Social Network Analysis for Computer Scientists

F.W. Takes

Leiden University, The Netherlands

Lecture 4 — Web structure and propagation-based centrality measures
So . . .

- Everyone has chosen a topic; we have 15 topics
- Did you pick a paper from the list of papers of your particular topic?
- Did you pick a presentation slot?
- Notice that there will be no class on October 9 and November 20?
- Any questions about Assignment 1?
- Or questions about the course project (presentation and paper)?
World Wide Web

Figure: Tim Burners-Lee
World Wide Web

- Around since 1990
- Chaos of webpages: how to order?
World Wide Web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories
Around since 1990

Chaos of webpages: how to order?

Webdirectories
  - Submission services
World Wide Web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories
  - Submission services
- Search engines based on term frequency
World Wide Web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories
  - Submission services
- Search engines based on term frequency
  - Keyword stuffing
World Wide Web

- Around since 1990
- Chaos of webpages: how to order?
- Webdirectories
  - Submission services
- Search engines based on term frequency
  - Keyword stuffing
- Search engines based on “smart” webpage ranking
  - So how?
Webgraph

- **Webgraph**: directed unweighted network \( G = (V, E) \)
- Nodes \( V \) are webpages
- Links \( E \) are “hyperlinks” to other pages
- Many dense subgraphs . . .
Webgraph

- **Webgraph**: directed unweighted network \( G = (V, E) \)
- Nodes \( V \) are webpages
- Links \( E \) are “hyperlinks” to other pages
- Many dense subgraphs . . .
- Pages may belong to the same domain
- Alternative: draw webgraph with only (sub)domains as nodes
- Idea: search engine ranks the webpages using the structure of the webgraph
- Centrality measures
Centrality measures

- **Distance/path-based measures:**
  - Degree centrality \(O(n)\)
  - Closeness centrality \(O(mn)\)
  - Betweenness centrality \(O(mn)\)
  - Eccentricity centrality \(O(mn)\)

- **Propagation-based measures:**
  - Hyperlink Induced Topic Search (HITS)
  - PageRank and Random Walks
Hyperlink Induced Topic Search

- A link to a page is a “vote” for that page
- But how important is the page casting the vote?
- **Hyperlink Induced Topic Search (HITS)**
  - **Hubs**: pages that link to good authorities
  - **Authorities**: contain useful information and are therefore linked from many good hubs

Hyperlink Induced Topic Search

Query: Top automobile makers

Hyperlink Induced Topic Search

Leskovec, Stanford CS224W (http://cs224w.stanford.edu)
Hubs and Authorities

- A “good webpage” is either a hub or an authority
- Each page $v \in V$ has two scores:
  - **Hub score** $h(v)$
  - **Authority score** $a(v)$
- Iterative algorithm
- Rules/definitions are somewhat “recursive”
- **Propagation model** that updates state at time $t + 1$ based on $t$
HITS algorithm

- For all nodes $v \in V$, at $t = 0$ initialize $a^0(v) = h^0(v) = 1/\sqrt{n}$
- Repeat:
  1. $t = t + 1$
  2. Update the authority scores, so for all nodes $v \in V$:
     $a^{t+1}(v) = \sum_{v \in N'(v)} h^t(v)$
  3. Update the hub scores, so for all nodes $v \in V$:
     $h^{t+1}(v) = \sum_{v \in N(v)} a^t(v)$
  4. Normalize both scores so that
     $\sum_{v \in V} (a^{t+1}(v))^2 = \sum_{v \in V} (h^{t+1}(v))^2 = 1$

- Until scores converge:
  $\sum_{v \in V} (a^{t+1}(v) - a^t(v))^2 < \epsilon$ and
  $\sum_{v \in V} (h^{t+1}(v) - h^t(v))^2 < \epsilon$
For some small $\epsilon$. 
HITS algorithm (easy mode)

- For all nodes $v$, **initialize** the hub and authority scores equally
- Repeat:
  1. $t = t + 1$
  2. **Update the authority score** of all nodes $v$ to the sum of the hub scores of the nodes pointing to $v$
  3. **Update the hub score** of all nodes $v$ to the sum of the authority scores of the nodes to which $v$ points
  4. **Normalize** both scores so that they sum to 1
- Until values **converge**: between iteration $t$ and $t + 1$ the values of both scores differ less than $\epsilon$
HITS complexity

- Memory: 2 lists of size $n$ for hub and authority scores, so $O(n)$
- Time: Update and normalize $n$ values in each iteration based on their neighborhoods of average size $(m/n)$, so $O(n \cdot (m/n)) = O(m)$
- Usually 100 iterations for convergence, so $100 \cdot m$
- Compare this to betweenness or closeness centrality which takes $O(mn)$ time . . .
Winner takes it all
PageRank

- A link to a page is a “vote” for that page
- But how important is the page casting the vote?
- PageRank answer: that just depends on how many other pages vote for that page
- **PageRank**: number from 0 (low) to 10 (high) that indicates the importance of a page
- Similar to eigenvector centrality
Towards PageRank

- Each page has its own importance $PR(v)$
- Each page $v$ casts equal votes of size $\frac{PR(v)}{\text{deg}(v)}$ for all other pages $w \in N(v)$ that it links to
  (in practice, rel="nofollow" prevents this)
- The amount of importance $PR(v)$ that a page receives depends on the pages that link to it: $PR(v) = \sum_{w \in N'(v)} \frac{PR(w)}{\text{deg}(w)}$
- Again recursive
- Does it converge?
Towards PageRank Example
Challenges

\[ PR(v) = \sum_{w \in N'(v)} \frac{PR(w)}{\text{deg}(w)} \]

- **Spider traps**: links back and forth:
  
  \[ \circ u \leftrightarrow \circ v \]

- **Dead ends**: pages that do not have outgoing links

\[ \circ w \rightarrow \circ v \rightarrow \circ x \]

\[ \circ w \]
Towards PageRank

- **Random Surfer** model
- Idea: a user browsing the web either clicks a link or opens an arbitrary page
- With probability $p$, follow a link to a neighbor
- With probability $1 - p$, jump to a random node
- In practice: $p = 0.85$ and thus $1 - p = 0.15$ ("follow five links and jump")
PageRank algorithm (semi-easy mode)

- For all nodes $v \in V$, at $t = 0$ initialize $PR^0(v) = (1/n)$
- Repeat:
  1. $t = t + 1$
  2. $PR^t(v) = \frac{1-p}{n} + p \cdot \sum_{w \in N'(v)} \frac{PR^{t-1}(w)}{deg(w)}$
  3. Normalize so that $\sum_{v \in V} PR(v) = 1$
     (just divide each value by the sum of all values)
- Until scores converge:
  $\sum_{v \in V} |PR^t(v) - PR^{t-1}(v)| < \epsilon$
PageRank centrality

- PageRank $C_{PR}(v)$, which is the value of $PR(v)$ after iteratively and simultaneously applying:

$$
PR(v) = \frac{1-p}{n} + p \left( \sum_{w \in N'(v)} \frac{PR(w)}{\text{deg}(w)} \right)
$$

for each of the nodes $v \in V$ and then normalizing the values so that they sum to 1, where $PR(v)$ is initialized to $1/n$ and $N'(v)$ is the set of nodes that links to node $v$ and $p = 0.85$

- 100 iterations is usually enough

- Time $100 \cdot m$, roughly equal to HITS.
And more . . .

