Information Networks and the WWW

Structure of the Web

Information Network & WWW

- Information Network
- History of WWW
- Structure of WWW

Information Network & WWW

- Information Network
- History of WWW
- Structure of WWW

Information Network

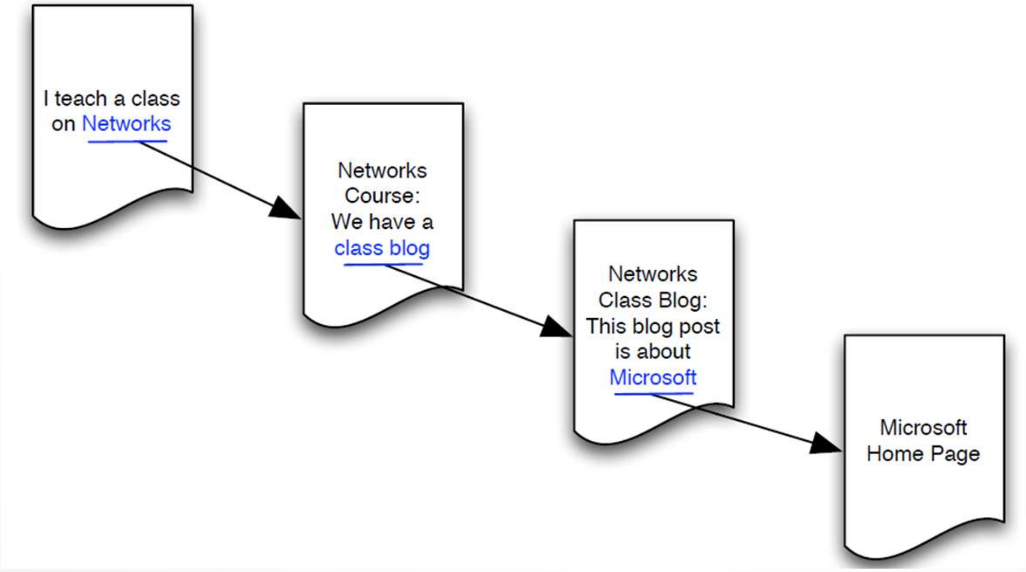
- Similarity with Social Networks:
  - Both graphs
  - Have some similar properties (Connected Components, Power Laws)

- Differences with Social Network:
  - Nodes represent information, not people
  - Links mostly one-directional

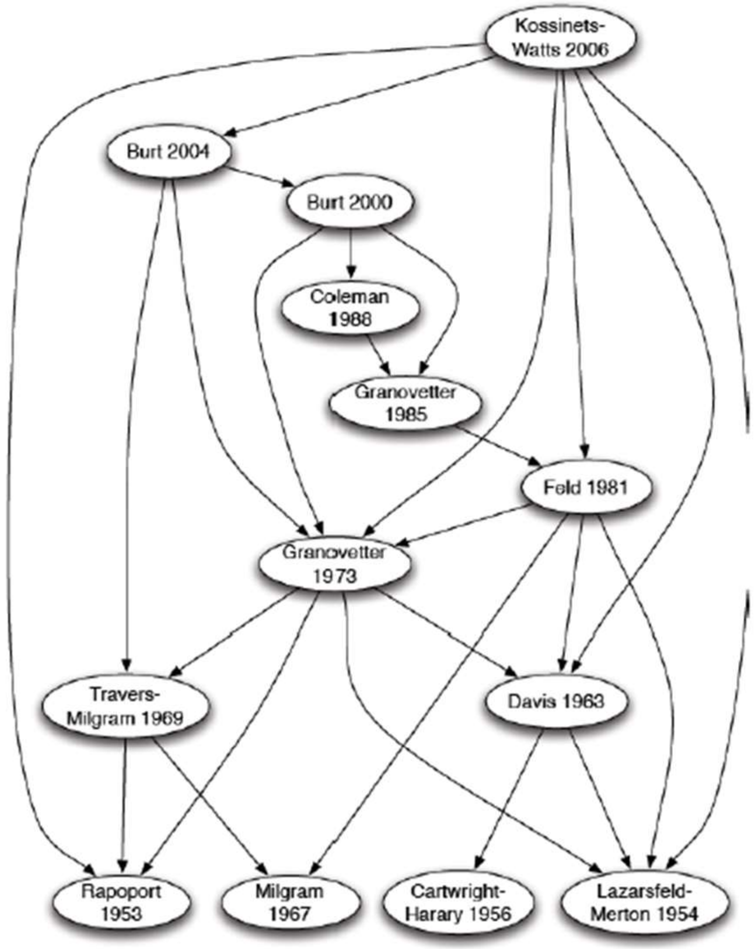
- Which Social Networks are mostly unidirectional?
  - Name Recognition Network
  - Twitter Follower Network

