DS501: Graph Data

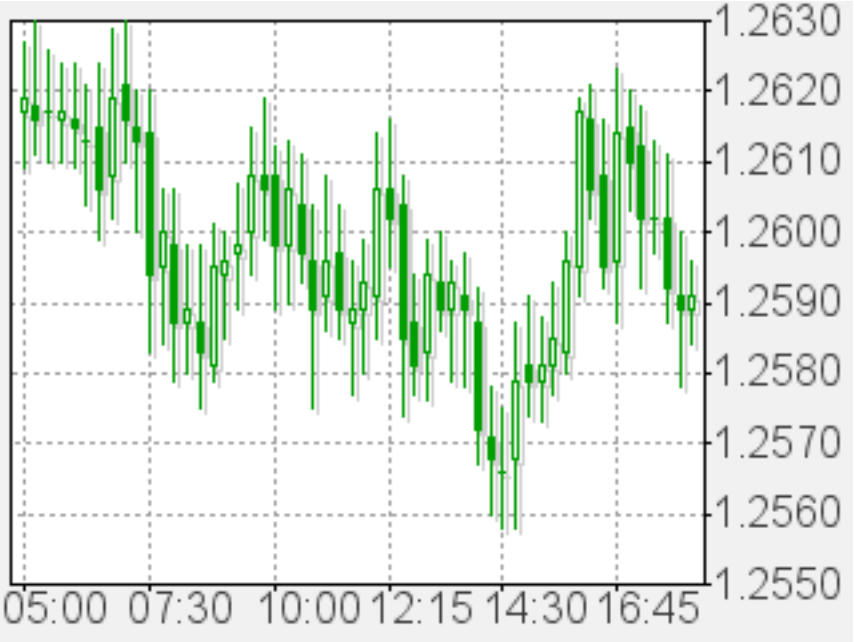
Prof. Randy Paffenroth
rcpaffenroth@wpi.edu

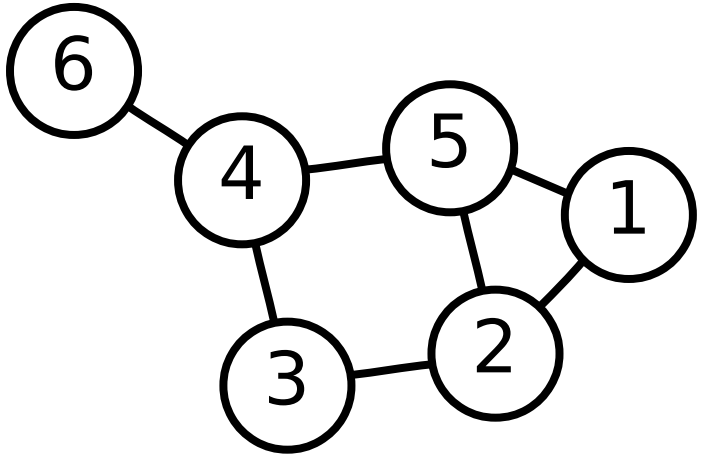
Worcester Polytechnic Institute

Announcements

- Case Study 3 due today!

Graph or Graph?



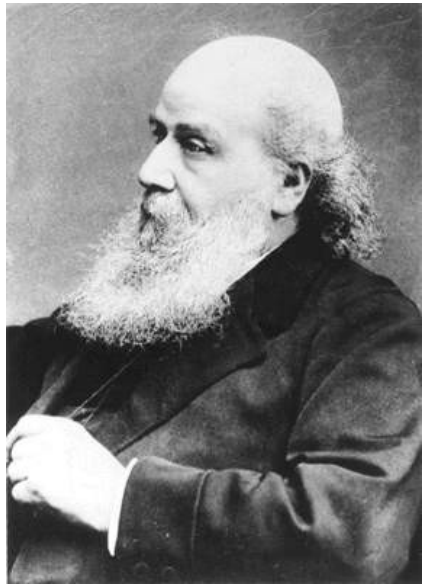


History

- The paper written by Leonhard Euler on the Seven Bridges of Königsberg and published in 1736 is regarded as the first paper in the history of graph theory.
- The term "graph" was introduced by Sylvester in a paper published in 1878 in Nature, where he draws an analogy between invariants in algebra and molecular diagrams.

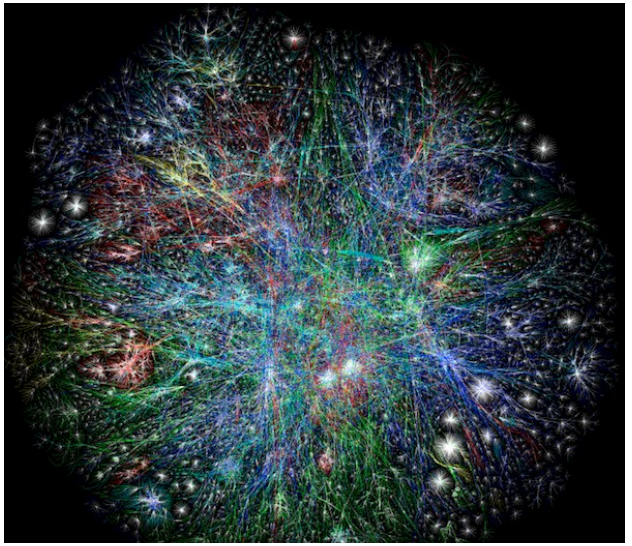
https://en.wikipedia.org/wiki/Graph_theory