- Jump to relevant pages with higher probability
- Choose a relevant neighbor with higher probability
- **Relevance** based on keywords, previous visits, geo aspects, . . .
- Computation and definition using matrices
- Many other PageRank variants . . .
- Personalized PageRank
PageRank and Beyond
Actual Google (Page)Rank

- PageRank $PR(v)$
- Relevant keywords
- User’s search history
- Local aspects
- “Rewards and punishments”
Elaine Washburn aan info

Hi Frank,

Please see proposal for [redacted] below:

We represent several industries that might interest you:

- Online gaming: you would receive 150 USD per year
- Finance, telecommunications, tourism or health: you would receive 100 USD per year

The advert will be text, not a visual banner. It will appear on a single page of your website. We aim to complete payment via secure payment partners Paypal or Moneybookers within 48 hours of the advert going live on your site.

Additionally, please read our terms and conditions: [www.moredigital.com/terms.pdf](http://www.moredigital.com/terms.pdf)

Please let me know which industry you prefer, we'll then let you know which client fits your site best and draft an advert!

Best regards,
Elaine
The Internship
WHAT THE HELL IS SEO?

GOOGLE TOP SEARCH RESULT:
SEO - STORA ENSO OYJ IS A
PAPER AND BOARD
MANUFACTURER IN EUROPE..

AND I THOUGHT THAT WAS
SOMETHING ABOUT BEING
ON TOP OF GOOGLE...

SEO COMIC

by JOZZUA - ISULONG-SEOPH.NET
The Web

http://www.cheswick.com/ches/map/gallery/index.html
Why study the webgraph?

- Understanding social mechanisms that govern growth
- Designing ranking methods
- Devising better crawling algorithms
- Creating accurate models of the web’s structure

Webgraph in 1999

- Altavista: 200 million nodes
- 186 million nodes in the weakly connected component (90% of the links)
- 56 million nodes in the strongly connected component
- Power law degree distribution
- Average distance of 16 (if there is a path, 25% of the cases)
- Average (undirected) distance of 6.83
- Diameter is 28

Webgraph in 1999

Crawling the webgraph in 2012

- Crawled by Common Crawl Foundation
- First half of 2012
- Breadth-first visiting strategy
- Heuristics to detect spam pages
- Seeded with the list of domains from a previous crawl and a set of URLs from Wikipedia

Webgraph in 2012

- Page graph
- Host graph
- PLD graph (Pay-Level Domain (PLD): a subdomain of a public top-level domain, for which users have to pay. PLDs identify a single user or organization)

<table>
<thead>
<tr>
<th>Granularity</th>
<th># Nodes in millions</th>
<th># Arcs in millions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page Graph</td>
<td>3 563</td>
<td>128 736</td>
</tr>
<tr>
<td>Host Graph</td>
<td>101</td>
<td>2 043</td>
</tr>
<tr>
<td>PLD Graph</td>
<td>43</td>
<td>623</td>
</tr>
</tbody>
</table>

Table 1: Sizes of the graphs

Degree

- Compared to 1999, average degree increased from 7.5 to 36.8
- Perhaps due to use of content management systems (they tend to create dense websites)

