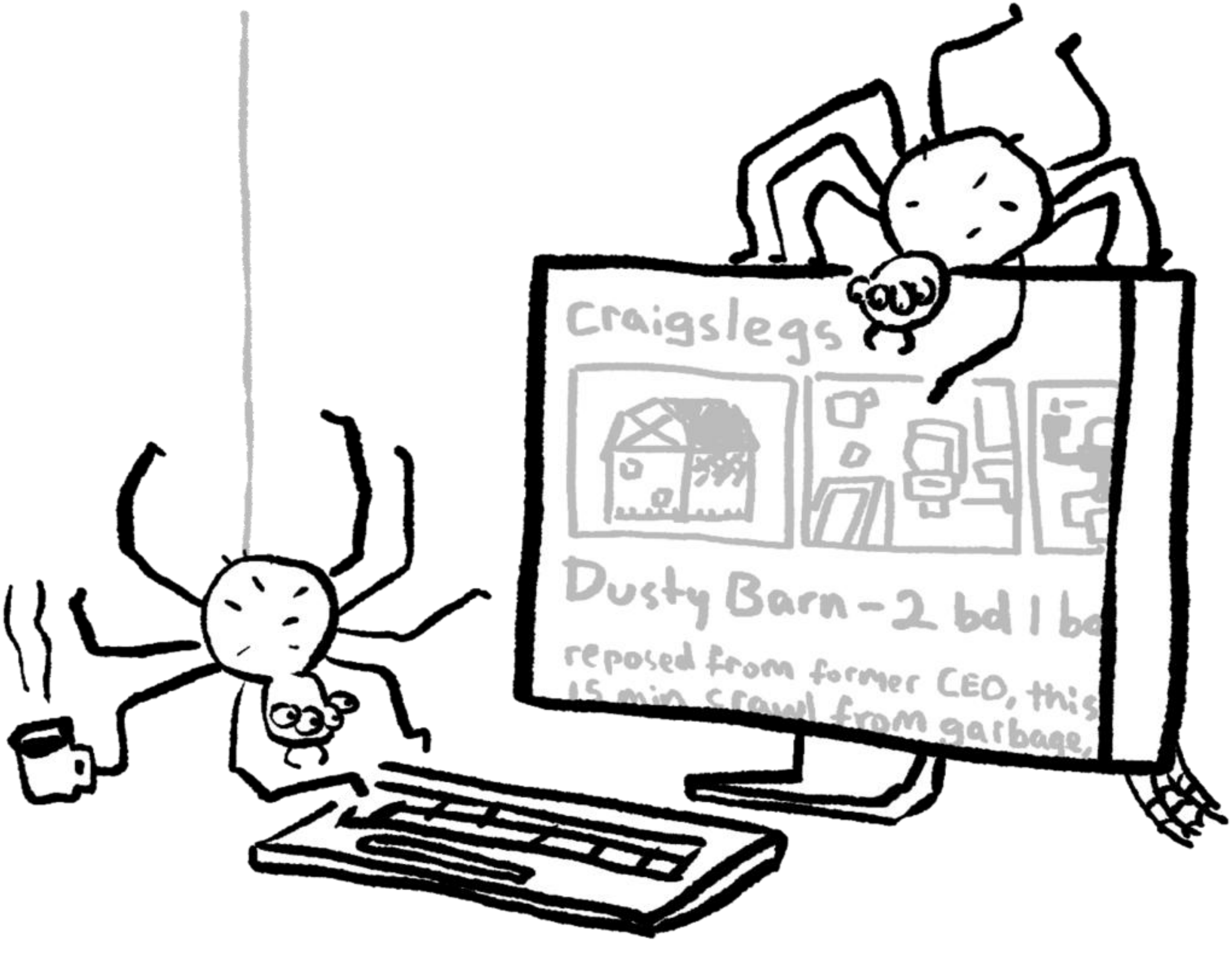
Crawling The Web

Powered by @Brilliant App

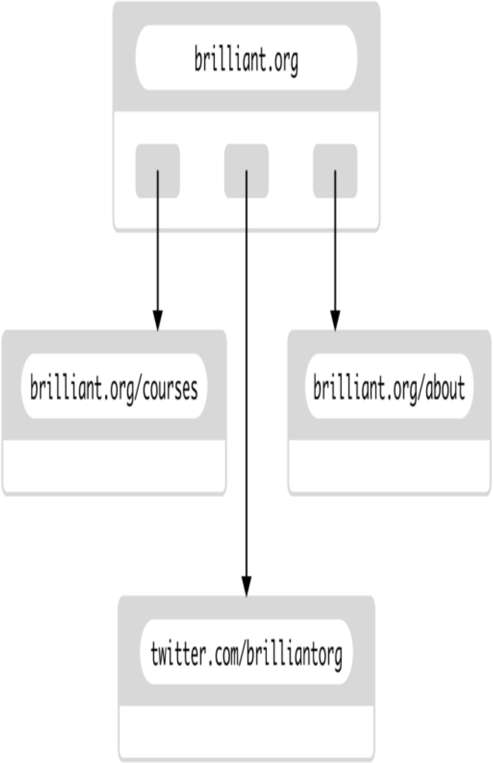
Lecture prepared and delivered by

Dr Syed Khaldoon Khurshid

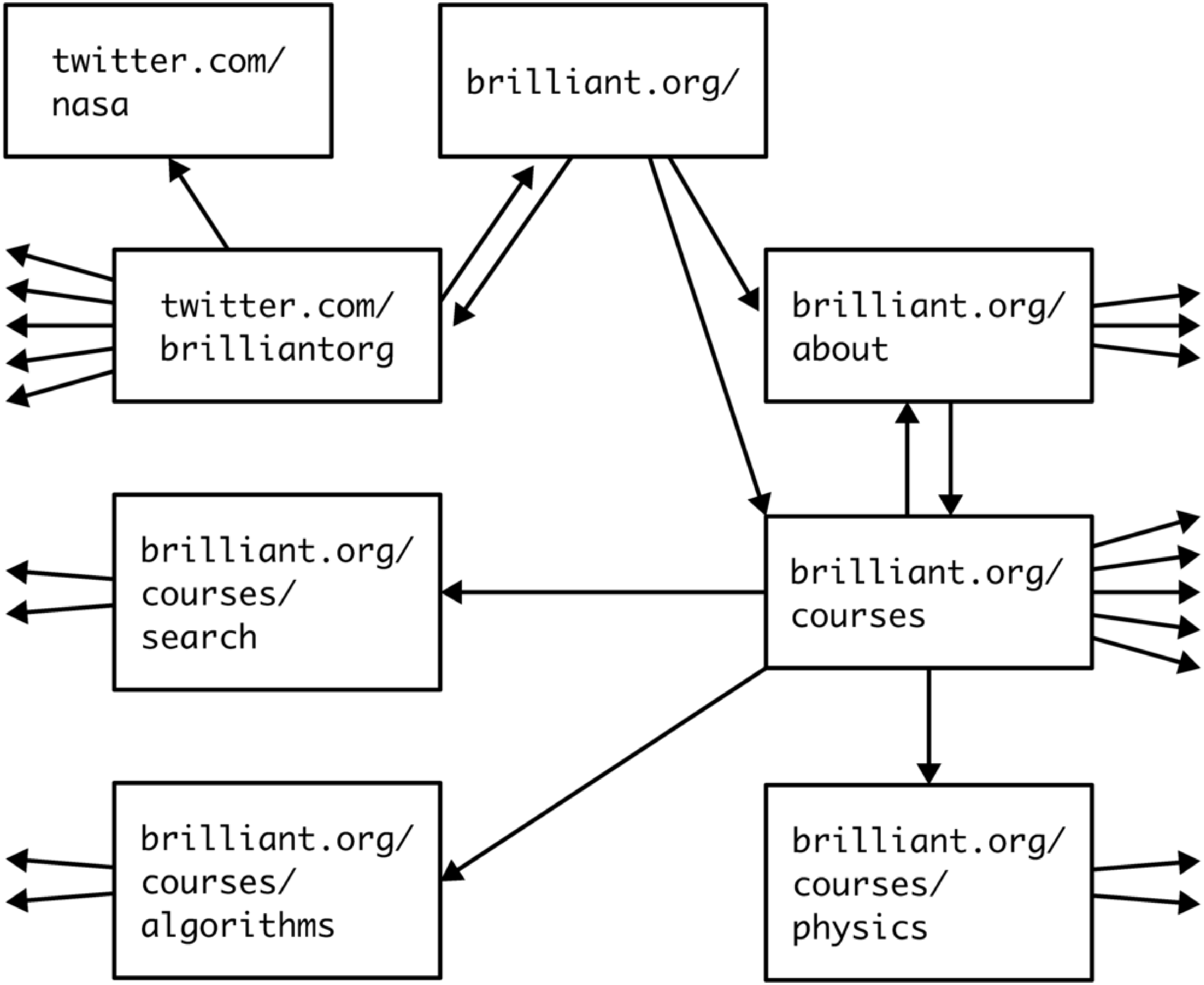
- If you want to build a card catalog, you need to look at all the books in a library or collection. If you want to build a concordance for a book, you need to look at all the words in the book.
- If you want to build a search engine for every website on the internet, where do you even start? Where do you find web pages to put in your web search engine?
- The answer involves Graphs and Spiders.