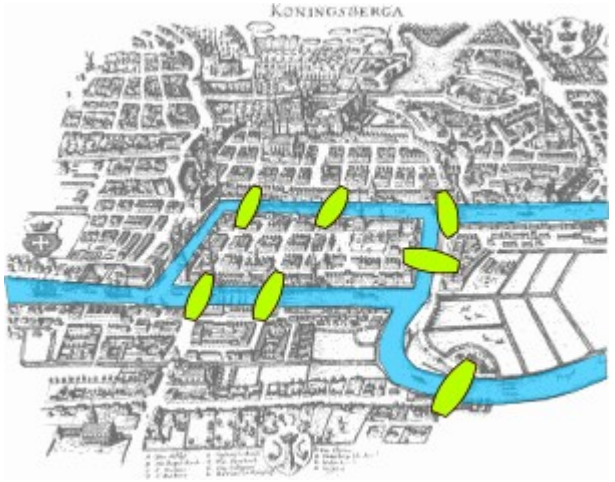
Leonhard Euler



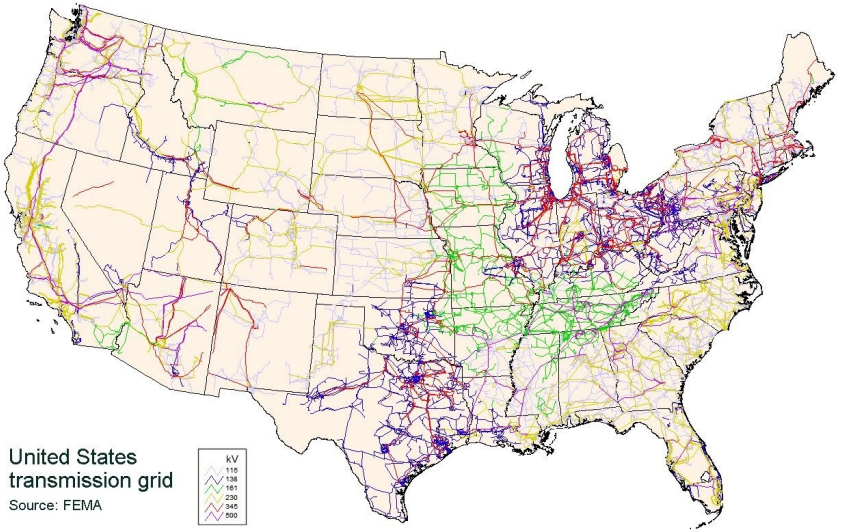
James Joseph Sylvester

Graphs Everywhere





By Bogdan Giuşcă - Public domain (PD), based on the image, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=112920



What graphs can you think of?

Road network

Social network

Flight Paths

neural network

authors For Papers

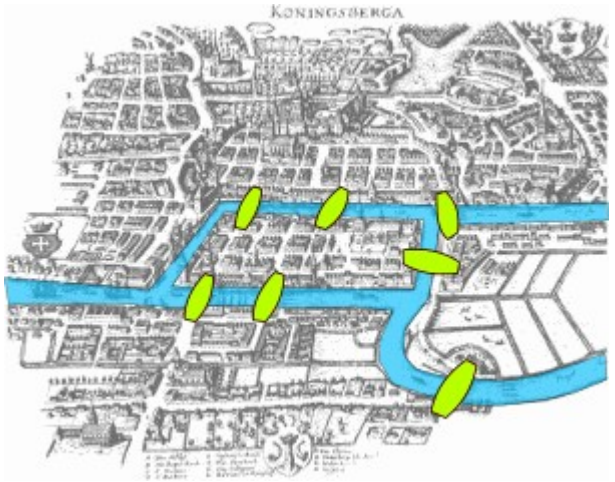
(Loss) and students

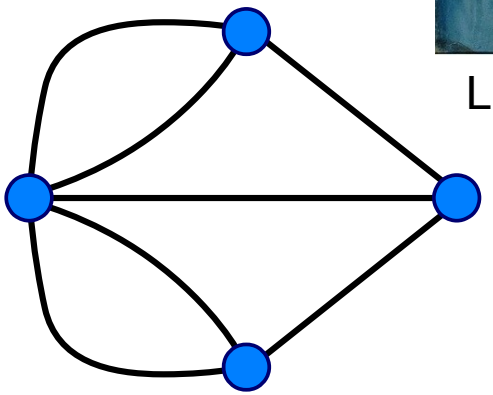
twitter

That seems simple, so why all the interest?

- Graphs lead to many deep and beautiful mathematical concepts!

