CS4423 - Networks

Prof. Götz Pfeiffer
School of Mathematics, Statistics and Applied Mathematics
NUI Galway

6. Power Laws and Scale-Free Graphs

Lecture 17: The Structure of the World Wide Web

Happy 30th Birthday, WWW (https://webfoundation.org/2019/03/12/web-birthday-30/)!

So far, the networks that have been discussed most of the time consisted of people or organizations, connected by links representing opportunities for interactions. The World Wide Web is an example of a network of a different kind, a so-called **information network**.

In [1]:
```
import numpy as np
import pandas as pd
import networkx as nx
import matplotlib.pyplot as plt
```

Information Networks

Information networks connect pieces of information, like documents, or parts of documents, through links that represent references of some kind. Such links, in contrast to social relationships which are typically symmetric, only point in one direction. The underlying graph of an information network thus is a **directed graph**.

Information networks have existed before the internet. Some prominent examples include:

- **Academic Publications.** In the scientific literature it is customary to give credit to sources that have been used in the form of references to those publications that contain those sources. This practice creates a network whose nodes are the publications, and whose links represent the references, pointing from the citing article back to the cited article. A large part of this network in the mathematical literature is captured on MathSciNet (http://www.ams.org/mathscinet).
- **Mathematical Proofs.** In mathematics, the proof of a particular theorem usually relies on theorems that have already been proved. Citing a theorem in a proof thus creates a link from the theorem being proved back to the theorem being used, in a network of mathematical theorems. In a similar way, a complex computer program, consisting of several subroutines, can be regarded as a network of subroutines, pointing to each other through links that arise from one subroutine calling another.
- **Technical Documentation.** The documentation of complex systems, such as computer software, typically consists of a collection of articles (manual pages), each describing one aspect of the system, frequently using cross-references to each other. Here the network consists of the manual pages, and the links represent those cross references. In a similar way, an encyclopedia (or a dictionary) organizes its content as a sequence of articles, sorted alphabetically, with supporting cross-references.

Hypertext

The **World Wide Web** arose out of the desire to make technical documentation more easily accessible by using the physical infrastructure of the rapidly growing internet. It was conceived by Tim Berners-Lee (https://en.wikipedia.org/wiki/Tim_Berners-Lee) around 1990 as information management system at CERN (http://info.cern.ch/hypertext/WWW/TheProject.html).

In this system, documents are **web pages**, that anyone can create and store in a publicly accessible place on their computer. Moreover, it supplies a **web browser**, a piece of software that can retrieve the web pages from those public spaces, allowing others to easily access those documents.

Web pages contain **hypertext**, that is a mixture of plain text and **hyperlinks**. Here, a hyperlink (or just link) is a reference to another document that the reader can follow by clicking. Hyperlinks have a **source** (the document they are contained in) and a **target** (the document they reference). This creates a network of documents as nodes and hyperlinks as **directed edges** between them.

There are many alternative ways to organize information: alphabetically (like the telephone book), hierarchical (in folders like the files on a computer), ... Certainly, the physical constraints of the environment (like the fact that books need to be stored on shelves, that pages in a book come in order) have an influence on how well a particular solution works.

Hypertext makes logical relationships within and between texts (which traditionally are only implicit) explicit first class objects. The browser makes links immediately actionable.

Hypertext originates from the works of the visionaries Vannevar Bush (https://en.wikipedia.org/wiki/Vannevar_Bush) (the Memex, 1945) and Ted Nelson (https://en.wikipedia.org/wiki/Vannevar_Bush) (Xanadu Project, 1965).

It will be useful to distinguish between **navigational links** (providing access to related pages) and **transactional links** (which exist more for a side effect, like ordering a book, or sending an email, than for the sake of leading to the next page). The distinction is not always clear, but transactional links are the kind that is of little interest for search engines. It's the navigational links that form the edges of the directed graph that turns the Web into an information network.

As with undirected graphs, an interesting question in directed graphs is: which nodes can be reached from a given node?

Reachability in Directed Graphs

Recall that a **directed graph** is a pair $G = (X, E)$ with **vertex set** X and **edge set** $E \subseteq X^2 = X \times X$. For an edge $(x, y) \in E$ we sometimes write $x \to y$.

