So . . .

- Send me an e-mail at ftakes@liacs.nl if you never got my e-mail in the first week.
- Did you pick a topic and a presentation time slot?
- Questions about the homework assignment?
- 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
World Wide Web

- 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 \(O(1)\)
  - 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
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
        $$\sum_{v \in V} (h^{t+1}(v) - h^t(v))^2 < \epsilon$$
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 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 \( \frac{PR(v)}{deg(v)} \) 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: \( PR(v) = \sum_{w \in N'(v)} \frac{PR(w)}{deg(w)} \)
- 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:
  \[ u \rightarrow v \]

- **Dead ends**: pages that do not have outgoing links
  \[ W \rightarrow V \rightarrow X \rightarrow W \]
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:

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

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.
PageRank
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 . . .
PageRank and Beyond
Actual Google (Page)Rank

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

Also, please read our terms and conditions: 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
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.

- ...

- Would require new algorithms for ranking ...
Webgraph

- 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

Diameter . . .

- Example presentation
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
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
Credits

Slides partially based on course by Jure Leskovec — Stanford CS224W — http://cs224w.stanford.edu