Bridges of Konigsberg

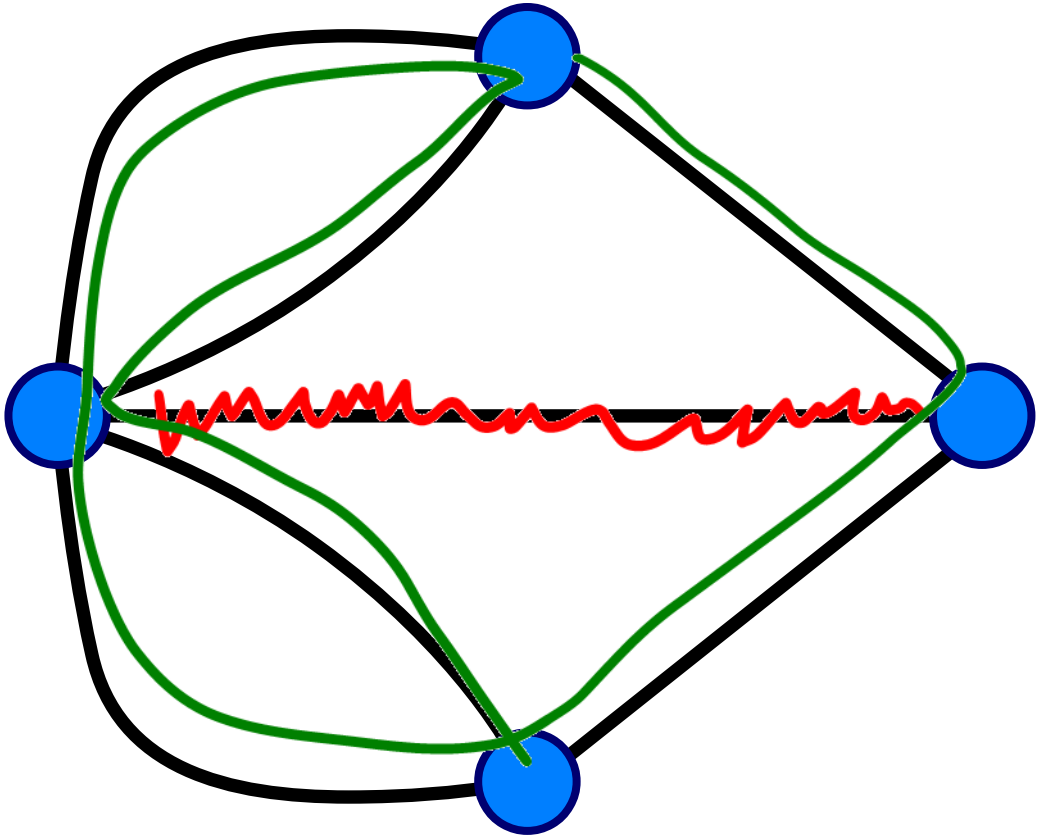




Leonhard Euler

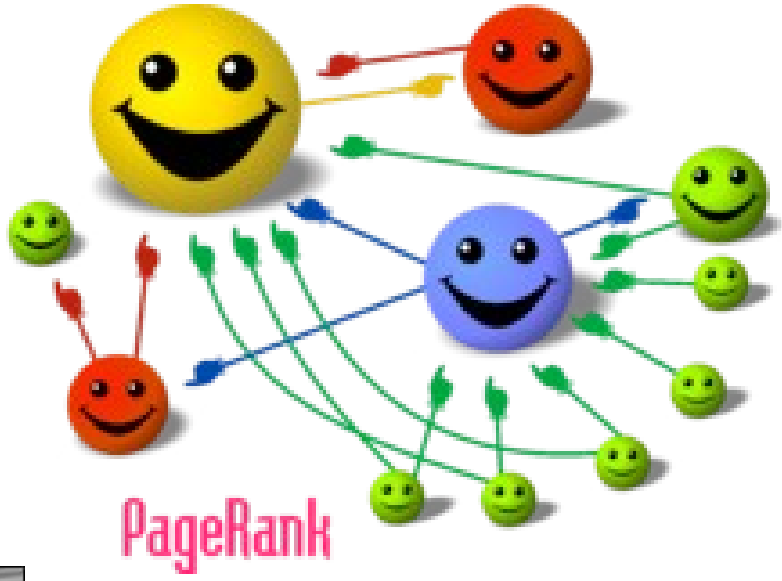
Example

- Can you cross every bridge exactly once?



1997

The desire to *automate* search



Larry Page and Sergey Brin

The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

1 Introduction and Motivation

The World Wide Web creates many new challenges for information retrieval. It is very large and heterogeneous. Current estimates are that there are over 150 million web pages with a doubling life of less than one year. More importantly, the web pages are extremely diverse, ranging from "What is Joe having for lunch today?" to journals about information retrieval. In addition to these major challenges, search engines on the Web must also contend with inexperienced users and pages engineered to manipulate search engine ranking functions.