World Wide Web

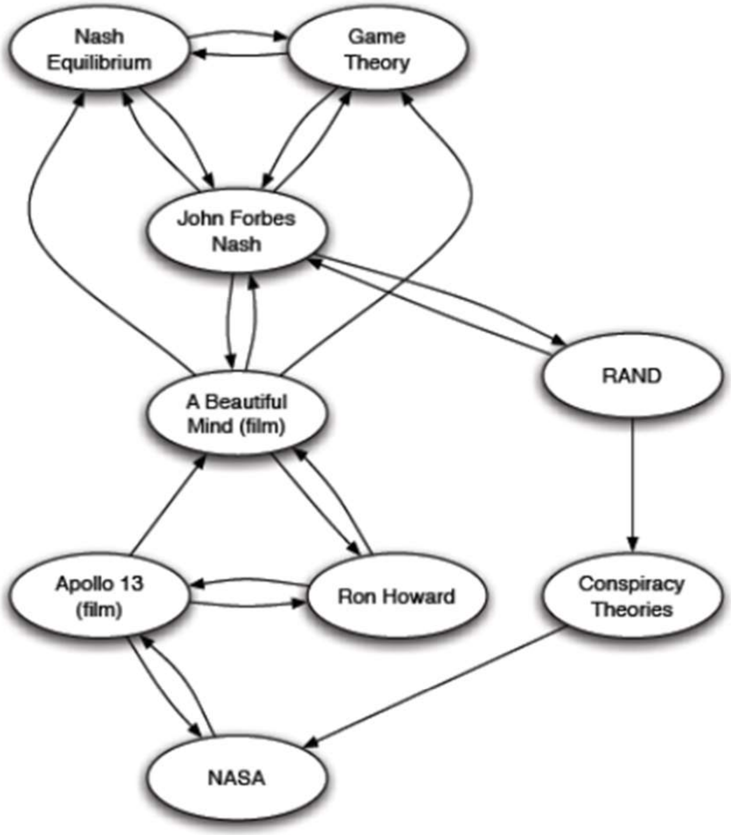
- Web is an application developed to let people share information over the Internet
- Created by Tim Berners-Lee during the period 1989-1991 [1]
- First, it provided a way for you to make documents easily available to anyone on the Internet, in the form of Web pages that you could create and store on a publically accessible part of your computer.
- Second, it provided a way for others to easily access such Web pages, using a browser that could connect to the public spaces on computers across the Internet and retrieve the Web pages stored there.


I teach a class on Networks
Networks Course: We have a class blog
Networks Class Blog: This blog post is about Microsoft
Microsoft Home Page

Other Information Networks



Citations



References in an Encyclopedia

Information Network & WWW

- Information Network
- History of WWW
- Structure of WWW

Hypertext

- Precursor of WWW
  - Since middle of the twentieth century [2]
  - Replace the traditional linear structure of text with a network structure, in which any portion of the text can link directly to any other part.

Memex

- Vannevar Bush and his seminal 1945 article in the Atlantic Monthly entitled "*As We May Think*" [3]
- Traditional methods for storing information in a book, a library, or a computer memory are highly linear — they consist of a collection of items sorted in some sequential order.
- Our conscious experience of thinking, on the other hand, exhibits associative memory like a semantic network represents
  - You think of one thing; it reminds you of another; you see a novel connection; some new insight is formed.
- Bush called for the creation of information systems that mimicked this style of memory Memex
  - Functioned very much like the Web, consisting of digitized versions of all human knowledge connected by associative links
- Bush's article foreshadowed not only the Web itself, but also many of the dominant metaphors that are now used to think about the Web:
  - the Web as universal encyclopedia
  - the Web as giant socio-economic system
  - the Web as global brain
- Vannever Bush's vision was so accurate is not in any sense coincidental:
  - Bush occupied a prominent position in the U.S. government's scientific funding establishment
- Tim Berners-Lee invoked Bush's ideas when he set out to develop the Web

Evolution of Web

- In the 1990s, the 1st decade of the Web, in which it grew rapidly from a modest research project to a vast new medium with global reach.
- In the early phase of this period:
  - most pages were relatively static documents
  - most links served primarily navigational functions to transport you from one page to another, according to the relational premise of hypertext
- The Web has increasingly outgrown the simple model
- Links on the Web of two types:
  - Navigational: serving the traditional hypertextual functions of the Web