Craigslegs
Dusty Barn - 2 bd 1 ba
reposed from former CEO, this
15 min crawl from garbage,

- ***Links* are part of everyday life. In the mid-1980s, links were a big new idea, and one that made the World Wide Web very different from previous internet services**. Email existed without links for more than a decade before the web was invented.
- Links are also the key mechanism that lets an internet search engine find web pages.



- One way to think about the web is to ignore where the links are on every page, and ignore the text on every page, just keeping track of which pages link to other pages.
- If you make a box for every URL, and then you draw an arrow from the box where a link appears to the box where the link leads, you end up with a structure that is familiar to computer scientists: a Graph.

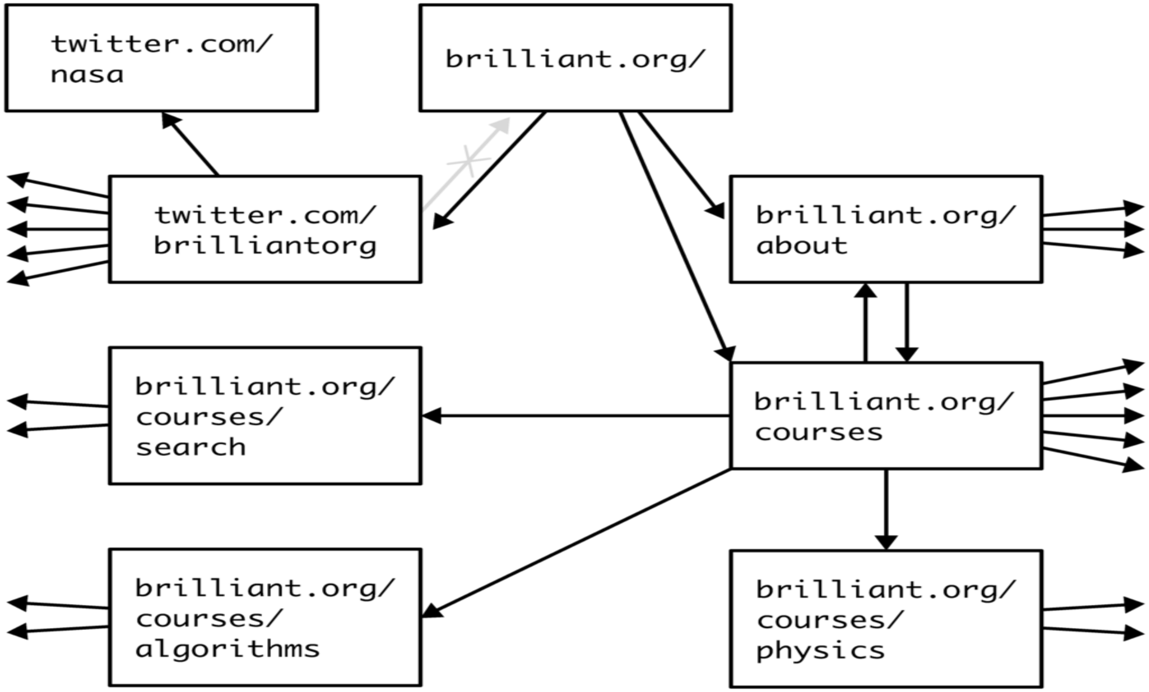

twitter.com/
nasa
brilliant.org/
twitter.com/
brilliantorg
brilliant.org/
about
brilliant.org/
courses/
search
brilliant.org/
courses
brilliant.org/
courses/
algorithms
brilliant.org/
courses/
physics

- The vertices in this graph are marked with URLs. The arrows are also called *edges*. Every edge starts on a page that has a link, and the arrow points to the URL where the link directs you.
- If Twitter removes the link from twitter.com/brilliantorg/ to brilliant.org/, what happens to this graph of the web?

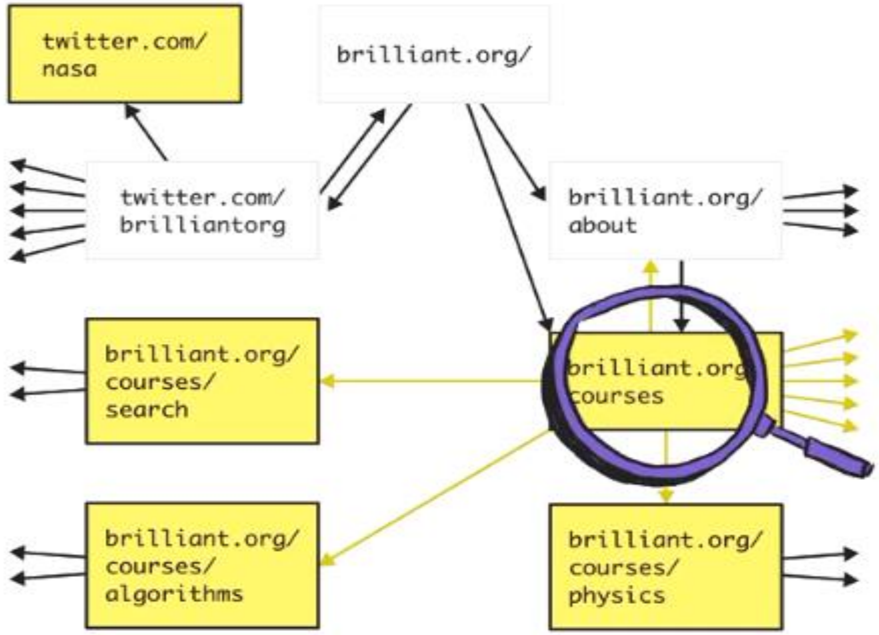
Correct answer: **One edge gets deleted.**

Explanation

- When Twitter removes a link to Brilliant, the link in the other direction still exists.
- Therefore, only one edge has been removed from the web graph

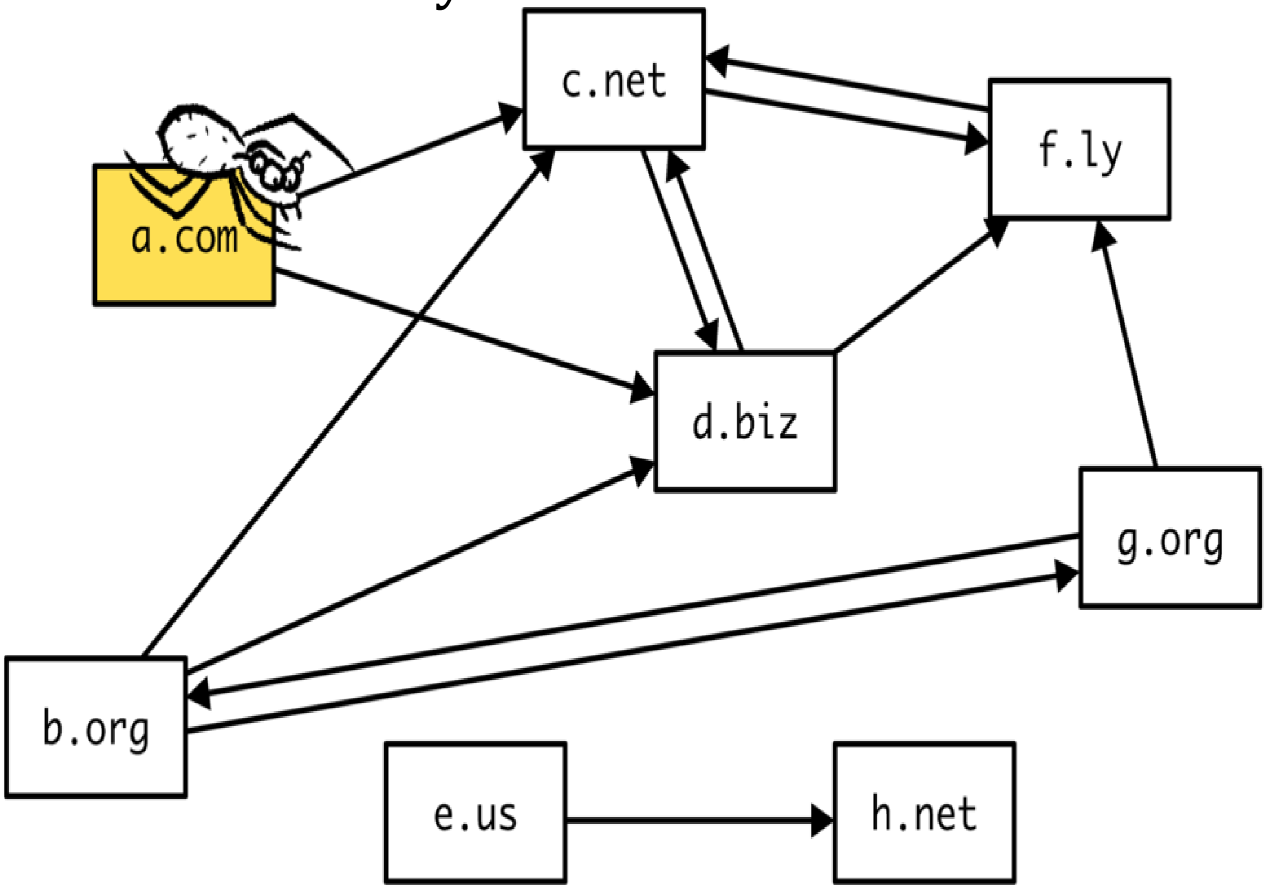

twitter.com/ nasa
brilliant.org/
twitter.com/ brilliantorg
brilliant.org/ about
brilliant.org/ courses/ search
brilliant.org/ courses
brilliant.org/ courses/ algorithms
brilliant.org/ courses/ physics

- The below illustrates one way for an algorithm to explore the graph of the web. The algorithm keeps track of a group of URLs that it knows about but hasn't yet read, the **"frontier,"** shown in yellow.
- On each round, a URL is selected and read. All of the previously unexplored URLs linked on that page are now known to the algorithm, **so they are added to the frontier.**


twitter.com/ nasa
brilliant.org/
twitter.com/ brilliantorg
brilliant.org/ about
brilliant.org/ courses/ search
brilliant.org/ courses
brilliant.org/ courses/ algorithms
brilliant.org/ courses/ physics

Programs that perform search on the graph of the web quickly became called "**web crawlers**," which then got renamed "**spiders**."

- A web spider starts crawling the web graph above at a.com and keeps crawling until it has run out of websites to visit. Select *all* websites that will be visited by a spider that starts in this way.



- Web spiders were the first tool that took advantage of the web as a graph. Following links, significant parts of the web could be identified, read, and incorporated into search engines. **The process was a never-ending one, because the web graph is constantly changing and growing as URLs are added and links are modified.**

- Almost all of the early search engines in the **early 1990s took advantage of web crawlers**, using the graph structure of the web in order to find and index pages that could then be searched by users.
- In the late 1990s, Google got its start by using the same graph structure in a different way. **Google figured out a way to analyze the web graph and measure a page's popularity**. This popularity metric, **PageRank**, made Google's search engine incredibly effective at ranking different web pages that matched a search query.

Any
Questions
?