However, unlike "flat" document collections, the World Wide Web is hypertext and provides considerable auxiliary information on top of the text of the web pages, such as link structure and

The PageRank Paper

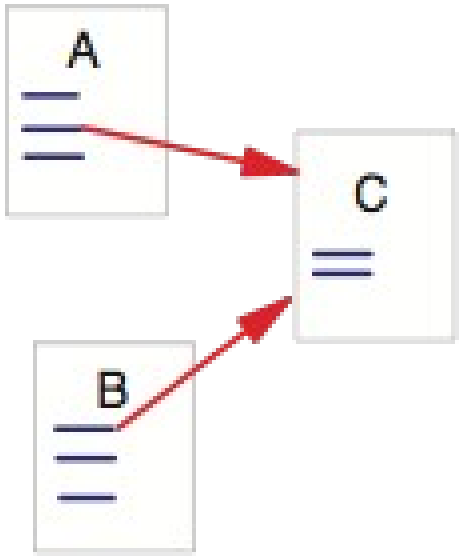


Figure 1: A and B are Backlinks of C

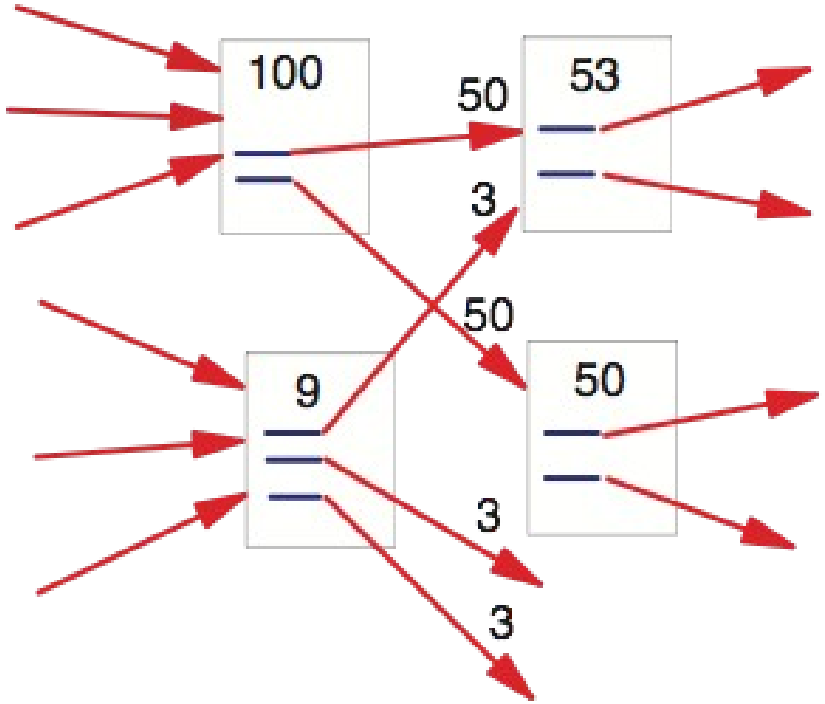


Figure 2: Simplified PageRank Calculation

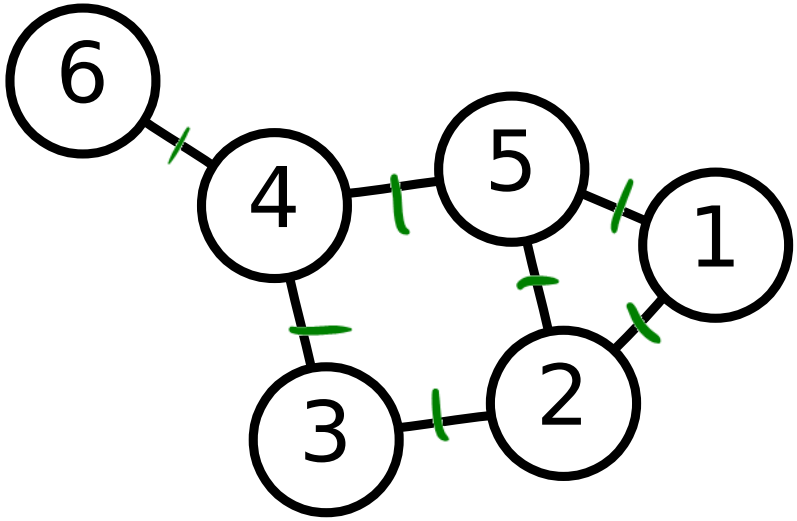
Graph Representation

- Mathematical object consisting of
  - V : A set of nodes/vertices/point
  - E : A set of connection/links between pairs of nodes
- G(V, E)

$V = \{A, B, C\}$

$E = \{\{A, B\}, \{A, C\} \ldots\}$

Example



V = {1, 2, 3, 4, 5, 6}

E = { {1,2}, {1,5}
{2,5}, {2,3}
{5,6}, {3,4}
{4,6} }

Types of graphs

- Edge types:
  - Undirected
    - Example?
  - Directed
    - Example?
  - Loops

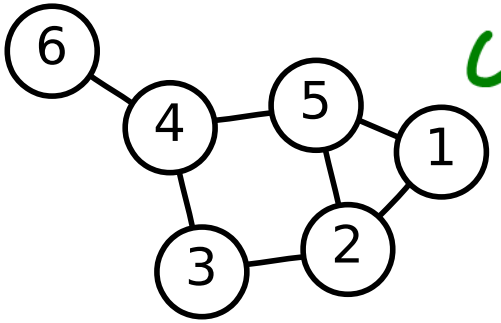
Distances between cities

Friendship :-
citations flights

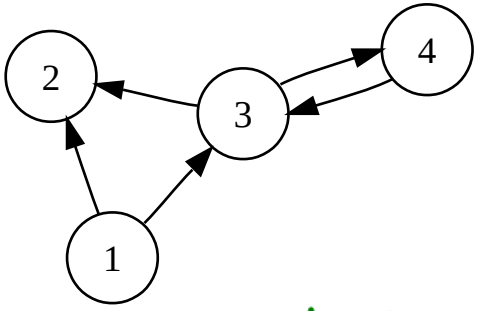
liking, web page, Boss

Types of graphs

- Graph types
  - Undirected
  - Directed
  - Mixed
  - Multi-graphs

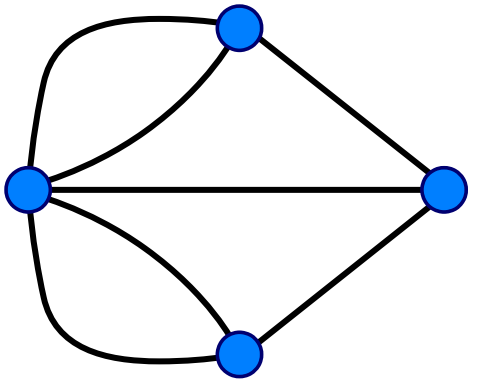


Undirected



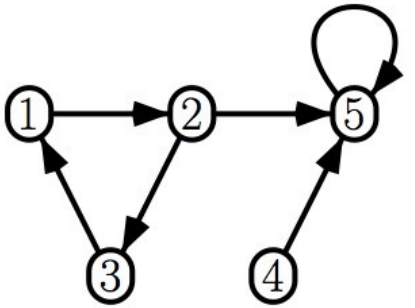
Directed Graph

multi graph

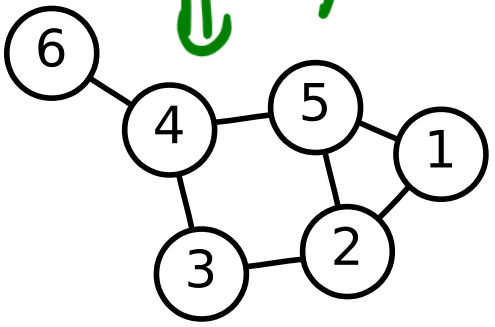


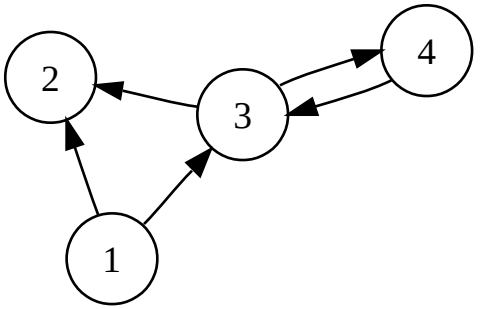
Types of graphs

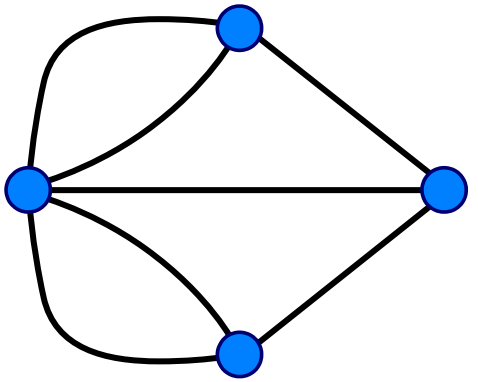
- Simple graphs
  - Undirected
  - No loops or multiple edges



simple graph

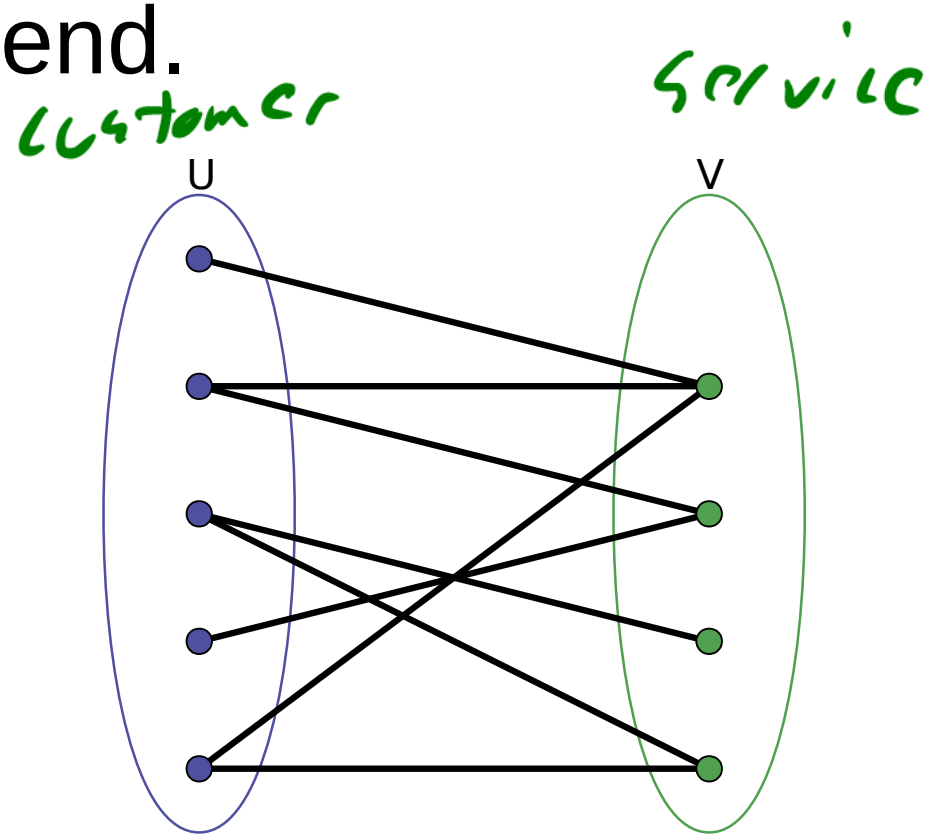






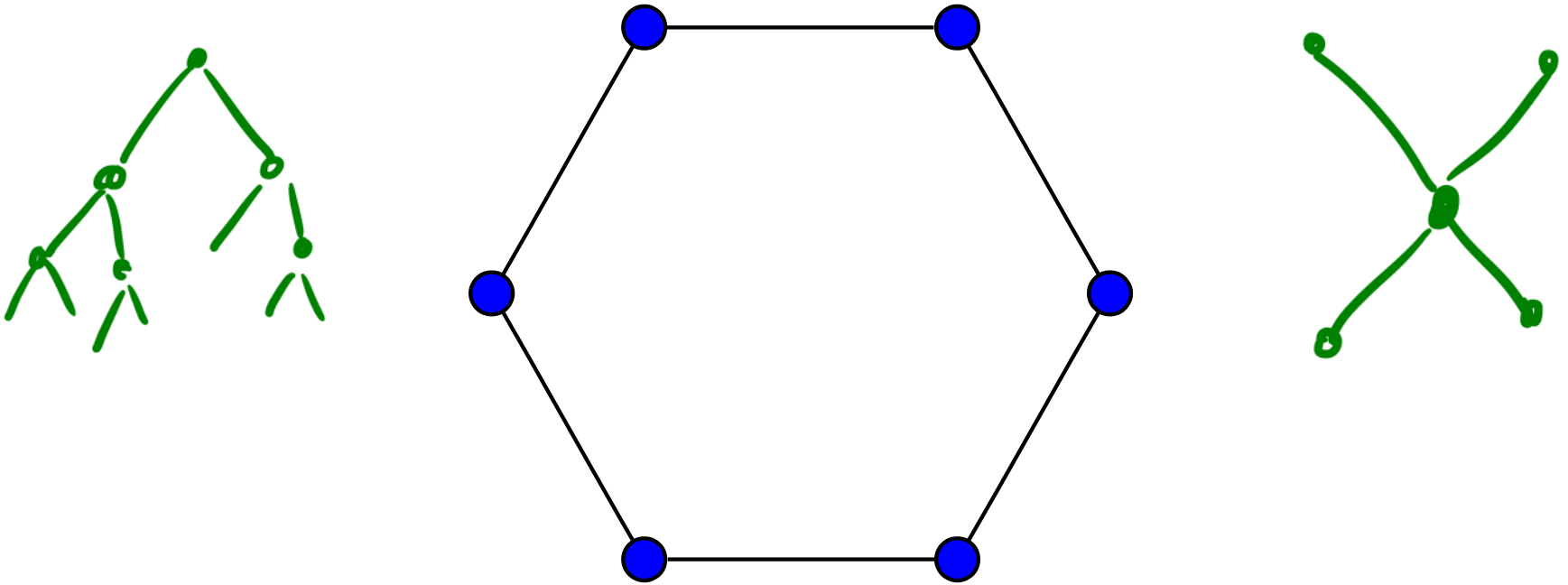
Special graphs: Bi-partite

- Every edge has one “blue” end and one ”green” end.



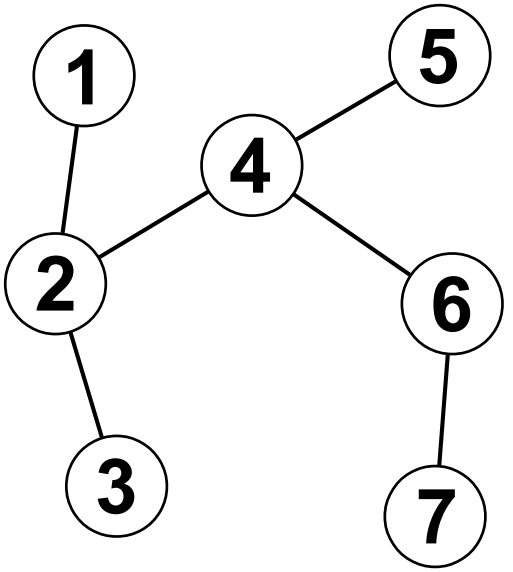
Special graphs: Tree

- Simple cycles
  - Some number of vertices connected in in a closed chain.