## Centrality in the Webgraph

<table>
<thead>
<tr>
<th>PageRank</th>
<th>Indegree</th>
<th>Harmonic Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>gmpg.org</td>
<td>wordpress.org</td>
<td>youtube.com</td>
</tr>
<tr>
<td>wordpress.org</td>
<td>youtube.com</td>
<td>en.wikipedia.org</td>
</tr>
<tr>
<td>youtube.com</td>
<td>gmpg.org</td>
<td>twitter.com</td>
</tr>
<tr>
<td>livejournal.com</td>
<td>en.wikipedia.org</td>
<td>google.com</td>
</tr>
<tr>
<td>tumblr.com</td>
<td>tumblr.com</td>
<td>wordpress.org</td>
</tr>
<tr>
<td>en.wikipedia.org</td>
<td>twitter.com</td>
<td>flickr.com</td>
</tr>
<tr>
<td>twitter.com</td>
<td>google.com</td>
<td>facebook.com</td>
</tr>
<tr>
<td>networkadvertising.org</td>
<td>rtalabel.org</td>
<td>apple.com</td>
</tr>
<tr>
<td>promodj.com</td>
<td>wordpress.com</td>
<td>vimeo.com</td>
</tr>
<tr>
<td>skriptmail.de</td>
<td>rtalabel.org</td>
<td>creativecommons.org</td>
</tr>
<tr>
<td>parallels.com</td>
<td>mp3shake.com</td>
<td>amazon.com</td>
</tr>
<tr>
<td>tistory.com</td>
<td>w3schools.com</td>
<td>adobe.com</td>
</tr>
<tr>
<td>google.com</td>
<td>domains.lycos.com</td>
<td>myspace.com</td>
</tr>
<tr>
<td>miibeian.gov.cn</td>
<td>staff.tumblr.com</td>
<td>w3.org</td>
</tr>
<tr>
<td>phpbb.com</td>
<td>club.tripod.com</td>
<td>bbc.co.uk</td>
</tr>
<tr>
<td>blog.fc2.com</td>
<td>creativecommons.org</td>
<td>nytimes.com</td>
</tr>
<tr>
<td>tw.yahoo.com</td>
<td>vimeo.com</td>
<td>yahoom.com</td>
</tr>
<tr>
<td>w3schools.com</td>
<td>miibeian.gov.cn</td>
<td>microsoft.com</td>
</tr>
<tr>
<td>wordpress.com</td>
<td>facebook.com</td>
<td>guardian.co.uk</td>
</tr>
<tr>
<td>domains.lycos.com</td>
<td>phpbb.com</td>
<td>imdb.com</td>
</tr>
</tbody>
</table>

WCC of the Webgraph

- Weakly Connected Component (WCC)
- 91.8% in 1999, 94% in 2012
- Component size distribution

SCC of the Webgraph

- Strongly Connected Component (SCC): 51.3% of the nodes
- Computation required 1TB of RAM
- Graph compression framework WebGraph was used

Bow-tie structure in 2012

Bow-tie structure in 2012

<table>
<thead>
<tr>
<th>Component</th>
<th>Common Crawl 2012</th>
<th>Broder et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># nodes (in thousands)</td>
<td>% nodes (in %)</td>
</tr>
<tr>
<td>LSCC</td>
<td>1 827 543</td>
<td>51.28</td>
</tr>
<tr>
<td>IN</td>
<td>1 138 869</td>
<td>31.96</td>
</tr>
<tr>
<td>OUT</td>
<td>215 409</td>
<td>6.05</td>
</tr>
<tr>
<td>TENDRILS</td>
<td>164 465</td>
<td>4.61</td>
</tr>
<tr>
<td>TUBES</td>
<td>9 099</td>
<td>0.26</td>
</tr>
<tr>
<td>DISC.</td>
<td>208 217</td>
<td>5.84</td>
</tr>
</tbody>
</table>

Table 3: Comparison of sizes of bow-tie components

Distances and diameter

- Diameter lower bound: 5,282

Figure: Cumulative Distance Distribution

Webgraph conclusions

- Measurements on the largest webgraph available to the public
- Average degree has significantly increased, almost by a factor of 5
- Connectivity has increased, average distance has decreased
- Structure of the web appears dependent on the specific web crawl
- The distribution of indegrees and outdegrees is extremely different

Making a “better” webgraph

- Not just an unweighted unlabeled directed network
- **Resource Description Framework (RDF):** link is a triple 
  `[subject] [predicate] [object]`
- Link weighting: define a weight for outgoing links (to give hints to PageRank algorithm)
- Link annotation: make more use of the `rel=""` attribute to describe the kind of link: alternate, search, next, etc.
- Requires new algorithms for ranking …
For next week . . .

- Finish the homework assignment
- Take a look at the paper(s) for your course project
- Start preparing your presentation
- Report any questions or unclarities w.r.t. the course deliverables