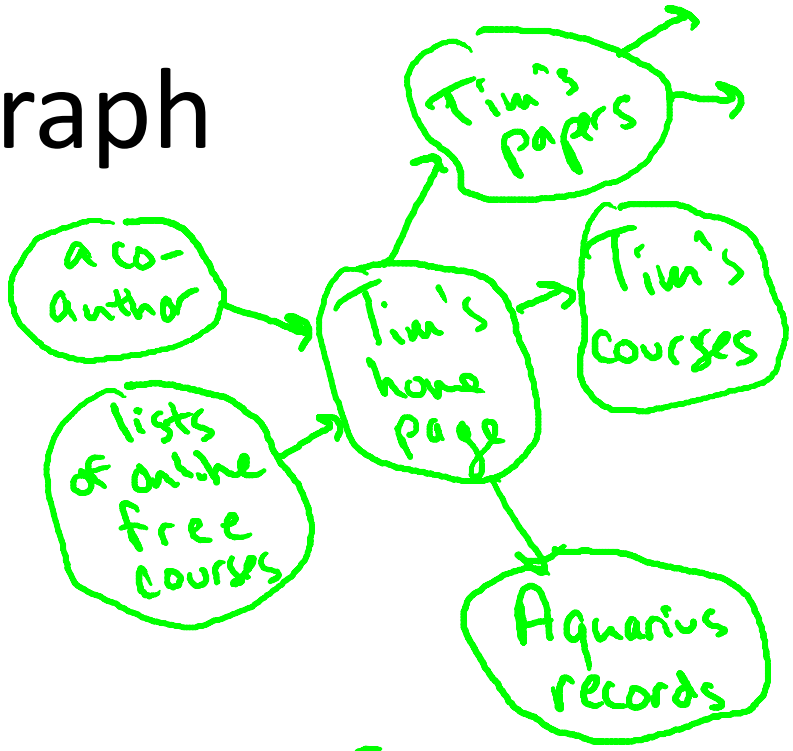
Graph Primitives

Structure of the Web

Design and Analysis
of Algorithms I

The Web graph

- vertices = Web pages
- (directed) edges = hyperlinks



Question: what does the Web graph look like?

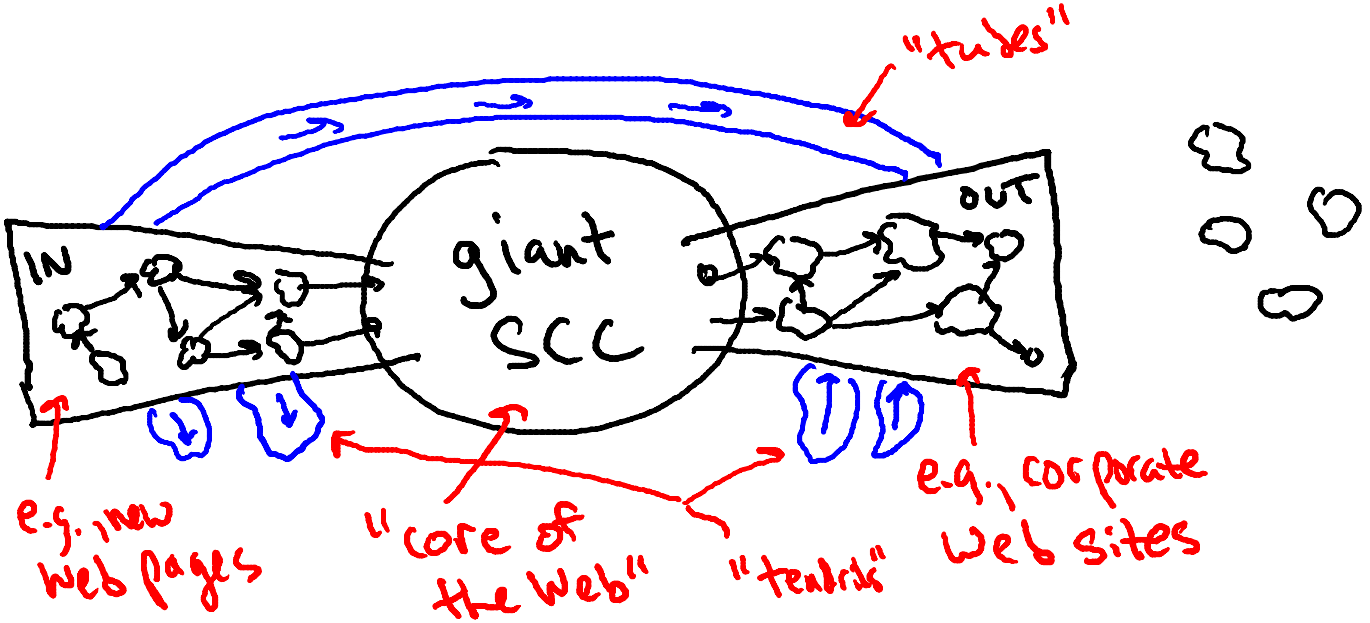
(assume you've already "crawled" it)

Size: ≈ 200 million nodes, ≈ 1 billion edges

Reference: [Broder et al WWW 2000]

computed the SCCs of the Web graph. (pre-Map-Reduce/Hadoop)

The Bow Tie


"tubes"
IN
giant SCC
OUT
e.g., new Web pages
"core of the Web"
"tendrils"
e.g., corporate Web sites

Main Findings

1. all 4 parts (giant, IN, OUT, tubes+tendrils) have roughly the same size
2. within CORE, very well connected (has the "small world" property) [Milgram]
3. outside, surprisingly poorly connected

Modern Web Research

(1) temporal aspects – how is the Web graph evolving over time?

(2) informational aspects – how does new information propagate throughout the Web (or blogosphere, or Twitter, etc.)

(3) finer-grained structure – how to define and compute "communities" in information and social networks?

Recommended Reading: Easley + Kleinberg, "Networks, Crowds, + Markets".