Special graphs: Tree

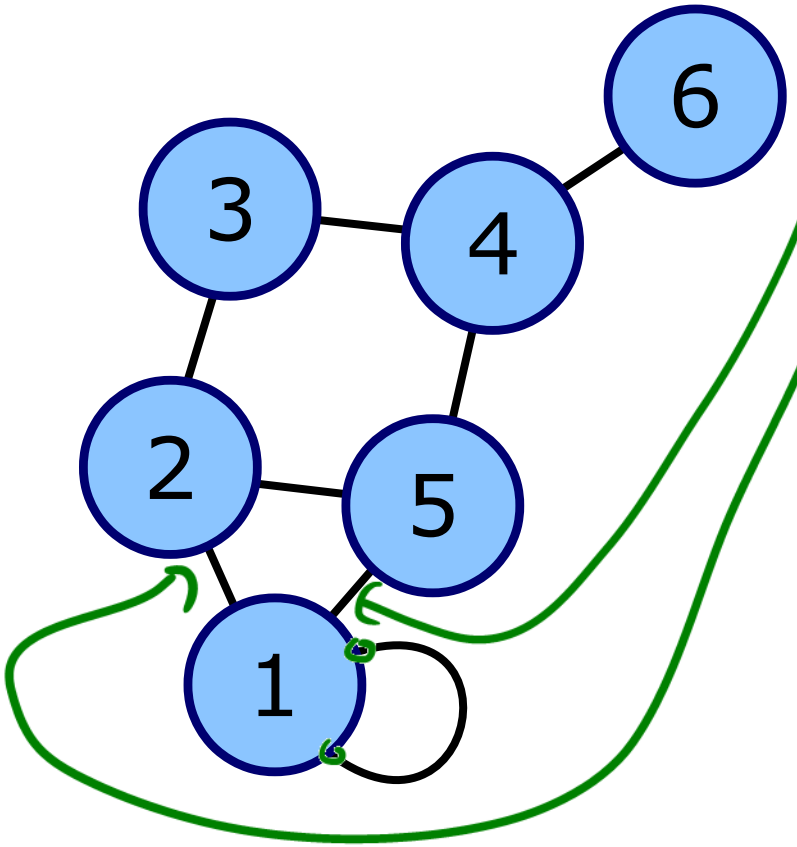
- No cycles



From before: Federal budget

http://www.brightpointinc.com/interactive/budget/index.html?source=d3js

Two views of a graph



for undirected, symmetric

Adjacency matrix

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 2 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

“Large Graph”

- Of course, in Data Science we are interested in “Large Graphs”
  - What makes a graph large?
  - Can you give some examples?

Brain network, web,
Road network

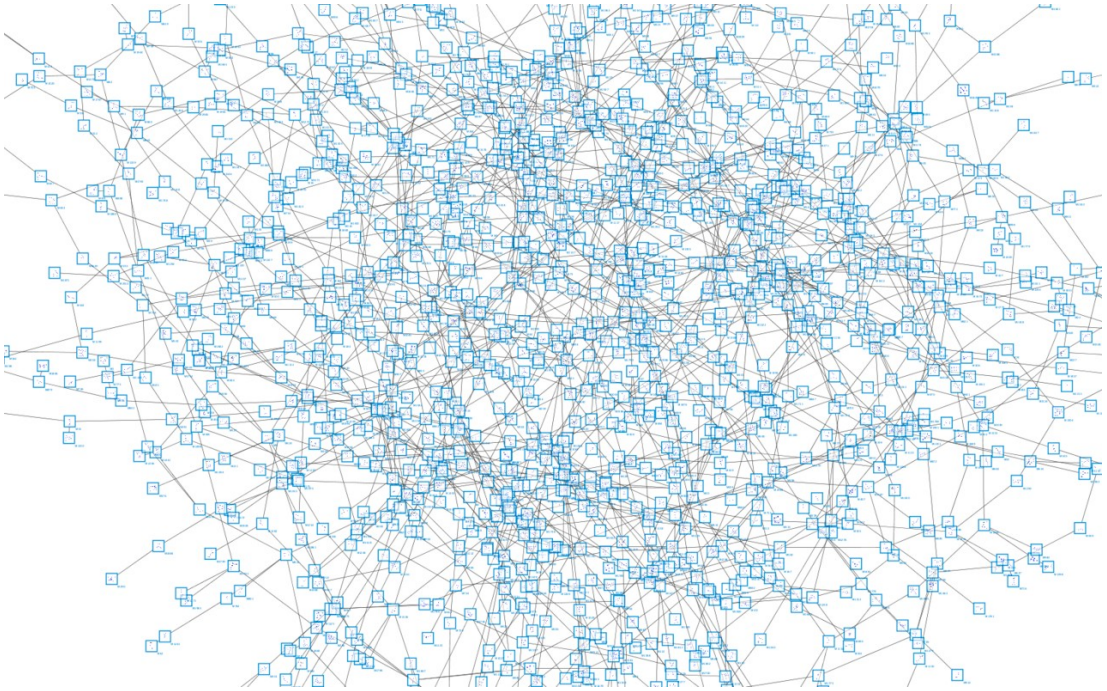
Warm up problem

- Given a **simple** graph with 10 vertices, how many edges can you have?