A **path** in a directed graph $G = (X, E)$ is a sequence of nodes $(x_0, x_1, \ldots, x_l)$ with $x_{i-1} \to x_i$ for $i = 1, \ldots, l$. The number l is called the **length** of the path. We write $x \rightsquigarrow y$ if there exists a path (possibly of length 0) from x to y in G.

A directed graph G is **weakly connected** if, when considerd as undirected graph, it is connected. The **weakly connnected components** (WCCs) of G are its connected components, when considered as undirected graph.

A directed graph G is **strongly connected** if, for each pair of vertices $x, y \in X$, there is a path from x to y in G, i.e., if $x \rightsquigarrow y$.

A **strongly connected component (SCC)** of a directed graph G is a subset C of X which is (i) strongly connected, and (ii) not part of a larger strongly connected subset of X.

In general, a directed graph is a collection of WCCs. Each WCC in turn is a SCC.

When a directed graph G is regarded as a **relation** on the set X, strongly connected components can be described as the **equivalence classes** of an equivalence relation that is obtained as follows.

First note that the relation $x \rightsquigarrow y$ is the reflexive and transitive closure of the edge relation $x \to y$. Thus, by construction it is reflexive and transitive. It might not be anti-symmetric, though, meaning that there might be vertices x and y with $x \rightsquigarrow y$ and $y \rightsquigarrow x$.

However, the new relation $x \equiv y$, defined as $x \rightsquigarrow y$ and $y \rightsquigarrow x$ is an equivalence relation (why?) and its equivalence classes are the strongly connected components of G. Denote the class of $x \in X$ by $[x]$.

Moreover, there is a partial order relation $\leq$ (a relation which is reflexive, transitive and anti-symmetric) on the set of equivalence classes, $[x] \leq [y]$ if $x \rightsquigarrow y$.

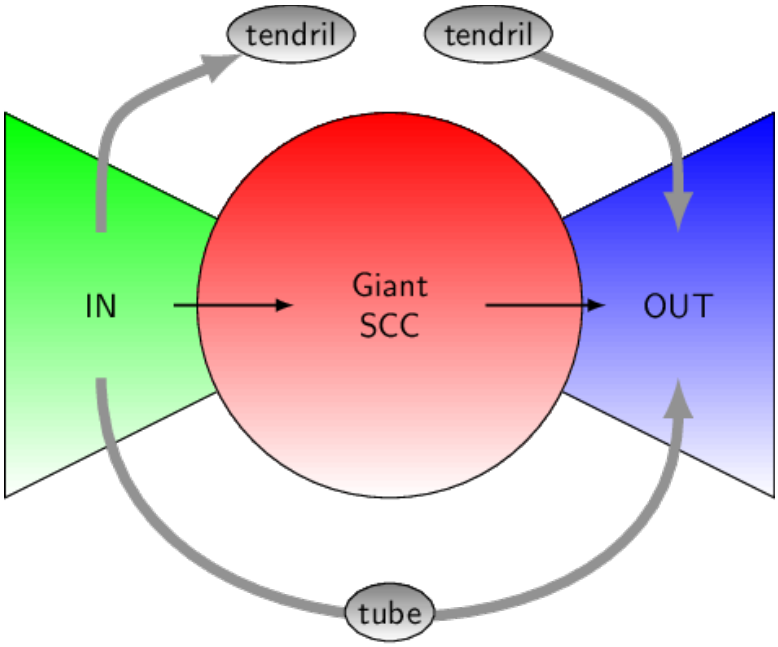
The Bow-Tie Structure of the WWW

Like the giant component in a simple graph, a directed graph with sufficiently many edges has a giant SCC.

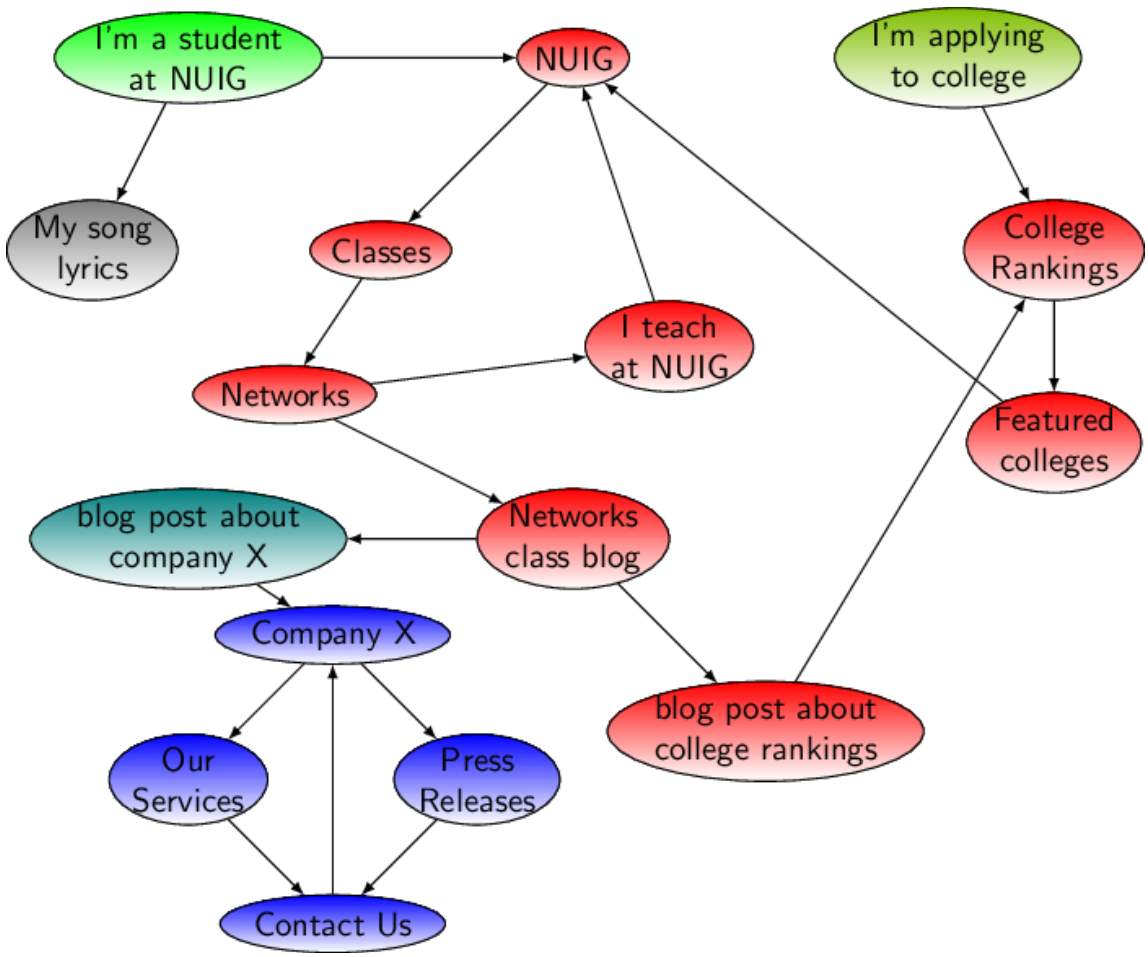
The remainder of the graph consists of four more sets of components of nodes.

1. IN: upstream components, the set of all components C with $C <$ SCC.
2. OUT: downstream components, the set of all components C with $C >$ SCC.
3. tendrils: the set of all components C with either $C >$ IN and $C \not<$ OUT or $C <$ OUT and $C \not>$ IN;
   and tubes: components C with $C >$ IN, $C <$ OUT but $C \not<$ SCC.
4. disconnected components.

Thus, in any directed graph with a distinguished SCC, the containing WCC necessarily has the following global bow-tie structure:


tendril
tendril
IN
Giant
SCC
OUT
tube

For example,


I'm a student at NUIG
NUIG
I'm applying to college
My song lyrics
Classes
College Rankings
I teach at NUIG
Networks
Featured colleges
blog post about company X
Networks class blog
Company X
blog post about college rankings
Our Services
Press Releases
Contact Us

Research on available data on the Web in 1999 has confirmed this bow tie structure for the World Wide Web, with a large Giant SCC covering less than $\frac{1}{3}$ of the vertex set, and the three parts IN, OUT and the tendrils and tubes roughly of the same size. One can assume that this proportion has not changed much over time, although the advent of social media has brought many new types of links and documents to the Web.

In [ ]: