CORSO DI LAUREA
MAGISTRALE IN
INGEGNERIA INFORMATICA

SOCIAL NETWORKS ANALYSIS
A.A. 2021/22

INFORMATION NETWORKS AND THE STRUCTURE OF THE WEB

Information Networks

- In previous lectures we dealt with social networks
  - Nodes represent social entities interacting with other subjects and taking decisions
  - Links represent (opportunity of) social or economic relations between two nodes
- In this lecture we are interested into information networks
  - Networks where nodes are pieces of information (documents) that are interconnected
  - There is a link between two documents if and only if they are interconnected
- An example of an information network
  - World Wide Web (the main object of our study)

Social and Information Networks

- Information networks significantly differ from social networks
  - Nodes in a social network are autonomous agents taking their own decisions
  - Nodes in an information network are documents and they do not take decisions
  - Links in an information network are unidirectional
    - The is a link between X and Y if document X refers to document Y
    - It may be not the case that Y does not refer to X
- Information networks show several important characteristics in common with social networks
  - Giant component
  - A lot of short paths
  - Centrality measures
  - Connections with matching markets

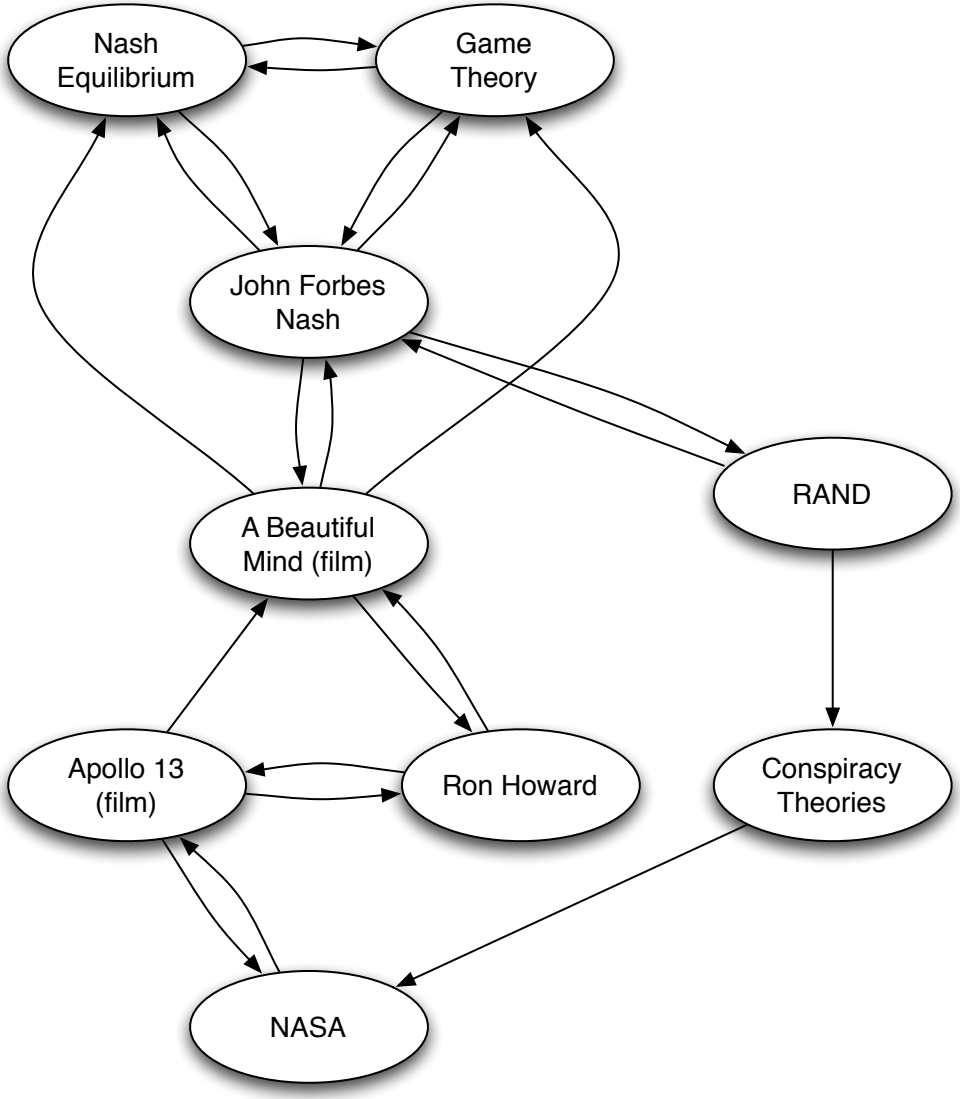
The World Wide Web

- The Web is the largest and most popular information network
  - Created by Tim Barners-Lee in 1989-1991 as a tool to make results of different research groups at CERN easily available to other groups
- The basic idea
  - Documents are represented in a standard textual format (HTML) and made available in a computer space publicly accessible (Web server)
  - A specialized application (browser) reads the pages, displays their contents and surf through them
- Nowadays the Web is a huge information system consisting of several billions of pages
  - Pages are published and maintained by different subjects that decide autonomously when to publish them and which pages to link

Hypertexts

- The Web is an hypertext
  - You can read the web pages in a non sequential order
  - You can move from a page to another through their links
    - Links are chosen by the page's author
- Hypertexts were introduced at half of the 20-th century to make texts more fruible
- Some pioneers of hypertexts
  - Citation networks and cross-references
    - Differ from hyperlinks since a document can only refer older documents
  - Wiki
  - Associative memories
  - Semantic networks

Reference networks



Vannevar Bush and the Memex

- Most of the characteristics of the Web as we know it were predicted and accurately described by Vannevar Bush in 1945
  - As we May Think – Atlantic Monthly

- In Bush vision in future computers the traditional linear memorization methods for information storing would have replaced by associative methods
  - More similar to how the human brain works

- Bush described a prototype of this new computer, called the Memex
  - Very similar to our Web
  - Contained digital representation of the whole human knowledge interconnected through associative links

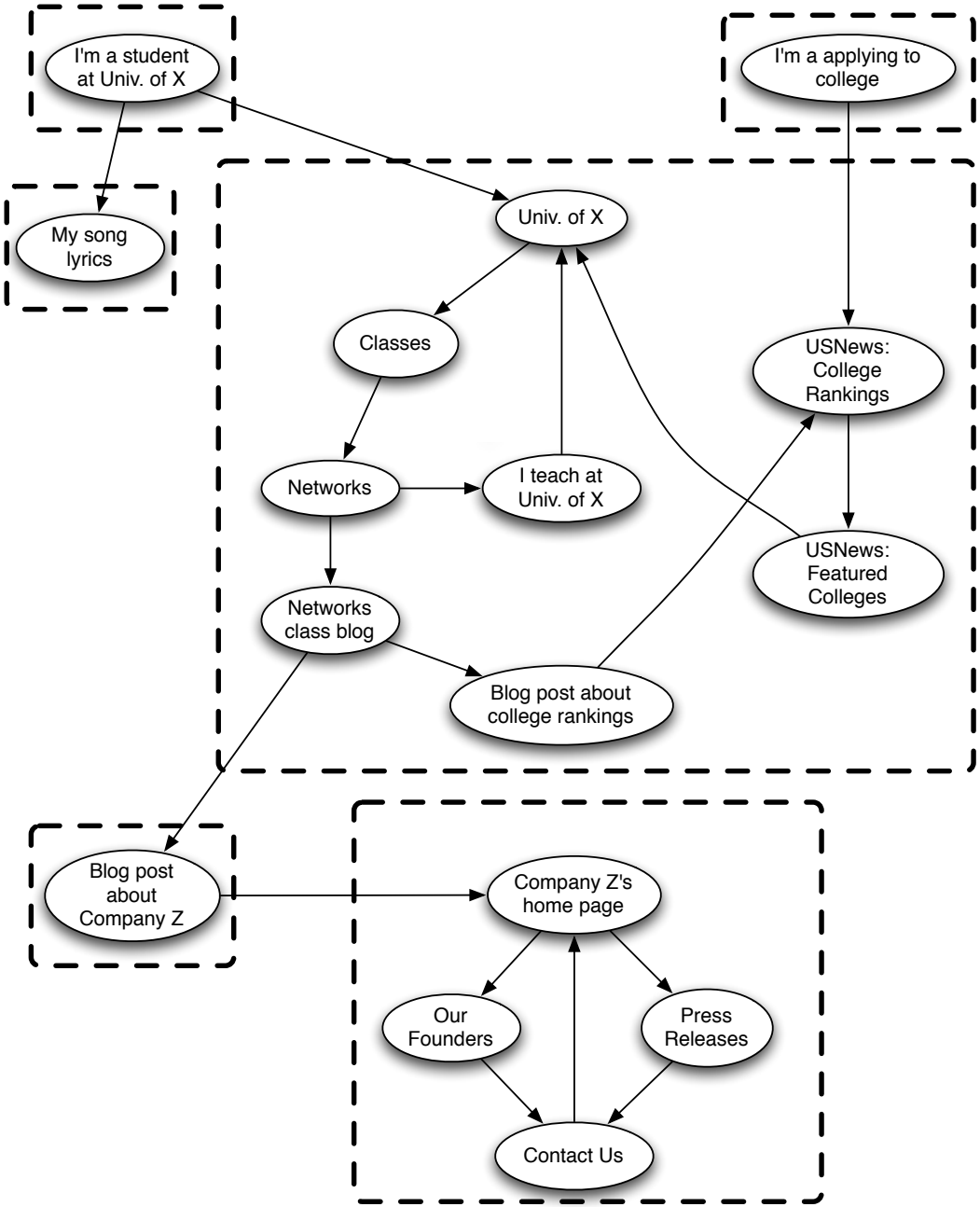
The Evolution of the Web

- In the last 25 years the Web evolved into a huge computational platform
- Nowadays there are two distinct families of pages
  - Static pages providing pieces of information
  - Dynamic pages representing the output of a program running on the server and called through the network
- Similarly, links can be of two types
  - Navigational links
    - Let the users navigate through static pages and reach dynamic pages
    - They represent the real structure of the Web
  - Transactional links
    - Let the users call a remote application
- Search engines take into account only pages reachable through navigational links

The Web as a Directed Graph

- The Web and all the information networks can be represented as directed graphs
  - Consider only navigational links
- The graph of the Web is
  - Directed
  - Not connected
- We can apply to the Web all the graph theory concepts
  - paths, cycles, connectivity, connected components
- A strongly connected component S of a directed graph is a subset of nodes such that
  - Every node in S is connected to all the other nodes in the set
  - This subset is maximal

An Example


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

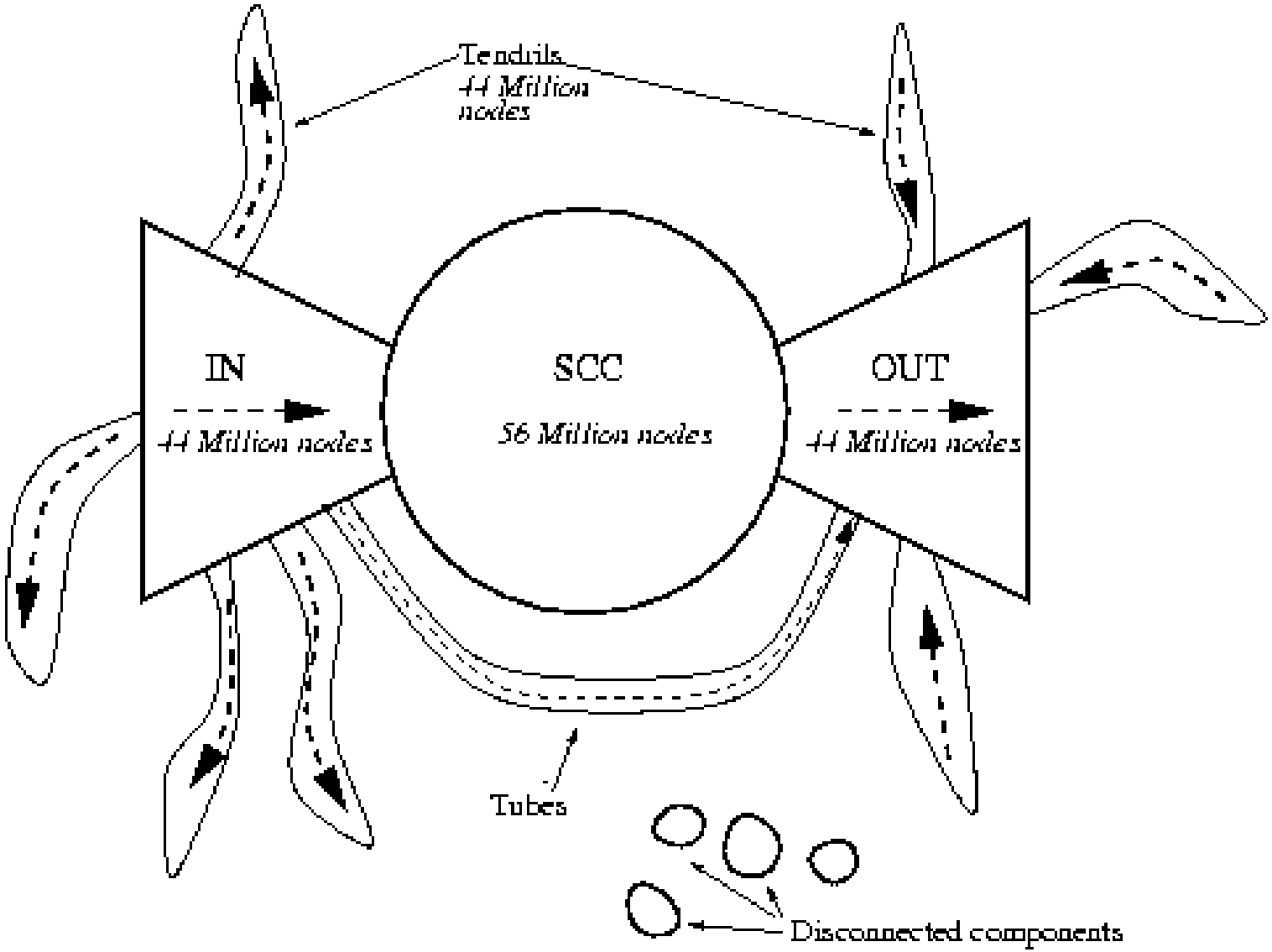
The structure of the Web

- In 1999 Broder et al. created a map of the Web at that time
  - Based on data memorized in the index of the Altavista search engine
- The Broder experiment was replicated several times, both on the Web an on other information networks
  - Google
  - Wikipedia
  - Citation networks of scientific papers
- All the experiments showed these networks have similar characteristics
  - Giant component

A Classification of the Components of the Web

- The Broder group classified the strongly connected components of the Web in three classes
  - CORE: a giant component where each page can reach any othr page through navigational paths
  - IN: pages that can reach the CORE but cannot be reached from it
  - OUT: pages that can be reached from the CORE but cannot reach pages in it
  - TENDRILS: pages that can be reached by IN or reach OUT
  - ISOLATED pages: everything else
- When new links are created pages in IN and OUT can move to the CORE and replaced by new just created pages

The Bow-Tie Structure



The Emergence of the Web 2.0

- The increasing richness of Web content observed in the last 10 years fueled a series of further significant changes in the Web
- Three major forces behind these changes
  - the growth of Web authoring styles that enabled many people to collectively create and maintain shared contents (wikipedia)
  - 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 (google services, cloud)
  - the growth of linking styles that emphasize on-line connections between people, not just between documents (social networks)
- Tim 'O Reilly described all these transformations as the emergence of the Web 2.0
  - "an attitude, not a technology"
- Designers of Web sites today need to think not just about organizing information, but also about the social feedback effects
  - users who are able to interact directly with one another
  - Wisdom of crowds, long tail, cascading behaviours
- The social-networking aspects of Web 2.0 sites provide rich data for large-studies of social network structure

Searching in the Web

- The Web is a huge information network
  - Its utility depends on how efficiently we can search for information in it
  - The search engine is a crucial element of the Web
- Since the first days of the Web people studied ways to organize it and make the search more efficient

First Approaches to the Search Problem

- Web Directories
  - Yahoo, DMOZ
  - Index pages Pages supervised by an editorial staff providing links to relevant pages for each specific subject

- Information Retrieval
  - Techniques developed since 60's to search in archives of structured documents through keywords
    - Newspapers' articles, scientific articles, patents, verdicts, laws, ecc.
  - Keywords are a very limited way to express a complex information and they suffer from the problems of synonymy and polysemy
  - Till 80s there were people specialized in bibliographic searches

Searching in the Web

- Information retrieval techniques revealed to be ineffective for searching in the Web
  - Do not scale to the dimensions of the Web
  - Web pages have a limited structure and are written with very different authoring styles
  - rich diversity in the set of people issuing queries
    - the problem of multiple meanings becomes particularly severe
  - dynamic and constantly-changing nature of Web content
- The main problem when searching in the Web
  - Is not to find a few elements in a large dataset …
  - But to filter the huge quantity of information in the network and select the most relevant ones
- We need good techniques to classify web pages (**ranking**) with respect to a subject

Ranking Web Pages

- Modern search engines can efficiently search in the Web and index thousands of pages satisfying a query
- The average user checks only the first 5-10 pages presented by the search engine
  - If he cannot find what he's looking for, he presents a new query or move to another search engine
- For a search engine it's crucial to rank pages with respect to their "**relevance**" with respect to a specific query
  - The first page returned by the search engine contains the most relevant pages
  - It has to maximize the probability that users can find what they are looking for

Ranking and Centrality

- Centrality measures describe how a node in the graph is important with respect to a given property

- To rank web pages we need a centrality measure that measures pages' relevance with respect to a query

When is a Page Relevant?

- The relevance of a page can be evaluated by a human operator by reading its content
  - Not scalable solution
- We need automatic mechanisms to evaluate the relevance of a page with respect to a query
- When is a page relevant?
  - If it linked by other relevant pages

Link Analysis

- We can consider graph links as votes on the relevance of pages
  - If I link a page I'm implicitly voting for its relevance
- Analyzing links in the graph we can extract information about page relevance
- Not all the links are relevant
  - I can link pages that I want to criticize
  - I can link advertisements
  - I can link pages that are out of contest
- But most of them are relevant
- However, we assume that the relevance of a page is proportional to the number of incoming links

Link Analysis Algorithms

- The more used link analysis algorithms are
  - HITS (Hubs and Authorities)
  - PageRank
  - Different versions of Pagerank customized on a particular subject
- We will discuss them in details in the next lecture