Large-Scale Networks

The Structure of the Web

Dr Vincent Gramoli | Lecturer
School of Information Technologies
Last week we looked at structural balance
- Signed graphs indicate simple + and – relations between individuals
- Signed graphs are too simplistic to measure trust in online ratings

This week, we will talk about hubs and authority
- What being authoritative mean
- How the authority gets transferred from one node to another
Information network: nodes are pieces of information and links join pieces of information that are related to each other

The World Wide Web is probably the most prominent information network

Although different from social networks studied before, it shares many similarities
The World Wide Web
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 [BCL+94].

It features two components:
- Makes the Internet available 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
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 Web uses the network metaphor
- Each page 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

NB. Information networks predate the development of computers and the creator of hypertext was motivated by earlier large-scale networks
Information on the Web is organized using a network metaphor: The links among the pages turns the web into a directed graph.
There are other ways of structuring information

- Alphabetic classification
- With folders

The Web makes logical relationship with the text (traditionally implicit) into explicit links
A precursor of hypertext is *citation*
- For authors to credit the source of an idea

Let’s consider the citations among a set of sociology papers that provided some of the key ideas mentioned earlier:

- Triadic closure
- The small-world phenomenon
- Structural balance
- Homophily
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.
The cross-references among a set of articles in an encyclopedia form another kind of information network that can be represented as a directed graph.

The figure shows the cross-references among a set of Wikipedia articles on topics in game theory, and their connections to related topics including popular and government agencies.
Associative memory

Understand how humans navigate Wikipedia

Get an idea of how people connect concepts
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.
A directed graph is **strongly connected** if there is a path from every node to every other node.

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 [BKM+00]

- They use SCC as building blocks
- They used the index of pages and links of AltaVista, one of the largest commercial search engine

This study was replicated on

- the larger index of Google’s search engine and
- large research collection of web pages
Their findings include:

1. The Web contains a giant strongly connected component (SCC):
   - search engines are like big indexes that connect big companies, governmental agencies, educational institutions)
   - It is frequent to find some of them that link back to the same site so that they are connected with each other

2. **IN**: Some nodes can reach the SCC but cannot be reached from it, these nodes are *upstream* of it

3. **OUT**: Some nodes can be reached from the SCC but cannot reach it, these nodes are *downstream* of it

4. There are nodes that can neither reach the SCC nor be reached from it
   - *Tendrils* are nodes that can be reached from IN that cannot reach the giant SCC and the one that can reach out but cannot be reached from the SCC
   - *Disconnected* are nodes that have no path to the giant SCC (even if ignoring directions)
A schematic picture of the bow-structure of the Web. Although the numbers are now outdated, the structure has persisted. [BKM+00]

Large-Scale Networks

Hubs and Authorities

Dr Vincent Gramoli | Lecturer
School of Information Technologies
The Problem of Ranking
The problem of ranking

- Type “University of Sydney” in Google’s search engine

The University of Sydney

sydney.edu.au/

Australia's leading higher education and research University.

4.3 ★★★★★ 161 Google reviews · Write a review · Google+ page

Uni Sydney Postgraduate - sydney.edu.au

Ad www.sydney.edu.au/postgraduate

Apply by 31 Jan 2015 to commence a postgrad degree at Uni of Sydney.
The problem of ranking

› How does Google’s search engine know which page to print first?

› Search engines only exploit information from the Web (no external info)

=> There should be enough information *intrinsic* to the Web to *rank* results
Automated information retrieval is old (1960’s)

...take **keyword queries** as input and outputs...
- Newspaper articles
- Scientific papers
- Patents
- Legal abstracts

Common problems:
- **Synonymy**: two words share the same meaning (e.g., funny, hilarious)
- **Polysemy**: one word has two meanings (*tiger* is an animal, Mac OS X version)
What has changed since the 1960’s?

1960’s
- Search queries were made by experts:
  - Reference librarians
  - Patent attorneys
  - People used to search collections of documents
- They knew how to use a controlled style and vocabulary

Nowadays
- Since ~2000 (Google): everyone is a searcher
- Since ~2005 (Web 2.0): everyone is an author
Diversity

- Diversity in authoring style makes it harder to rank documents
- Different searchers have different expectations for a given keyword

Inaccuracy

- Typing “World Trade Center” on the 11 September 2001 would lead to old results
- Google would rank based on pages previously indexed days/weeks ago

⇒ News search in search engine (not yet as fast as Twitter)

Needle-in-the-haystack

- Large volume of web pages, impossible to be accurate
- Most people are interested in filtering out most, how to select these?
Link Analysis
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
A list-find technique

› How to make deeper use of the network structure of the Web?

› Voting is not enough
  - Type “newspapers”, you may get high scores for prominent newspapers
  - Along with irrelevant highly ranked pages (Yahoo!, Facebook, Amazon)
A list-find technique

- Counting in-links to pages for the query “newspapers”.
- Unlabeled 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)

⇒ Vote number is a too simple measure
A list-find technique

What other information can complement vote measure?

What are the pages that compile lists of resources relevant to the topic?
- Such list 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 a the query “newspapers”

Let’s go back to the previous figure
- Few pages voted for many of the highly voted pages
- Pages have some sense of where the good answers are
- Page value as a list is the sum of the votes received by all pages for which it voted
Finding good lists for the query “newspapers”: each page’s value as a list is written as a number inside it. Again this number indicates the sum of the votes received by all pages for which it voted.
If we believe that **pages scoring well as lists** have a better sense for where the good results, we should **weight their votes more heavily**

Similarly, **people recommending lots of restaurants** may act as high-value lists so that you end up **giving them more value**
The principle of repeated improvement

- Re-weighting 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)
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.

To each page $p$, we assign pairs

- $\text{hub}(p)$ and
- $\text{auth}(p)$

Each page starting with $(1,1)$
Voting: Using the quality of hubs to refine our estimate for the quality of authorities

› **Authority update rule**: For each page $p$, update $\text{auth}(p)$ to be the sum of the hub scores of all pages that point to it.

List-finding: Using the quality of the authorities to refine our estimate of the quality of the hubs

› **Hub update rule**: For each page $p$, update $\text{hub}(p)$ to be the sum of the authority scores of all pages that it points to.
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 them by the sum of all scores)
After 1 application of the Authority Update Rule (assuming that hub(p) = 1 for every p)
Hubs and authorities

After second application of the Authority Update Rule

After normalization (sum of authorities was 125)
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 lowering changes

- The same limits are reached whatever initial values we choose for hubs and authorities
- They are properties of the link structure (not initial values)

- Ultimately, we read 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
Limiting hub and authority values for the query “newspapers”
Applications beyond the Web
Application of link analysis

› Link analysis techniques have divers applications:
  - In any domain where information is connected by a network structure
  - Citation analysis
  - U.S. Supreme Court citations

› Citation analysis
  - Garfield’s impact factor is a standard measure for scientific journals
    - Average number of citations received by a paper in the given journal over the past two years
  - U.S. Supreme Court Citations
Impact factor

- A standard measure for scientific journals

- Garfield’s IF: Average number of citations received by a paper in the given journal over the past two years [Gar72]

- **Repeated improvement**: In the 1970’s, extension of impact factor to weight citations from journals of high impact factor

- Led to influence weight for journals [PN76]
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
Raising and falling of some key 5th Amendment cases (20th century)
- 1936 Brow vs. Mississippi about confessions obtained under torture
- 1966 Miranda vs. Arizona: the need for citations to the former quickly declined
Significant decisions can vary widely in the rate at which they acquire authority.

- Roe vs. Wade grew in authority rapidly.
- Brown vs. Board of Education only began acquiring significant authority a decade after it was issued because it was initially weak and then strengthened by the Civil Rights Act in 1964.
Spectral Analysis of Hubs and Authorities
Set of $n$ pages represented as nodes labeled 1, 2, 3, ..., $n$

Links are encoded in an adjacency $n \times n$ matrix $M$
- $M_{ij}$ (ith row, jth column) = 1 if there is a link from node i to j
- $M_{ij}$ = 0 otherwise

Example: the directed hyperlinks among Web pages represented as a $n$ adjacency matrix $M$
Let’s consider the hub and authority rules in terms of matrix multiplication

› For every node $i$,
  - its hub score is denoted $h_i$
  - its authority score is denoted $a_i$

› Hub vector is denoted $h$, and authority vector is $a$

› **Hub Update Rule (formalized with matrix notation):**

\[
h_i = M_{i1} \cdot a_1 + M_{i2} \cdot a_2 + M_{i3} \cdot a_3 + \ldots + M_{in} \cdot a_n. \quad (1)
\]

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

\[
h = Ma.
\]
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)
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):**

$$a_i = M_{1i} \cdot h_1 + M_{2i} \cdot h_2 + M_{3i} \cdot h_3 + \ldots + M_{ni} \cdot h_n.$$  \hspace{1cm} (2)

The roles of columns and rows are interchanged so we use the transpose of matrix $M$ denoted $M^T$ defined by the property that $(i,j)$ entry of $M^T$ is the $(j,i)$ entry of $M$ (i.e., $M_{ij}^T = M_{ji}$).

$$a = M^T h.$$
Let’s perform the k-step hub-authority computation for large values of $k$.

- Let $a^{(o)}$ and $h^{(o)}$ be the vectors of whose coordinates equal to 1.
- Let $a^{(k)}$ and $h^{(k)}$ denote the vectors of authority and hubs after $k$ application of Authority and then Hub Update Rules in order.

Following previous formula we find that:

$$a^{(1)} = M^T h^{(0)}$$

and:

$$h^{(1)} = Ma^{(1)} = MM^T h^{(0)}.$$

That’s the result of the one-step hub-authority computation.
In the 2\textsuperscript{nd} step, we therefore get:

\[ a^{(2)} = M^T h^{(1)} = M^T M M^T h^{(0)} \]

and:

\[ h^{(2)} = M a^{(2)} = M M^T M M^T h^{(0)} = (M M^T)^2 h^{(0)}. \]

In the 3\textsuperscript{rd} step, we get:

\[ a^{(3)} = M^T h^{(2)} = M^T M M^T M M^T h^{(0)} = (M^T M)^2 M^T h^{(0)} \]

and:

\[ h^{(3)} = M a^{(3)} = M M^T M M^T h^{(0)} = (M M^T)^3 h^{(0)}. \]

What do we observe?
Conclusion: \( a^{(k)} \) and \( h^{(k)} \) are products of the terms \( M \) and \( M^T \) in alternating order, where \( a^{(k)} \) begins with \( M^T \) and the expression for \( h^{(k)} \) begins with \( M \).

We can write:

\[
a^{(k)} = (M^T M)^{k-1} M^T h^{(0)}
\]

and:

\[
h^{(k)} = (M M^T)^k h^{(0)}.
\]

The authority and hub vectors are the results of multiplying an initial vector by larger and larger powers of \( M^T M \) and \( M M^T \), respectively.
Eigenvectors and Convergence
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 c and d such that:
- \( h^{(k)} / c^k \) and \( a^{(k)} / d^k \) converge to limits as k goes to infinity.
Let’s focus on the sequence of hub vectors first

- If $h^{(k)} / c^k = (MM^T)^k h^{(0)} / c^k$ is going to a limit $h^{(*)}$, then $h^{(*)}$ shouldn’t change when multiplied by $MM^T$, although its length may grow by a factor of $c$.

- That is, we expect: $(MM^T)^k h^{(*)} = ch^{(*)}$.

- 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 $h^{(*)}$ to be an eigenvector of the matrix $MM^T$ with $c$ a corresponding eigenvalue.
Eigenvectors and convergence

Let’s prove that the sequence of vectors $h^{\langle k \rangle} / c^k$ converges to an eigenvector of the matrix $MM^T$

› A square matrix $A$ is *symmetric* if it remains the same after transposing it:
  - $A_{ij} = A_{ji}$ for each choice of $i$ and $j$
  - in other words $A^T = A$

› Every symmetric $n \times n$ matrix $A$ has a set of $n$ eigenvectors that are all unit vectors and all mutually orthogonal; that is, they form a basis for the space $\mathbb{R}^n$. [LM06]

› Since $MM^T$ is *symmetric* we can apply this fact to it. Let’s write the resulting mutually orthogonal eigenvectors $z_1, z_2, \ldots, z_n$ with corresponding eigenvalues $c_1, c_2, \ldots, c_n$, respectively, assuming that $|c_1| \geq |c_2| \geq \ldots \geq |c_n|$. 


Eigenvectors and convergence

Given any vector \( x \), we can think of the matrix-vector product as \( (MM^T)x \)

\( x \) is a linear combination of the vectors \( z_1, \ldots, z_n \)

\[ x = p_1z_1 + p_2z_2 + \ldots + p_nz_n \] for coefficients \( p_1, \ldots, p_n \).

We have:

\[ (MM^T)x = (MM^T)(p_1z_1 + p_2z_2 + \ldots + p_nz_n), \]
\[ = p_1MM^TZ_1 + p_2MM^TZ_2 + \ldots + p_nMM^TZ_n, \]
\[ = p_1c_1z_1 + p_2c_2z_2 + \ldots + p_nc_nz_n, \]

where the third equality follows from the fact that \( z_i \) is an eigenvector.

What this says is that \( z_1, z_2, \ldots, z_n \) is a very useful set of coordinate axes for representing \( x \): multiplication by \( MM^T \) consists simply of replacing each term by \( p_i z_i \) in the representation of \( x \) by \( c_i p_i z_i \).
Eigenvectors and convergence

As each successive multiplication by $MM^T$ introduces an additional factor of $c_i$ in front of the $i^{th}$ term, we have

$$(MM^T)^k x = c_1^k p_1 z_1 + c_2^k p_2 z_2 + \cdots + c_n^k p_n z_n.$$  

$h^{(0)}$ can be represented as a linear combination $q_1 z_1 + q_2 z_2 + \cdots + q_n z_n$, so:

$$h^{(k)} = (MM^T)^k h^{(0)} = c_1^k q_1 z_1 + c_2^k q_2 z_2 + \cdots + c_n^k q_n z_n.$$  

Dividing both sides by $c_1^k$ leads to:

$$h^{(k)} / c_1^k = q_1 z_1 + (c_2/c_1)^k q_2 z_2 + \cdots + (c_n/c_1)^k q_n z_n.$$  

As we assumed that $|c_1| > |c_2|$ as $k$ goes to infinity, every term but the first goes to 0.

$h^{(k)} / c_1^k$ tends to $q_1 z_1$ as $k$ goes to infinity.
Let’s show that the **starting vector does not matter**

- Instead of $h^{(0)}$ with all coordinates equal to 1, let’s choose another vector $x$ with positive coordinates

- $(MM^T)^k x = c_1^k p_1 z_1 + c_2^k p_2 z_2 + ... + c_n^k p_n z_n$.

- So $h^{(k)} / c_1^k$ is converging to $p_1 z_1$. 

Let’s show that $q_1$ and $p_1$ are not zero to show that $q_1z_1$ is in fact a non-zero vector in the direction of $z_1$.

- We compute the inner product of $z_1$ and $x$

$$z_1 \cdot x = z_1 \cdot (p_1z_1 + \ldots + p_nz_n)$$

$$= p_1(z_1 \cdot z_1) + p_2(z_1 \cdot z_2) + \ldots + p_n(z_1 \cdot z_n)$$

$$= p_1.$$

- $p_1$ is just the inner product of $x$ and $z_1$

- So, if our starting hub vector $h^{(0)} = x$ is not orthogonal to $z_1$ then our sequence of vectors converges to a nonzero vector in the direction of $z_1$. 
Let’s show that our starting hub vector $h^{(0)} = x$ is not orthogonal to $z_1$

› It is not possible to have a positive vector (with all coordinates positive) orthogonal to $z_1$

1. It is not possible for every positive vector to be orthogonal to $z_1$, so there is some positive vector $x$ for which $(MM^T)^k x / c_1^k$ converges to a nonzero vector $p_1 z_1$.

2. Since $(MM^T)^k x / c_1^k$ only has nonnegative numbers that converge to $p_1 z_1$, $p_1 z_1$ has only nonnegative coordinates and at least one positive coordinate (as it is nonzero).

3. So if we consider the inner product of any positive vector with $p_1 z_1$, the result must be positive. No positive vector can be orthogonal to $z_1$.

› The sequence of hub vectors converge to a vector in the direction of $z_1$
References


