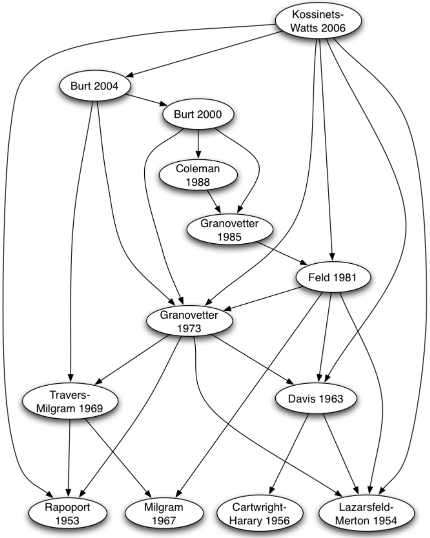
- The Web uses the network metaphor
  - Each page is a hypertext that can embed virtual links in any portion of the document
  - This virtual link allows a reader to move from one Web page to another



- The Web is thus a directed graph
  - Nodes are pages
  - The directed edges are the links from one page to another

Citation

- A precursor of hypertext is citation
  - For authors to credit the source of an idea
- The network of citations among a set of research papers forms a directed graph that, like the Web, is a kind of information network.
- In contrast to the Web, however, the passage of time is much more evident in citation networks since their links tend to point strictly backward in time.



Outline

Information Networks

Bow-tie Structure of the Web

Link Analysis: Hubs and Authorities

Applications beyond the Web

Spectral Analysis of Hubs and Authorities

Advanced Material (Optional)

Web as a Directed Graph

- Viewing social networks in terms of their graph structures have provided significant insights

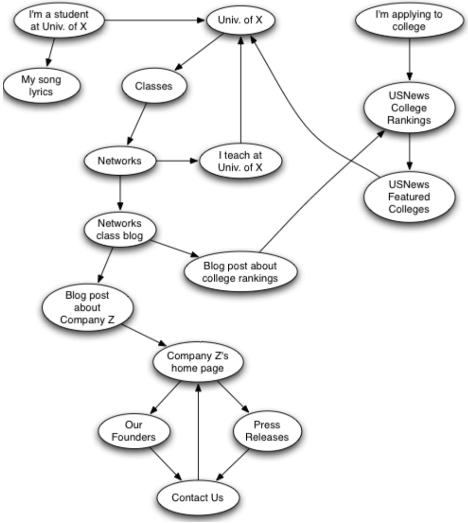
- The same is true for information networks such as the Web
  - It allows us to better understand the logical relationships expressed by its links
  - It helps to identify important pages as a step in organizing the results of Web searches

- The Web graph is directed
  - The edges point from one node to another, and are not symmetrical
  - Page A pointing to page B does not necessarily indicate page B pointing back to page A

Path in a Directed Graph

- A directed graph formed by the links among a small set of Web pages

- A path from A to B in a directed graph is a sequence of nodes, beginning with A and ending with B with the property that each consecutive pair of nodes in the sequence is connected by an edge pointing in the forward direction.



Strongly Connected Components

- A directed graph is strongly connected if there is a path from every node to every other node

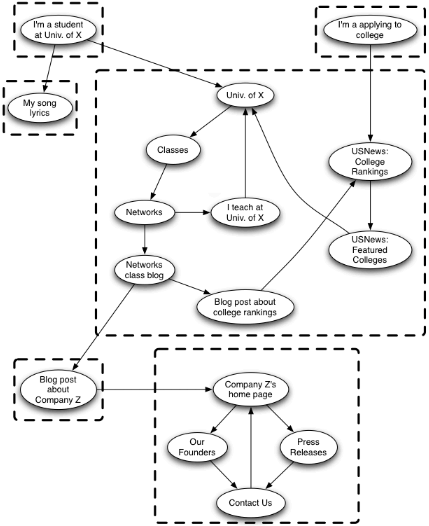
Strongly Connected Component (SCC)

A strongly connected component (SCC) in a directed graph is a subset of the nodes such that:

1. Every node in the subset has a path to every other; and
2. The subset is not part of some larger set with the property that every node can reach every other.

Strongly Connected Components

- A directed graph with its strongly connected components identified



The Bow-tie Structure of the Web

- In 1999, Broder et al. set out to build a global map of the Web [2]
  - They used the index of pages and links of AltaVista, one of the largest commercial search engine at that time
- This study was replicated on
  - the larger (early) index of Google's search engine and
  - large research collection of web pages

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener. Graph structure in the web. Computer networks, vol. 33, no. 1–6, pp. 309–320, 2000.