  - Transactional: perform transactions on the computers hosting the content.

Web 2.0

- In 2000s changes in Web:
  1. The growth of Web authoring styles that enabled many people to collectively create and maintain shared content;
  2. the movement of people's personal on-line data (including e-mail, calendars, photos, and videos) from their own computers to services offered and hosted by large companies;
  3. the growth of linking styles that emphasize on-line connections between people, not just between documents
- Showcases several "social phenomena":
  - Software that gets better the more people use It
  - The wisdom of crowd
  - The Long Tail

Information Network & WWW

- Information Network
- History of WWW
- Structure of WWW

Strongly Connected Component

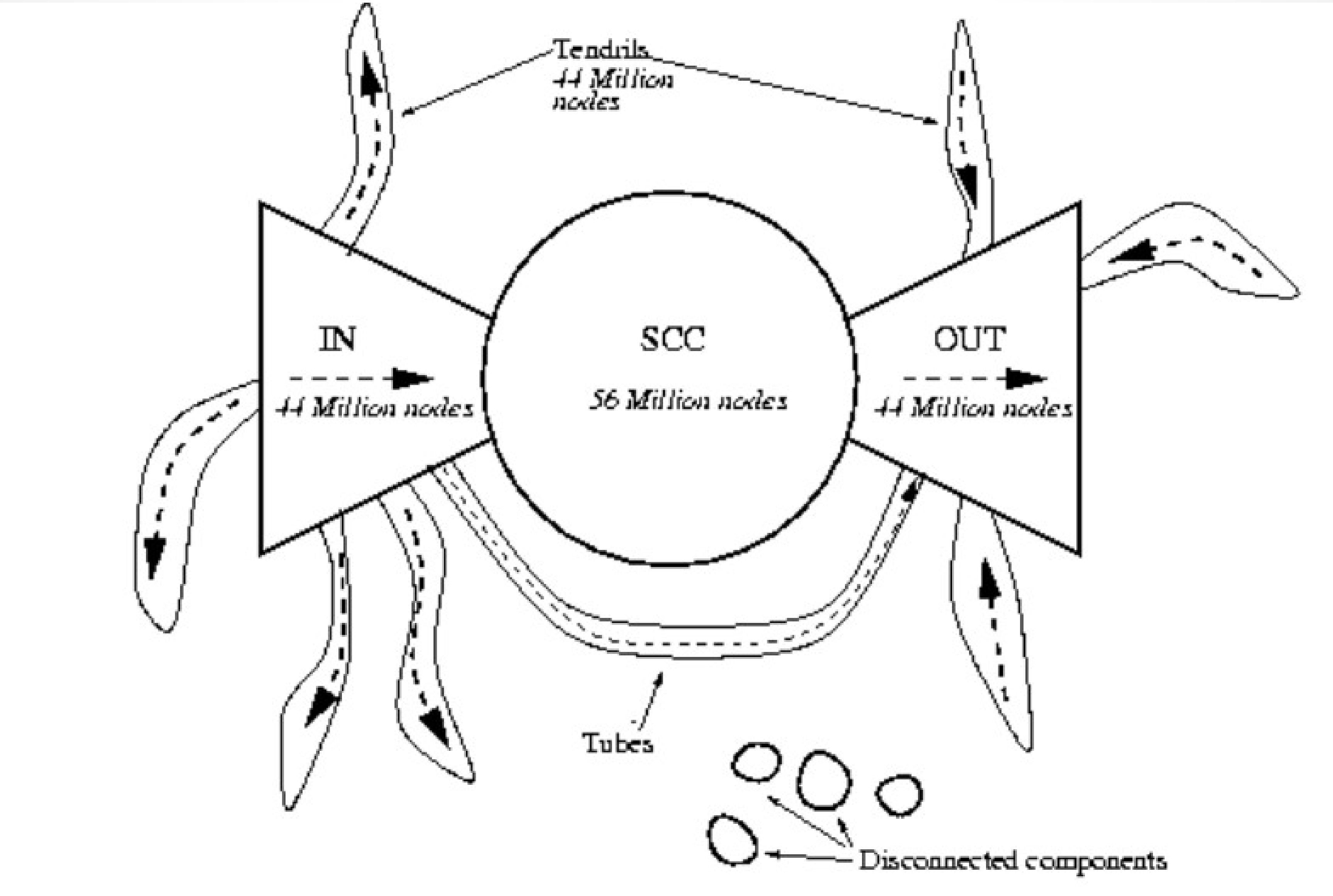
- Web is a directed graph
- SCC in a directed graph is a subset of the nodes such that:
  1. every node in the subset has a path to every other
  2. the subset is not part of some larger set with the property that every node can reach every other.
- Web contains a giant strongly connected component
  - Many naturally occurring undirected graphs have a giant connected component
  - Major "directory" sites -> home pages of major educational institutions, large companies, and governmental agencies -> pages in these sites ->Directory Pages
  - Thus, all these pages can mutually reach one another, and hence all belong to the same strongly connected component.
  - This SCC contains (at least) the home pages of many of the major commercial, governmental, and non-profit organizations in the world, it is easy to believe that it is a giant SCC.
  - If there were two giant SCCs —X & Y a single link from any node in X to any node Y and another link from any node in Y to any node in X, X and Y would merge

Structure of the Web

❑ [4]

- Crawl of Alta Vista Search Engine
- Only Navigational Pages
- The influential study has since been replicated on other, even larger snapshots of the Web:
  - An early index of Google's search engine
  - Large research collections of Web pages
- Similar analyses have been carried out for particular well-defined pieces of the Web:
  - The links among articles on Wikipedia
  - Complex directed graph structures arising in other domains, such as the network of interbank loans

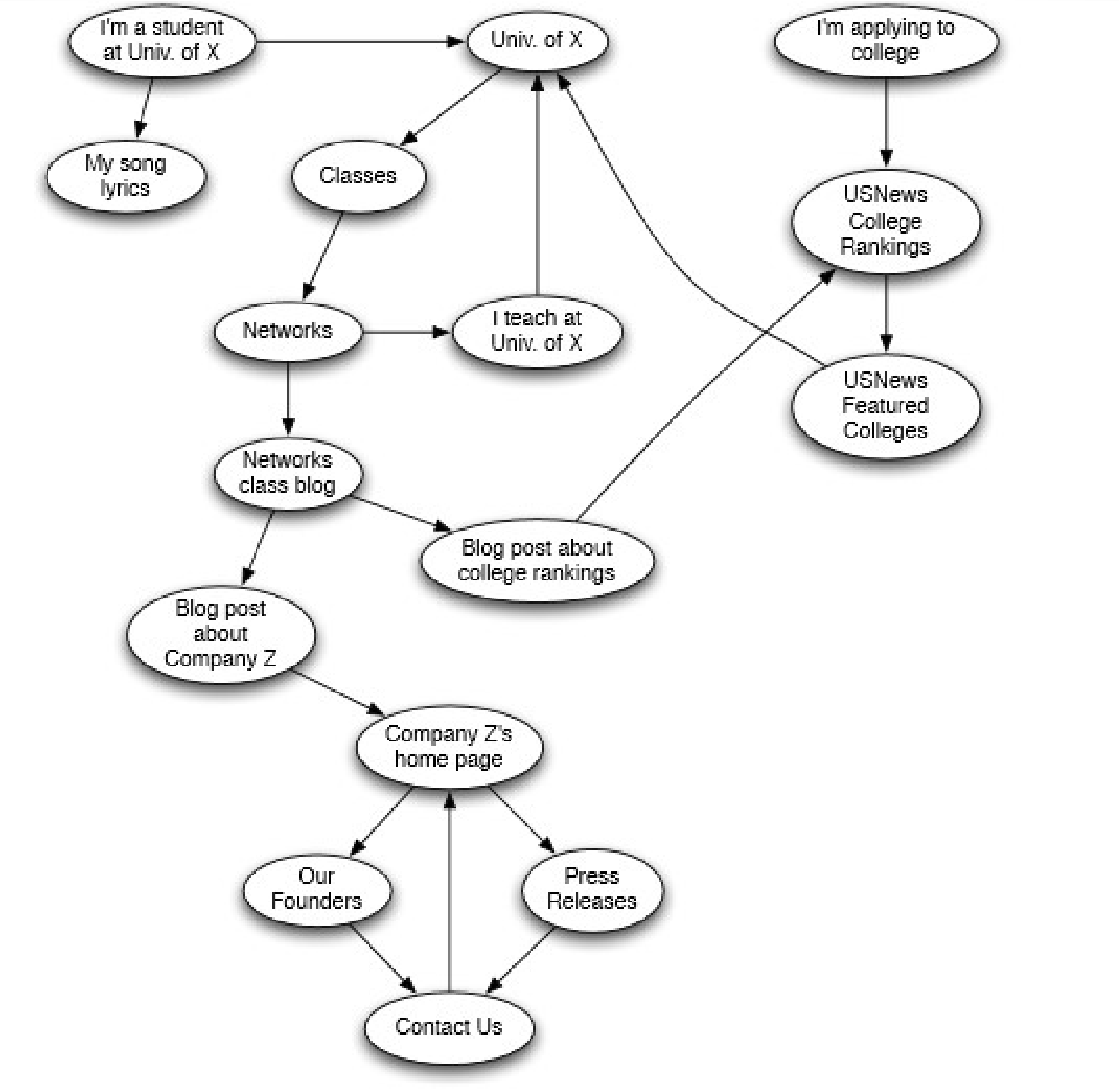
Bow-tie Structure of Web



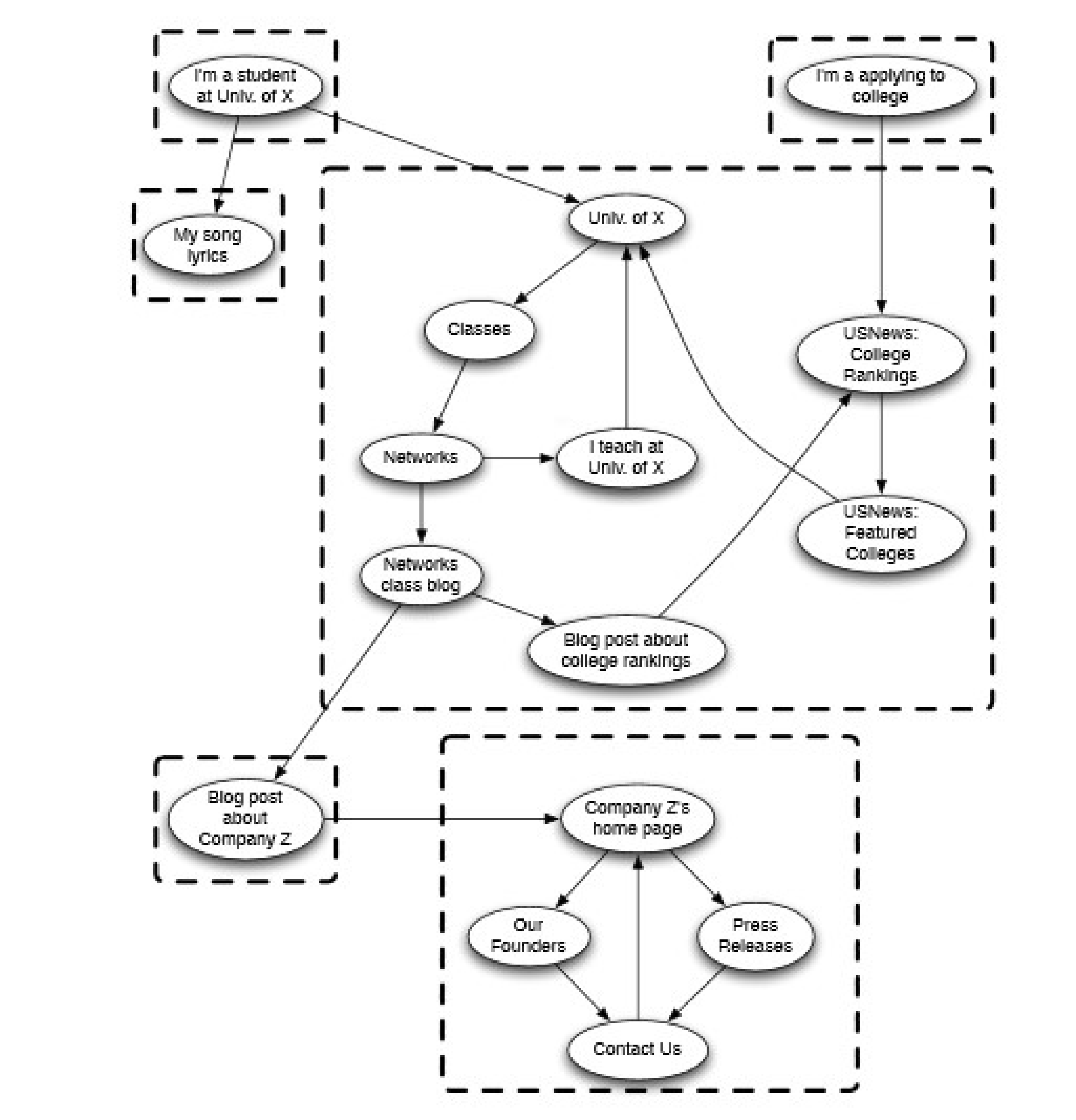
Components in the Bow-tie Structure

- IN: nodes that can reach the giant SCC but cannot be reached from it
- OUT: nodes that can be reached from the giant SCC but cannot reach it
- Tendrils: The "tendrils" of the bow-tie consist of
  a. The nodes reachable from IN that cannot reach the giant SCC,
  b. The nodes that can reach OUT but cannot be reached from the giant SCC. SCC
- Tube: A node that satisfy both (a) and (b) - travels from IN to OUT without touching the giant SCC
- Disconnected: Nodes that would not have a path to the giant SCC even if we completely ignored the directions of the edges
- Structure is dynamic

Exercise – Find Structure



SCCs


I'm a student at Univ. of X
I'm a applying to college
My song lyrics
Univ. of X
Classes
USNews: College Rankings
Networks
I teach at Univ. of X
USNews: Featured Colleges
Networks class blog
Blog post about college rankings
Blog post about Company Z
Company Z's home page
Our Founders
Press Releases
Contact Us

References

1. Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. The World-Wide Web. Communications of the ACM, 37(8):76-82, 1994.
2. Jakob Nielsen. The art of navigating through hypertext. Communications of the ACM, 33(3):296{310, 1990
3. Vannevar Bush. As we may think. Atlantic Monthly, 176(1):101{108, July 1945
4. Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. In Proc. 9th International World Wide Web Conference, pages 309–320, 2000

Reading

1. David Easley and Jon Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World.

   https://www.cs.cornell.edu/home/kleinber/networks-book/

   - Chapter 13

: