Google

PageRank – selecting the right documents

"PageRank reflects our view of the importance of web pages by considering more than 500 million variables and 2 billion terms."

the "perfect search engine"

"understands exactly what you mean and gives you back exactly what you want."

-- Larry Page, Google

making all the world's information universally accessible and useful

"knowledge is always good, and certainly always better than ignorance"

-- Sergey Brin, Google

Web of Links

- Uniform Resource Identifier
  - Uniform: there's a standard generic syntax
  - Resource: what a URI identifies -- anything at all
  - Identifier: URIs are identifiers, i.e., names, not addresses
  - Addresses: URLs are URI that locate resources

- Thanks to Henry Thompson

web links

http://www.ltg.ed.ac.uk/~ht/identity/ComponentGraph.svg?.....#.....

scheme | domain | path | query (optional) | fragment

More examples:

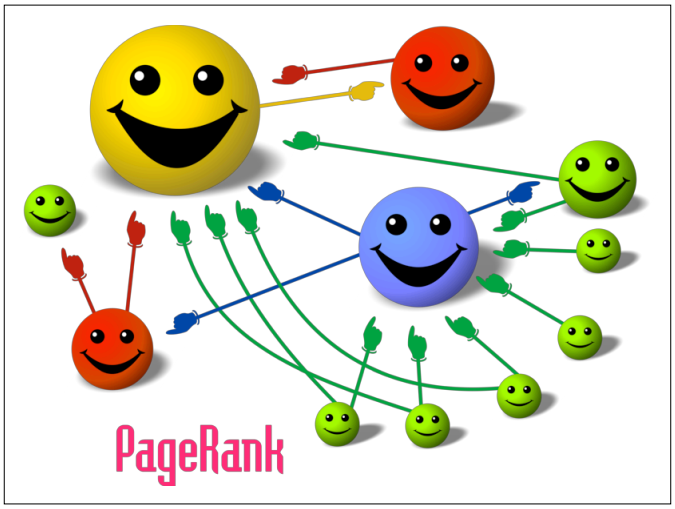
```
http://www.ed.ac.uk/
ftp://ftp.funet.fi/pub/mirrors/perl/
mailto:ht@inf.ed.ac.uk [that's a bit weird. . .]
file:///D:/Documents/HTalks/Inaugural/slides.html [this talk]
http://localhost/ht/Documents/HTalks/Inaugural/slides.html [likewise]
http://www.ltg.ed.ac.uk/~ht/travel.html#travel
http://maps.google.co.uk/maps?q=10+Crichton+Street,+Edinburgh&layer=c&
sll=55.944586,-3.187494&cbp=13,332.96,,0,1.84&cbll=55.944532,-3.18738&
hl=en&sspn=0.006295,0.006295&ie=UTF8&hq=&hnear=10+Crichton+St,+Edinburgh+EH8,+United+Kingdom&
ll=55.944532,-3.18738&spn=0.000012,0.006727&t=m&z=17& vpsrc=0&panoid=e8EPc2P5vtC6b-
oE8SlGiw
https://mail.google.com/mail/?ui=2&shva=1#inbox
```

A query is passed to the server and it can use this to decide what information to send back.

e.g., used to ask Google maps for a map showing a particular place

A fragment identifies a part of a web page

e.g., a named section identified by an id=”name”



Everyone knows that Google uses an algorithm called “page rank” (after Larry Page).

A way to rank “importance” of pages on the web.

Idea is simple – if lots of important pages point to a given page, then that page is important.

"random surfer" model

Repeat forever,

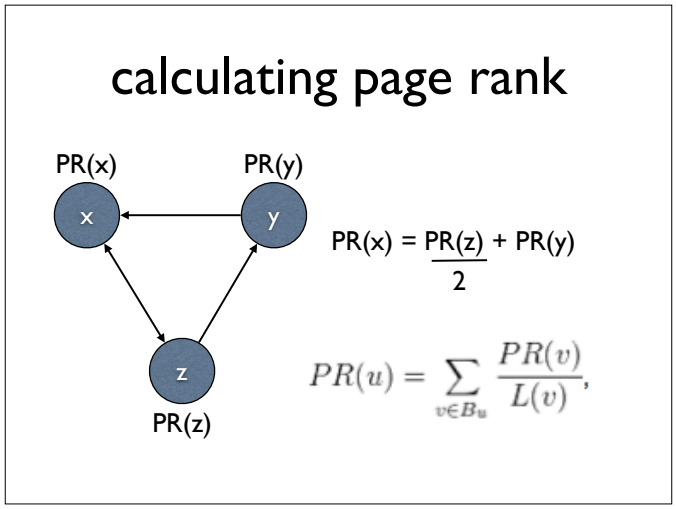
1. Choose a random number *0 < r <= 1*
2. *If r* < λ,
   - *Click the "surprise me" button*

   Else
   - *Click a link at random on the current page*

- Page rank of a page, p, is the proportion of times we expect the surfer to see page p

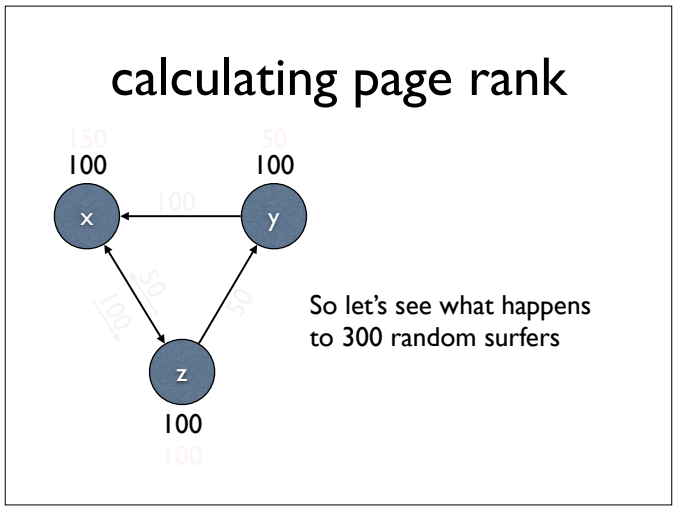
calculating page rank

- Each page sends an equal share of its PageRank to each of the pages it links to,
- then updates its PageRank by adding together the PageRank shares sent in all of the messages it receives



the PageRank for any page **u** is dependent on the PageRank values for each page **v** in the set $\mathbf{B_u}$ (this set contains all pages linking to page **u**), divided by the number $L(v)$ of links from page **v**.

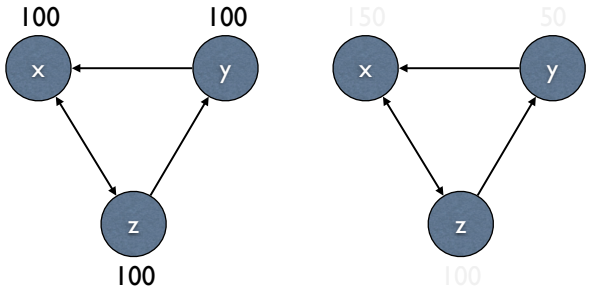
We could write down 3 equations and solve them with linear algebra, but in the general case, it gets a bit complicated



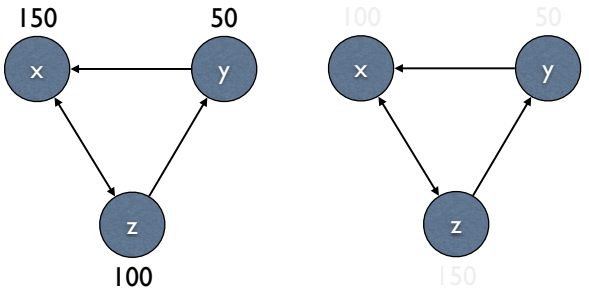
So instead let's use the random surfer model and see what happens....

We'll ignore the "surprise me" button for now, and just split up the surfers evenly between the links leaving a page

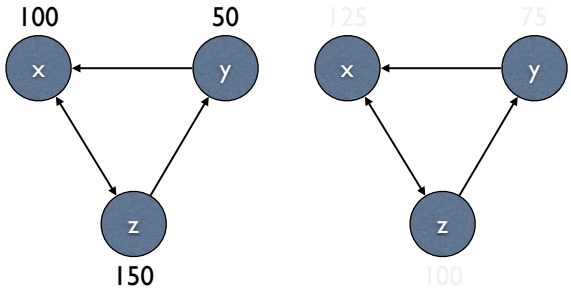
calculating page rank



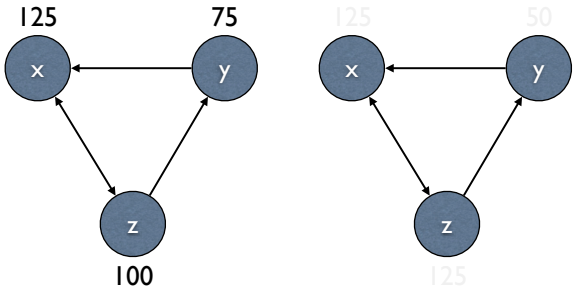
calculating page rank



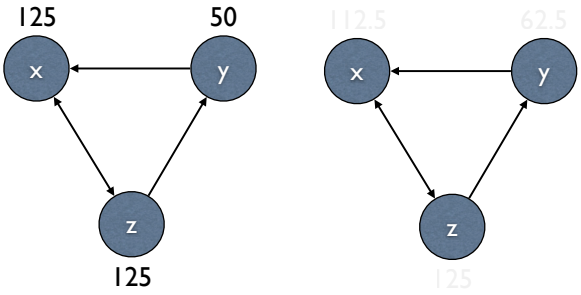
calculating page rank



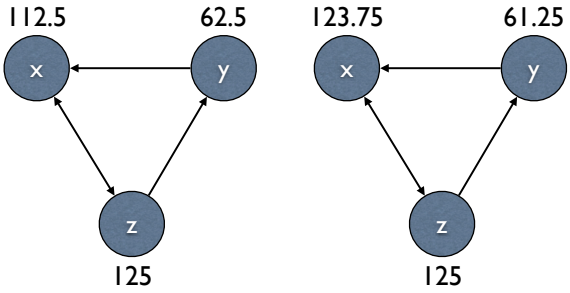
calculating page rank

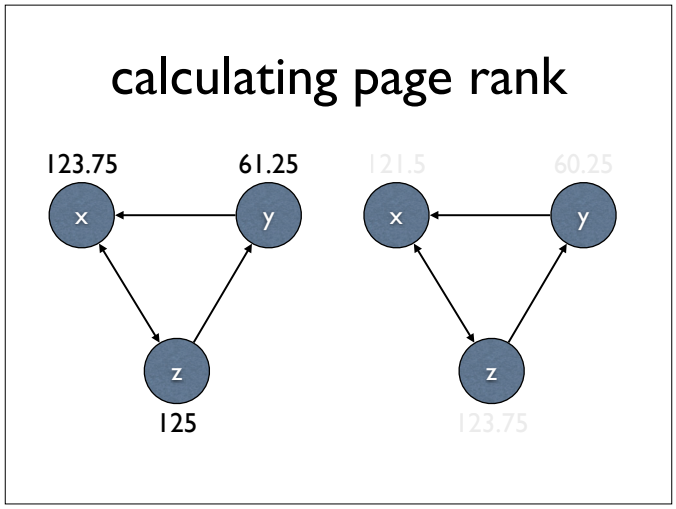


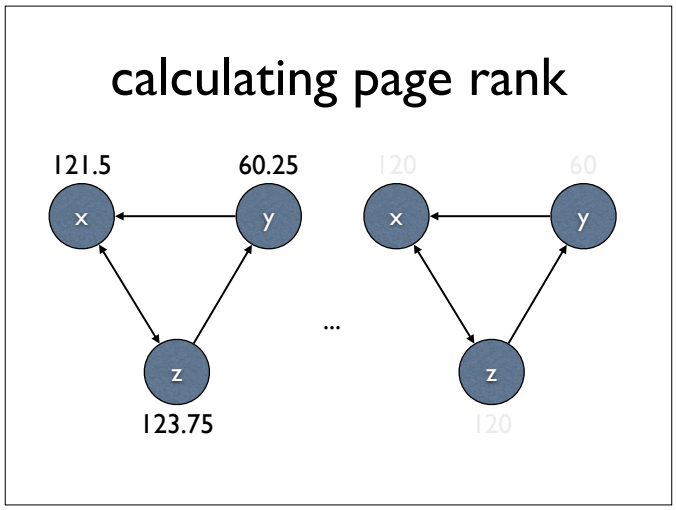
calculating page rank



calculating page rank

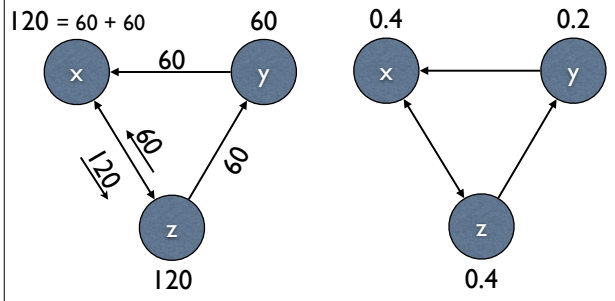




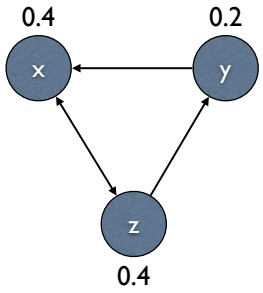


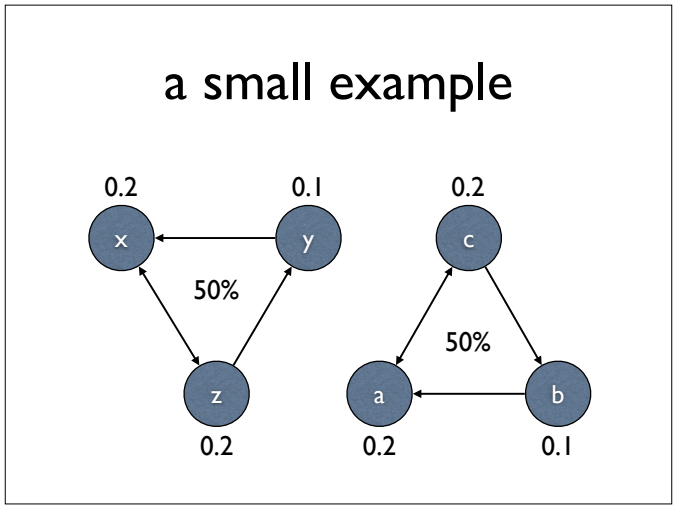
The numbers get closer and closer to the algebraic solution

calculating page rank



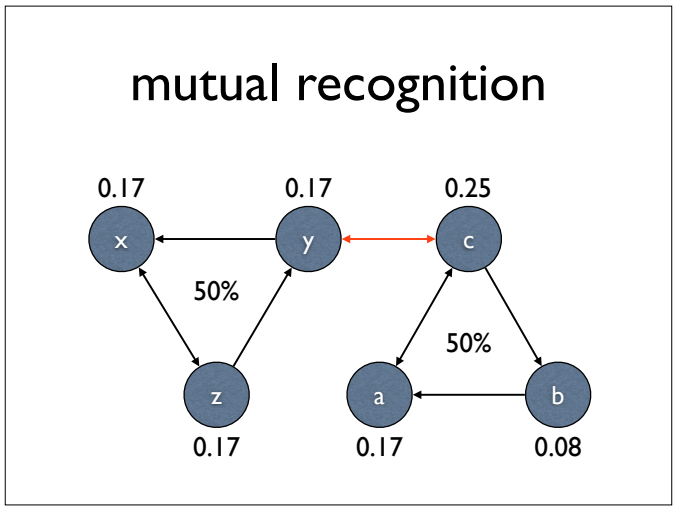
we're done!





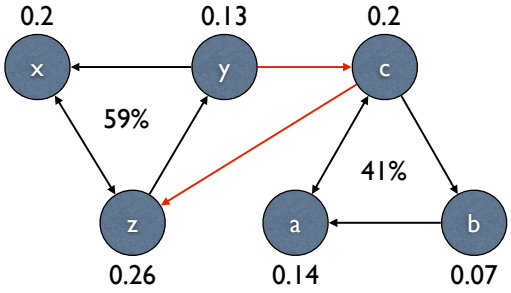
"Surprise me" button is needed to jump to disconnected sites and out of pages with no links.

A typical value of λ is 10 or 15%



The two sites linked to each other get a bigger share of page rank than before

another link



Some of c's PR is passed to z

Because of such effects, people try to fool the system by adding useless links to get better page rank.

Google watches out for such behaviour and black-lists offending sites (when they can be identified)

factors used by search engines

```
words on page
words in metadata
words in URL
links to page
'anchor' words on links to page
```

We will look at how a site Google ranks highly for "large cats" uses these ...

words on page

- used to generate 'big index'
- abused by
  - hidden text
    - Making Font Color the Same As Background Color
    - Making Font Color the Same Color As Background Image
    - Making Font Color Almost Match Background Color or Background Image
- Google says:
  - "Trying to deceive (spam) our web crawler by means of hidden text, ... compromises the quality of our results and degrades the search experience for everyone. We think that's a bad thing."

Hidden text is textual content which your visitors cannot see, but which is still readable by the search engines.

The idea is to load a Web page with keywords and keyword phrases that would be unsightly to visitors but that would improve the page's rankings in the search engine results, and to do so without letting your visitors see the text.

Hidden text is identified as search spam by each of the major search engines.

words in metadata

```
<head>
  <title>Largest Domestic Cat Breed</title>
  <meta name="description"
        content=
  "Possibly a definitive answer as to which cat is the
  largest domestic cat breed including ancillary
  information gathered along the way."
  />
  <meta name="keywords"
        content="Largest Domestic Cat Breed,largest domestic cat"
    />
  ...
</head>
```

Meta tags are HTML codes that are inserted into the header on a web page, after the title tag. They take a variety of forms and serve a variety of purposes, but in the context of search engine optimization when people refer to meta tags, they are usually referring to the meta description tag and the meta keywords tag.

There are other useful meta tags, including the meta copyright tag, and the meta author tag, among others. These tags are used to instruct user agents such as web browsers and search engine spiders on a variety of topics.

Unfortunately, so many unscrupulous webmasters have abused the meta description and meta keywords tag that search engines have had to de-emphasize their importance.

words in metadata

```
<img alt="large Maine Coon cat"
src="http://www.pictures-of-cats.org/
images/large-maine-coon-cat.jpg">


<a
 href="http://www.pictures-of-cats.org/
 largest-domestic-cat-breed.html"
  >Go to Largest Cat Breed</a>
```

The alt tag for images is designed to provide information to those who cannot view the image (eg for accessibility by the blind).

Using keywords in file-names can help.

Using keywords in anchor text for links from other pages (or other sites) can help.

Google's advice

- Make a site with a clear hierarchy and text links.
  - Every page should be reachable from at least one static text link.
- Offer a site map to your users with links that point to the important parts of your site.
  - If the site map has an extremely large number of links, you may want to break the site map into multiple pages.
- Keep the links on a given page to a reasonable number.
- Create a useful, information-rich site,
  - write pages that clearly and accurately describe your content.

Google's advice

- Think about the words users would type to find your pages,
  - make sure that your site actually includes those words within it.
- Try to use text instead of images to display important names, content, or links.
  - The Google crawler doesn't recognize text contained in images. If you must use images for textual content, consider using the "ALT" attribute to include a few words of descriptive text.
- Make sure that your <title> elements and ALT attributes are descriptive and accurate.
- Check for broken links and correct HTML.

Google's advice

- Make pages primarily for users, not for search engines.
  - Don't deceive your users or present different content to search engines than you display to users, which is commonly referred to as "cloaking."
- Avoid tricks intended to improve search engine rankings.
  - A good rule of thumb is whether you'd feel comfortable explaining what you've done to a website that competes with you. Another useful test is to ask, "Does this help my users? Would I do this if search engines didn't exist?"
- Don't participate in link schemes designed to increase your site's PageRank.
  - In particular, avoid links to web spammers or "bad neighborhoods" on the web, as your own ranking may be affected adversely by those links.
- Don't use unauthorized computer programs to submit pages, check rankings, etc.
  - Such programs consume computing resources and violate our Terms of Service. Google does not recommend the use of products such as WebPosition Gold™ that send automatic or programmatic queries to Google.

Google's advice

- Avoid hidden text or hidden links.
- Don't use cloaking or sneaky redirects.
- Don't send automated queries to Google.
- Don't load pages with irrelevant keywords.
- Don't create multiple pages, subdomains, or domains with substantially duplicate content.
- Don't create pages with malicious behavior, such as phishing or installing viruses, trojans, or other badware.
- Avoid "doorway" pages created just for search engines, or other "cookie cutter" approaches such as affiliate programs with little or no original content.
- Provide unique and relevant content that gives users a reason to visit your site.

Hiding text or links in your content can cause your site to be perceived as untrustworthy since it presents information to search engines differently than to visitors.

Cloaking refers to the practice of presenting different content or URLs to users and search engines. Serving up different results based on user agent may cause your site to be perceived as deceptive and removed from the Google index.

"Keyword stuffing" refers to the practice of loading a webpage with keywords in an attempt to manipulate a site's ranking in Google's search results. Filling pages with keywords results in a negative user experience, and can harm your site's ranking. Focus on creating useful, information-rich content that uses keywords appropriately and in context.

In some cases, content is deliberately duplicated across domains in an attempt to manipulate search engine rankings or win more traffic. Deceptive practices like this can result in a poor user experience, when a visitor sees substantially the same content repeated within a set of search results.

The term "malware" covers all sorts of malicious software designed to harm a computer or network. Kinds of malware include (but are not limited to) viruses, worms, spyware, and Trojan horses. Once a site or computer has been compromised, it can be used to host malicious content such as phishing sites (sites designed to trick users into parting with personal and credit card information).

Doorway pages are typically large sets of poor-quality pages where each page is optimized for a specific keyword or phrase. In many cases, doorway pages are written to rank for a particular phrase and then funnel users to a single destination.