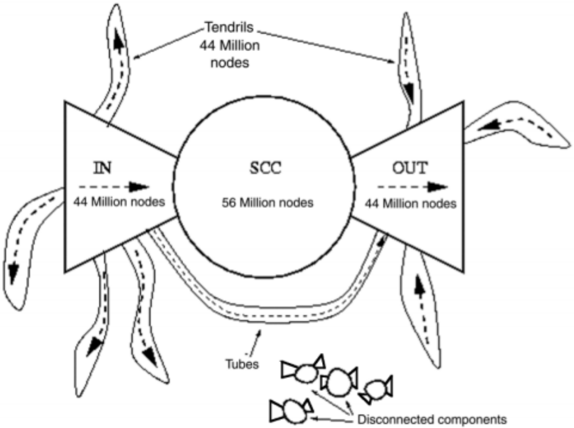
The Bow-tie Structure of the Web

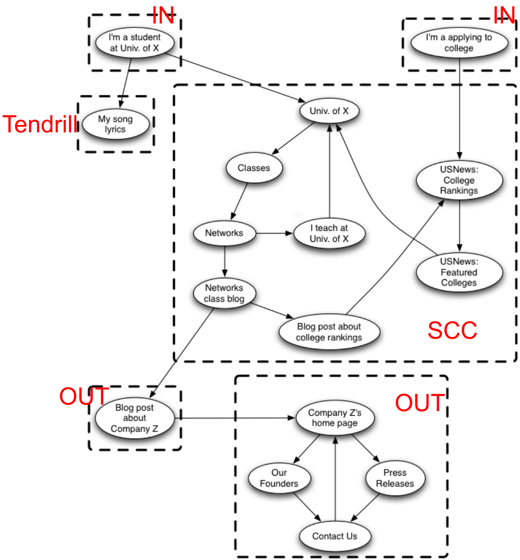
Their findings include:

- The Web contains a giant strongly connected component (SCC):
- **IN:** Some nodes can reach the giant SCC but cannot be reached from it, these nodes are upstream of the giant SCC
- **OUT:** Some nodes can be reached from the giant SCC but cannot reach it, these nodes are downstream of the giant SCC
- There are nodes that can neither reach the giant SCC nor be reached from it
  - Tendrils are nodes that can be reached from IN but cannot reach the giant SCC and the ones that can reach OUT but cannot be reached from the giant SCC
  - Disconnected are nodes that have no path to the giant SCC (even if ignoring directions)

The Bow-tie Structure of the Web

- A schematic picture of the bow-tie structure of the Web. Although the numbers are now outdated, the structure has persisted.





IN
I'm a student at Univ. of X
IN
I'm a applying to college
Tendrill
My song lyrics
Univ. of X
Classes
USNews: College Rankings
Networks
I teach at Univ. of X
USNews: Featured Colleges
Networks class blog
Blog post about college rankings
SCC
OUT
Blog post about Company Z
OUT
Company Z's home page
Our Founders
Press Releases
Contact Us

Outline

Information Networks

Bow-tie Structure of the Web

Link Analysis: Hubs and Authorities

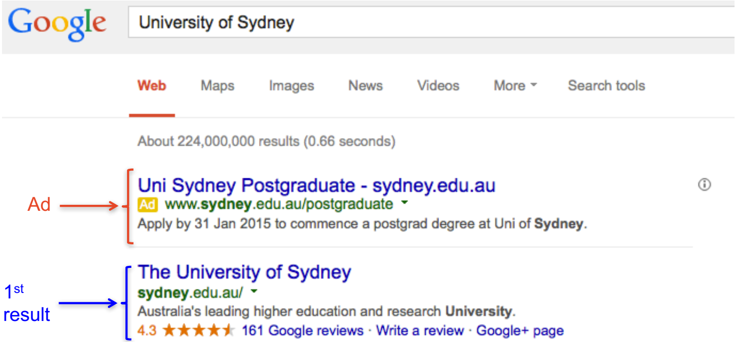
Applications beyond the Web

Spectral Analysis of Hubs and Authorities

Advanced Material (Optional)

Searching the Web: The Problem of Ranking

- Type "University of Sydney" in Google's search engine



- How does Google's search engine know which page to show first?
  - Search engines only exploit information from the Web (no external info)

$\Longrightarrow$ There should be enough information intrinsic to the Web to rank results

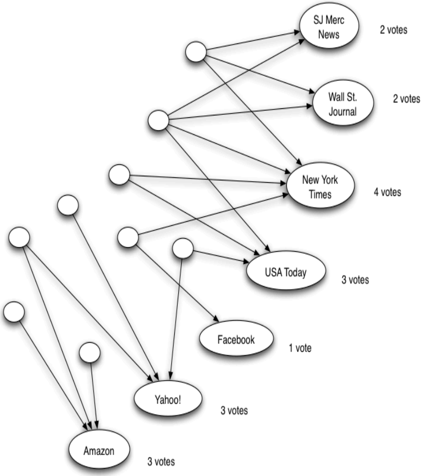
Voting by In-links

- Perspective
  - All pages with "University of Sydney" contain different numbers of occurrences
  - But all these webpages likely link to `sydney.edu.au`
- Links are essential
  - Some links may be off-topic, may be negative rather than positive ...
  - But overall, many incoming links means hopefully collective endorsement
- Let's list all relevant pages with the term "University of Sydney"
  - Consider links as votes from one webpage to another
  - What page receives the largest number of votes from other pages?
  - Ranking pages by decreasing number of votes works reasonably well

Voting by In-links

- Voting is not enough
- Type "newspapers", you may get high scores for prominent newspapers, along with irrelevant highly ranked pages
  - Unlabelled pages represent a sample of pages relevant to query newspapers
  - The most voted pages are
    - two newspapers (NYT, USA today)
    - two irrelevant results (Yahoo!, Amazon)

$\Longrightarrow$ Vote number is a too simple measure

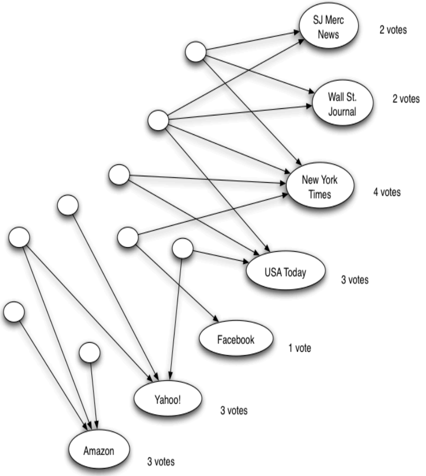


A List-finding Technique

- What other information can complement vote measure?
- What are the (list) pages that compile lists of resources relevant to the topic?
  - Such lists exist for most broad enough queries like “newspapers”
  - They would correspond to lists of links to online newspapers
  - Let’s try to find good list pages for the query “newspapers”

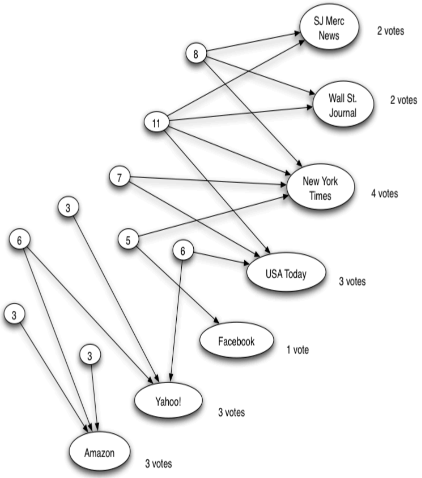
A List-finding Technique

- Let's consider the figure again
  - Few list pages voted for many of the highly voted pages
  - List pages have some sense of where the good answers are
  - Page's value as a list is the sum of the votes received by all pages for which it voted



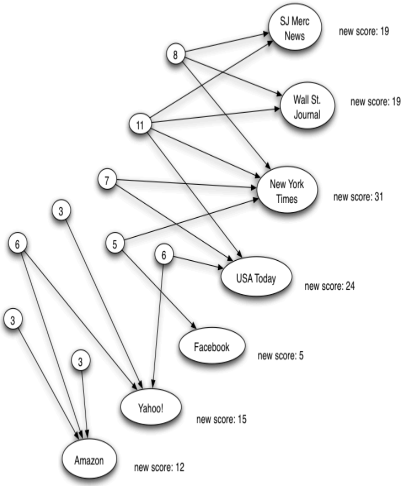
A List-finding Technique

- Finding good lists for the query "newspapers": each page's value as a list is written as a number inside it
- If we believe that pages scoring well as lists have a better sense for where the good results are, we should weight their votes more heavily
- Similarly, people recommending lots of good restaurants may act as high-value lists so that you end up giving them more value



The Principle of Repeated Improvement

- Re-weight votes for the query "newspapers": each of the labeled page's new score is equal to the sum of the values of all lists that point to it
- Why stop here?
  Can we refine the scores obtained on the left-hand side as well?
- This process can go back and forth forever *(repeated improvement)*



Hubs and Authorities

- This process suggests a ranking procedure that we can try to make precise, as follows
  - We call authorities the page with high score for the query
  - We call hubs the high-value list for the query
- For each page p, we assign pairs: $hub(p)$ and $auth(p)$
  - Each page starts with $(1, 1)$

- **Voting:** Use the quality of hubs to refine our estimate for the quality of authorities
  - **Authority Update Rule:** For each page p, update $auth(p)$ to be the sum of the hub scores of all pages that point to it.
- **List-finding:** Use the quality of the authorities to refine our estimate of the quality of the hubs.
  - **Hub Update Rule:** For each page p, update $hub(p)$ to be the sum of the authority scores of all pages that it points to.

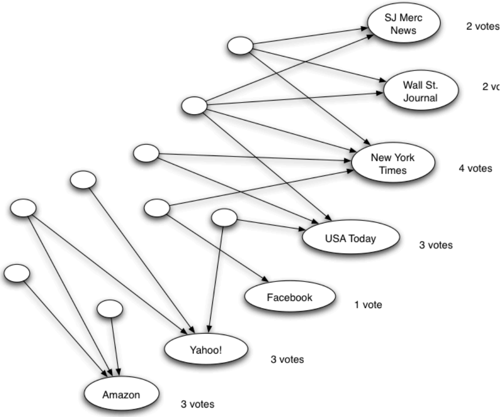
Hubs and Authorities

Algorithm

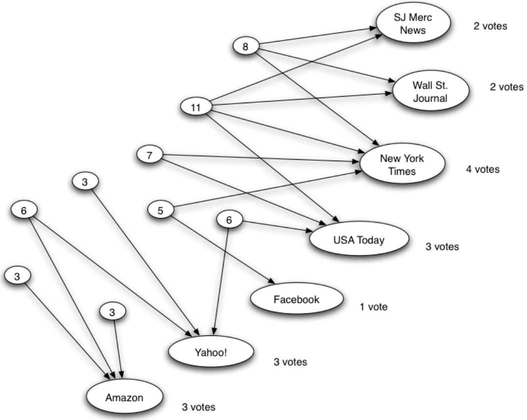
- We start with all hub scores and all authority scores equal to 1
- We choose a number of steps, k
- We then perform a sequence of k hub-authority updates

  Each update works as follows:
  - First apply the Authority Update Rule to the current set of scores
  - Then apply the Hub Update Rule to the resulting set of the scores
- At the end, the hub and authority scores may involve numbers that are very large so normalize them
  - divide each authority score by the sum of all authority scores
  - divide each hub score by the sum of all hub scores

Hubs and Authorities

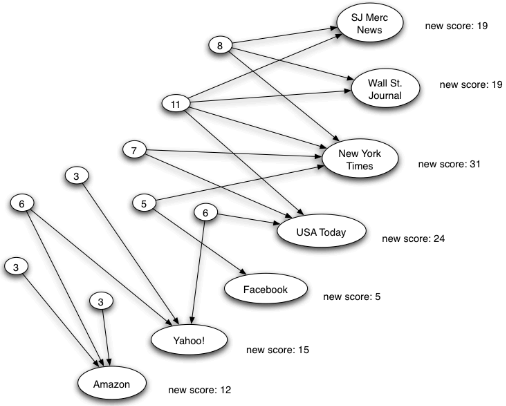


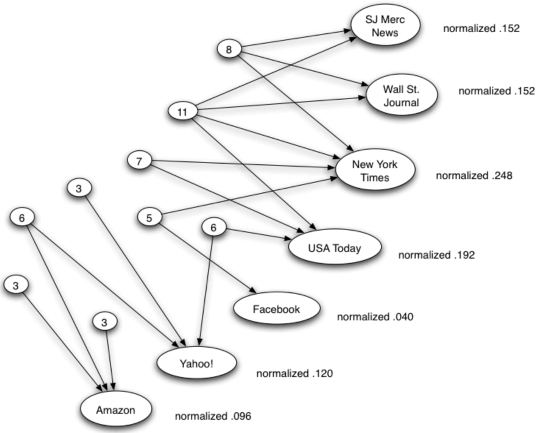
After 1 application of the Authority Update Rule (assuming that $hub(p) = 1$ for every p)



Then after 1 application of the Hub Update Rule

Hubs and Authorities





After second application of the Authority Update Rule

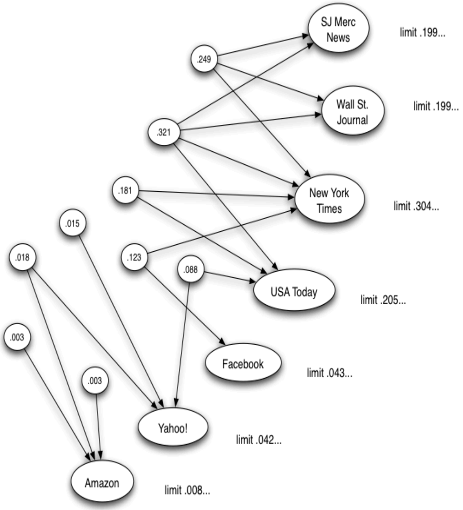
After normalization (sum of authorities was 125)

Hubs and Authorities

- What happens if we do this for larger and larger values of k?
  - Normalized values converge to limits as k goes to infinity
  - The result stabilizes as the improvement leads to smaller and smaller changes
- Ultimately, we reach an equilibrium
  - Your authority score is proportional to the hub scores of the pages that point to you
  - Your hub score is proportional to the authority scores of the pages you point to
- The same limits are reached whatever positive initial values we choose for hubs and authorities
  - The limiting values are properties of the link structure (not initial values)

Hubs and Authorities

- Limiting hub and authority values for the query "newspapers"



After normalization

Outline

Information Networks

Bow-tie Structure of the Web

Link Analysis: Hubs and Authorities

Applications beyond the Web

Spectral Analysis of Hubs and Authorities

Advanced Material (Optional)

Application of Link Analysis

- Link analysis techniques have diverse applications in any domain where information is connected by a network structure
  - Citation analysis (impact factor)
  - U.S. Supreme Court citations

US Supreme Court Citations

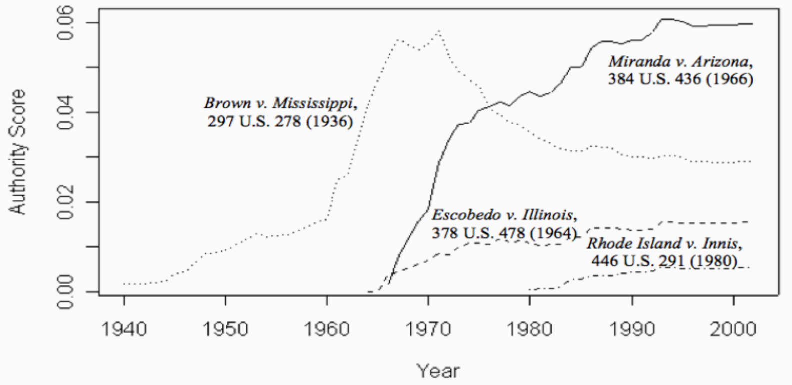
- Study of the network of citations among legal decisions by U.S. courts
- Citations are crucial in legal writing:
  - To ground a decision in precedent
  - To explain the relation of a new decision to what has come before

- Link analysis of citations helps identifying cases that play especially important roles in the overall citation structure

- Hub and authority measures used on all Supreme Court decisions (over 2 centuries)
  - Revealed cases that acquired significant authority according to these measures shortly after they appeared
  - But which took much longer to get recognition from the legal community
  - Showed how authority can change over long time periods

US Supreme Court Citations

- Rising and falling of some key 5th Amendment cases (20th century)
  - 1936 Brown vs. Mississippi about confessions obtained under torture
  - 1966 Miranda vs. Arizona: the need for citations to the former quickly declined



Outline

Information Networks

Bow-tie Structure of the Web

Link Analysis: Hubs and Authorities

Applications beyond the Web

Spectral Analysis of Hubs and Authorities

Advanced Material (Optional)

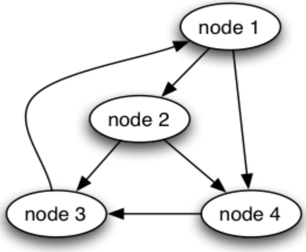
Spectral Analysis

- Spectral analysis is the use of eigenvalues and eigenvectors to study the structure of networks

- The limiting values of hub and authority values can be interpreted as coordinates in the eigenvectors of certain matrices derived from the network

Adjacency Matrix

- Set of n pages represented as nodes labeled $1, 2, 3, \ldots, n$
- Links are encoded in an adjacency $n \times n$ matrix $\mathbf{A}$
  - $A_{i,j}$ (i^{th} row and j^{th} column of $\mathbf{A}$) $= 1$ if there is a link from node i to j
  - $A_{i,j} = 0$ otherwise
- Example: the directed hyperlinks among Web pages represented as an $n \times n$ adjacency matrix $\mathbf{A}$



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

Hubs and Authorities Rules

Let's consider the hub and authority update rules in terms of matrix-vector multiplication

- For every node i,
  - its hub score is denoted h_i
  - its authority score is denoted a_i
- Hub vector is denoted $\mathbf{h}$, and authority vector is $\mathbf{a}$
  - Here we assume column vectors.
- **Hub Update Rule (formalized with matrix notation):**

$$h_i = A_{i,1} \times a_1 + A_{i,2} \times a_2 + A_{i,3} \times a_3 + \cdots + A_{i,n} \times a_n \tag{1}$$

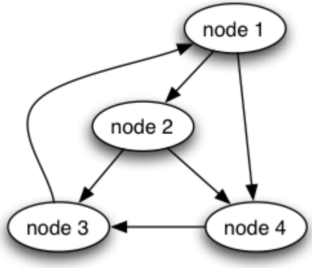
  - The values of $A_{i,j}$ as multipliers capture precisely the authority values to sum
  - Equation (1) is the definition of matrix-vector multiplication, hence we can write:

$$\mathbf{h} = \mathbf{A}\mathbf{a}$$

Hubs and Authorities Rules

Example

- The matrix representation allows to represent the Hub Update Rule as a matrix-vector multiplication
- The multiplication by a vector of authority scores $[2, 6, 4, 3]$ produces a new vector of hub scores $[9, 7, 2, 4]$



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 2 \\ 4 \end{bmatrix}$$

Hubs and Authorities Rules

- The Authority Update Rule is analogous to the Hub Update Rule, except that scores flow in the other direction across the edges
  - a_i is updated to be the sum of h_j over all nodes j that have an edge to i
- **Authority Update Rule (formalized with matrix notation):**

$$a_i = h_1 \times A_{1,i} + h_2 \times A_{2,i} + h_3 \times A_{3,i} + \cdots + h_n \times A_{n,i} \tag{2}$$

  - The roles of columns and rows are interchanged, so we use the transpose of matrix $\mathbf{A}$, denoted $\mathbf{A}^\top$, defined by the property that (i,j) entry of $\mathbf{A}^\top$ is the (j,i) entry of $\mathbf{A}$ (i.e., $(A^\top)_{i,j} = A_{j,i}$).

$$\mathbf{a} = \mathbf{A}^\top \mathbf{h} \quad \text{equivalently} \quad \mathbf{a}^\top = \mathbf{h}^\top \mathbf{A}$$

Hubs and Authorities Rules

Let's perform the k-step hub-authority computation for large values of k

- Let $\mathbf{a}^{(0)}$ and $\mathbf{h}^{(0)}$ be the vectors whose coordinates are all 1
- Let $\mathbf{a}^{(k)}$ and $\mathbf{h}^{(k)}$ denote the vectors of authority and hubs after k applications of Authority-and-then-Hub Update Rules in order
- Following previous formula we find that:

$$\mathbf{a}^{(1)} = \mathbf{A}^\top \mathbf{h}^{(0)}$$

and

$$\mathbf{h}^{(1)} = \mathbf{A}\mathbf{a}^{(1)} = \mathbf{A}\mathbf{A}^\top \mathbf{h}^{(0)}$$

That's the result of the one-step hub-authority computation.

Hubs and Authorities Rules

- In the 2nd step, we therefore get

$$\mathbf{a}^{(2)} = \mathbf{A}^\top \mathbf{h}^{(1)} = \mathbf{A}^\top \mathbf{A} \mathbf{A}^\top \mathbf{h}^{(0)}$$

  and

$$\mathbf{h}^{(2)} = \mathbf{A} \mathbf{a}^{(2)} = \mathbf{A} \mathbf{A}^\top \mathbf{A} \mathbf{A}^\top \mathbf{h}^{(0)} = (\mathbf{A} \mathbf{A}^\top)^2 \mathbf{h}^{(0)}$$

- In the 3rd step, we get

$$\mathbf{a}^{(3)} = \mathbf{A}^\top \mathbf{h}^{(2)} = \mathbf{A}^\top \mathbf{A} \mathbf{A}^\top \mathbf{A} \mathbf{A}^\top \mathbf{h}^{(0)} = (\mathbf{A}^\top \mathbf{A})^2 \mathbf{A}^\top \mathbf{h}^{(0)}$$

  and

$$\mathbf{h}^{(3)} = \mathbf{A} \mathbf{a}^{(3)} = \mathbf{A} \mathbf{A}^\top \mathbf{A} \mathbf{A}^\top \mathbf{A} \mathbf{A}^\top \mathbf{h}^{(0)} = (\mathbf{A} \mathbf{A}^\top)^3 \mathbf{h}^{(0)}$$

- What do we observe?

Hubs and Authorities Rules

- **Conclusion:** $\mathbf{a}^{(k)}$ and $\mathbf{h}^{(k)}$ are products of the terms $\mathbf{A}$ and $\mathbf{A}^\top$ in alternating order, where $\mathbf{a}^{(k)}$ begins with $\mathbf{A}^\top$ and the expression for $\mathbf{h}^{(k)}$ begins with $\mathbf{A}$.

- We can write:

$$\mathbf{a}^{(k)} = (\mathbf{A}^\top\mathbf{A})^{k-1}\mathbf{A}^\top\mathbf{h}^{(0)}$$

and

$$\mathbf{h}^{(k)} = (\mathbf{A}\mathbf{A}^\top)^{k}\mathbf{h}^{(0)}$$

- The authority and hub vectors are the results of multiplying an initial vector by larger and larger powers of $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$, respectively.

Outline

Information Networks

Bow-tie Structure of the Web

Link Analysis: Hubs and Authorities

Applications beyond the Web

Spectral Analysis of Hubs and Authorities

Advanced Material (Optional)

Eigenvectors and Convergence

- The magnitude of hubs and authorities increase at each step
- They only converge when we take normalization into account
- It is the direction of hubs and authorities that converges

- To show convergence, we need to show that there are constants c and d such that:
  - $\frac{1}{c^k}\mathbf{h}^{(k)}$ and $\frac{1}{d^k}\mathbf{a}^{(k)}$ converge to limits as k goes to infinity.

- Let's focus on the convergence of hub vectors $\frac{1}{c^k}\mathbf{h}^{(k)}$. The convergence proof of authority vectors $\frac{1}{d^k}\mathbf{a}^{(k)}$ is analogous.

Eigenvectors and Convergence

Let's focus on the sequence of hub vectors

- If $\frac{1}{c^k}\mathbf{h}^{(k)} = \frac{1}{c^k}(\mathbf{A}\mathbf{A}^\top)^k\mathbf{h}^{(0)}$ converges to a limit $\mathbf{h}^*$, then the direction of $\mathbf{h}^*$ shouldn't change when multiplied by $\mathbf{A}\mathbf{A}^\top$ although its length may grow by a factor of c.
  - That is, we expect $\frac{1}{c}\mathbf{A}\mathbf{A}^\top\mathbf{h}^* = \mathbf{h}^*$
  - or equivalently $\mathbf{A}\mathbf{A}^\top\mathbf{h}^* = c\mathbf{h}^*$

- Any vector satisfying this property (that does not change its direction when multiplied by a given matrix) is called an eigenvector of the matrix
  - The scaling constant c is called the eigenvalue corresponding to the eigenvector

- We expect $\mathbf{h}^*$ to be an eigenvector of the matrix $\mathbf{A}\mathbf{A}^\top$ with c a corresponding eigenvalue.

Eigenvectors and Convergence

Let's prove that the sequence of vectors $\frac{1}{c^k}\mathbf{h}^{(k)}$ converges to an eigenvector of the matrix $\mathbf{A}\mathbf{A}^\top$

- A square matrix $\mathbf{S}$ is *symmetric* if it remains the same after transposing it:
  - $S_{i,j} = S_{j,i}$ for every choice of i and j
  - in other words $\mathbf{S}^\top = \mathbf{S}$

- Every symmetric $n \times n$ matrix $\mathbf{S}$ has a set of n eigenvectors that are all unit vectors and all mutually orthogonal; that is, they form a basis for the space $\mathbb{R}^n$. [3]
  - Thus, $\mathbf{A}\mathbf{A}^\top$ has n mutually orthogonal eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ with corresponding eigenvalues $c_1, c_2, \ldots, c_n$, satisfying $|c_1| \geq |c_2| \geq \cdots \geq |c_n|$

- As $\mathbf{A}\mathbf{A}^\top$ is positive semidefinite (because $\mathbf{x}^\top\mathbf{A}\mathbf{A}^\top\mathbf{x} = (\mathbf{A}^\top\mathbf{x})^\top(\mathbf{A}^\top\mathbf{x}) \geq 0$), all the eigenvalues are non-negative, i.e., $c_1 \geq c_2 \geq \cdots \geq c_n \geq 0$.

[3] G. Strang. Linear Algebra and Learning from Data. 2019.

Eigenvectors and Convergence

$\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ is a set of mutually orthogonal unit vectors in $\mathbb{R}^n$.

- $\mathbf{h}^{(0)}$ can be represented as a linear combination of the vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$. That is, $\mathbf{h}^{(0)} = q_1\mathbf{z}_1 + q_2\mathbf{z}_2 + \cdots + q_n\mathbf{z}_n$ for coefficients $q_1, \ldots, q_n$
  - Here $\mathbf{h}^{(0)}$ can be an arbitrary vector in $\mathbb{R}^n$.
- We have:
$$\begin{aligned}(\mathbf{A}\mathbf{A}^\top)\mathbf{h}^{(0)} &= (\mathbf{A}\mathbf{A}^\top)(q_1\mathbf{z}_1 + q_2\mathbf{z}_2 + \cdots + q_n\mathbf{z}_n) \\ &= q_1\mathbf{A}\mathbf{A}^\top\mathbf{z}_1 + q_2\mathbf{A}\mathbf{A}^\top\mathbf{z}_2 + \cdots + q_n\mathbf{A}\mathbf{A}^\top\mathbf{z}_n \\ &= q_1c_1\mathbf{z}_1 + q_2c_2\mathbf{z}_2 + \cdots + q_nc_n\mathbf{z}_n\end{aligned}$$
  where the third equality follows from the fact that each $\mathbf{z}_i$ is an eigenvector with corresponding eigenvalue c_i of $\mathbf{A}\mathbf{A}^\top$.
- What this says is that $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ is a very useful set of coordinate axes for representing $\mathbf{h}^{(0)}$: multiplication by $\mathbf{A}\mathbf{A}^\top$ consists simply of replacing each term $q_i\mathbf{z}_i$ in the representation of $\mathbf{h}^{(0)}$ by $c_iq_i\mathbf{z}_i$.

Eigenvectors and Convergence

- As each successive multiplication by $\mathbf{A}\mathbf{A}^\top$ introduces an additional factor of c_i in front of the i^{th} term, we have

$$\mathbf{h}^{(k)} = (\mathbf{A}\mathbf{A}^\top)^k \mathbf{h}^{(0)} = c_1^k q_1 \mathbf{z}_1 + c_2^k q_2 \mathbf{z}_2 + \cdots + c_n^k q_n \mathbf{z}_n$$

- Dividing both sides by c_1^k leads to:

$$\frac{1}{c_1^k}\mathbf{h}^{(k)} = q_1 \mathbf{z}_1 + (\tfrac{c_2}{c_1})^k q_2 \mathbf{z}_2 + \cdots + (\tfrac{c_n}{c_1})^k q_n \mathbf{z}_n$$

- Assume that $c_1 > c_2$, then as k goes to infinity, every term but the first goes to 0.
  - $\frac{1}{c_1^k}\mathbf{h}^{(k)}$ tends to $q_1 \mathbf{z}_1$ as k goes to infinity.

Eigenvectors and Convergence

Let's show that the starting vector does not matter

- Instead of $\mathbf{h}^{(0)}$ being all coordinates equal to 1, let's choose another vector $\mathbf{x}$ with positive coordinates
  - Assume $\mathbf{x} = p_1\mathbf{z}_1 + p_2\mathbf{z}_2 + \cdots + p_n\mathbf{z}_n$
- $(\mathbf{A}\mathbf{A}^\top)^k\mathbf{x} = c_1^k p_1\mathbf{z}_1 + c_2^k p_2\mathbf{z}_2 + \cdots + c_n^k p_n\mathbf{z}_n$
- So $\frac{1}{c_1^k}\mathbf{h}^{(k)}$ is converging to $p_1\mathbf{z}_1$.
  - In other words, it is still converging to a vector in the direction of $\mathbf{z}_1$ despite the new choice for the starting vector $\mathbf{h}^{(0)} = \mathbf{x}$

Eigenvectors and Convergence

Let's show that q_1 and p_1 are not zero to show that $q_1\mathbf{z}_1$ is in fact a non-zero vector in the direction of $\mathbf{z}_1$

- We compute the inner product of $\mathbf{z}_1$ and $\mathbf{x}$

$$
\begin{aligned}
\langle \mathbf{z}_1, \mathbf{x} \rangle &= \langle \mathbf{z}_1, p_1\mathbf{z}_1 + \cdots + p_n\mathbf{z}_n \rangle \\
&= p_1\langle \mathbf{z}_1, \mathbf{z}_1 \rangle + p_2\langle \mathbf{z}_1, \mathbf{z}_2 \rangle + \cdots + p_n\langle \mathbf{z}_1, \mathbf{z}_n \rangle \\
&= p_1
\end{aligned}
$$

- p_1 is just the inner product of $\mathbf{x}$ and $\mathbf{z}_1$
- So, if our starting hub vector $\mathbf{h}^{(0)} = \mathbf{x}$ is not orthogonal to $\mathbf{z}_1$, then our sequence of vectors converges to a nonzero vector in the direction of $\mathbf{z}_1$

Eigenvectors and Convergence

Let's show that every positive vector $\mathbf{x}$ (i.e., with all coordinates positive) is not orthogonal to $\mathbf{z}_1$.

- Recall that $\mathbf{z}_1$ is a unit vector, and thus nonzero.
- $\mathbf{z}_1$ has only all nonnegative coordinates or all nonpositive coordinates.
  1. Since $\mathbf{z}_1$ is nonzero, there exists some nonnegative vector $\mathbf{x}$ (i.e., with all coordinates nonnegative) that is not orthogonal to $\mathbf{z}_1$, i.e., $p_1 = \langle \mathbf{x}, \mathbf{z}_1 \rangle$ is nonzero.
  2. Let $\mathbf{x}$ be any such a nonnegative vector. Since $\frac{1}{c_1^k}(\mathbf{A}\mathbf{A}^\top)^k\mathbf{x}$ has only nonnegative numbers and converges to $p_1\mathbf{z}_1$, $p_1\mathbf{z}_1$ has only nonnegative numbers.
- So if we consider the inner product of any positive vector with $\mathbf{z}_1$, the result must be nonzero. No positive vector can be orthogonal to $\mathbf{z}_1$.

$\Longrightarrow$ The sequence of hub vectors converge to a vector in the direction of $\mathbf{z}_1$

Reading

- Reading for this week
  - Chapter 13 of the textbook
  - Chapter 14 of the textbook, excluding the PageRank part

- Reading for next week
  - Chapter 14 of the textbook
  - Chapter 5 (link analysis) of Mining of Massive Datasets (3rd edition)
    http://www.mmds.org/