$10^2 \quad \because \quad \frac{10^2 - 10}{2} = 45$

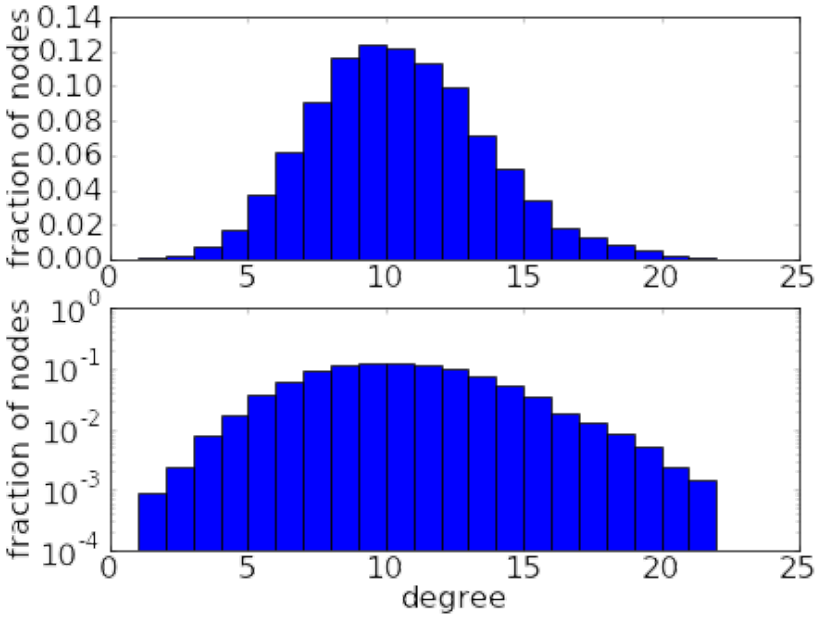
Properties of a Graph

- Degree Distribution
- Structural Properties



Degree Histogram

- Outdegree of a vertex = # outgoing edges
- For each number d, let n(d) = # vertices with outdegree d

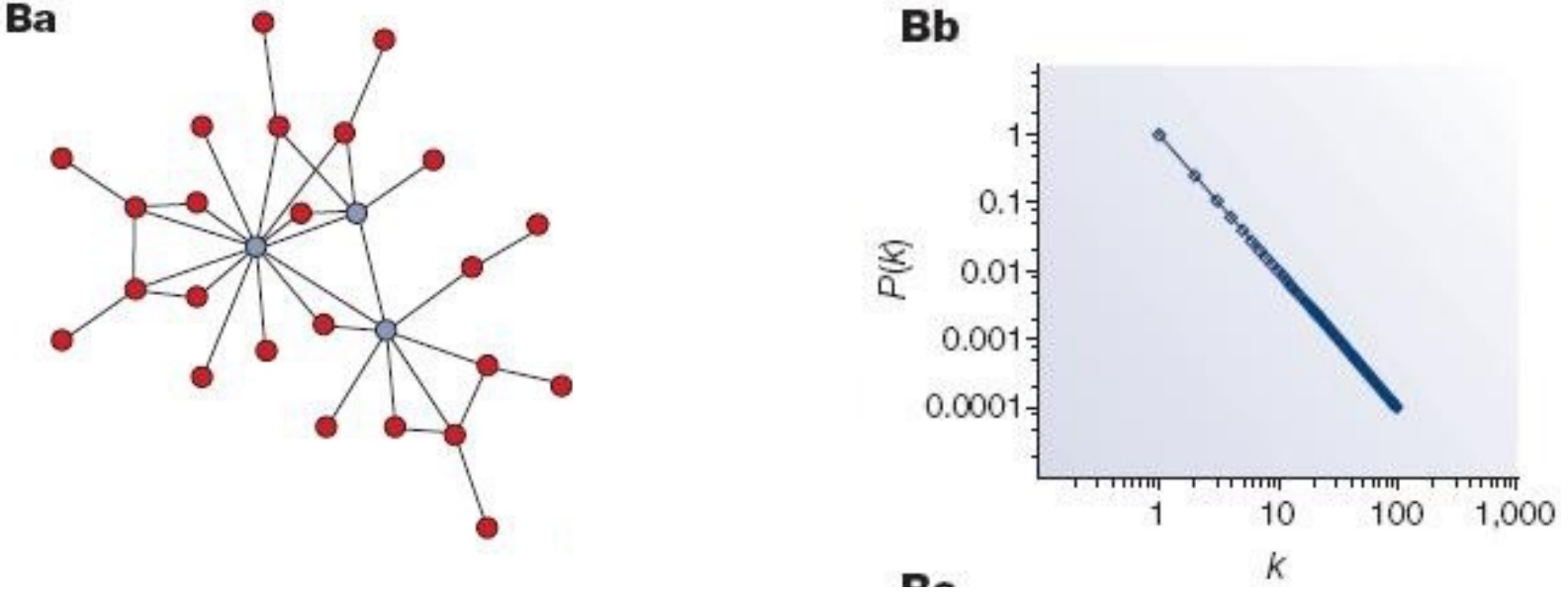


Degree Distribution

- Are there **large graphs** with “interesting” degree distributions?

Zipf Distribution (power law)

- n(d) ~ 1/ d^x for some value x > 0
- Human-generated data has Zipf distribution: letters in alphabet, words in vocabulary.
- In log-log scale



Distribution of the Web

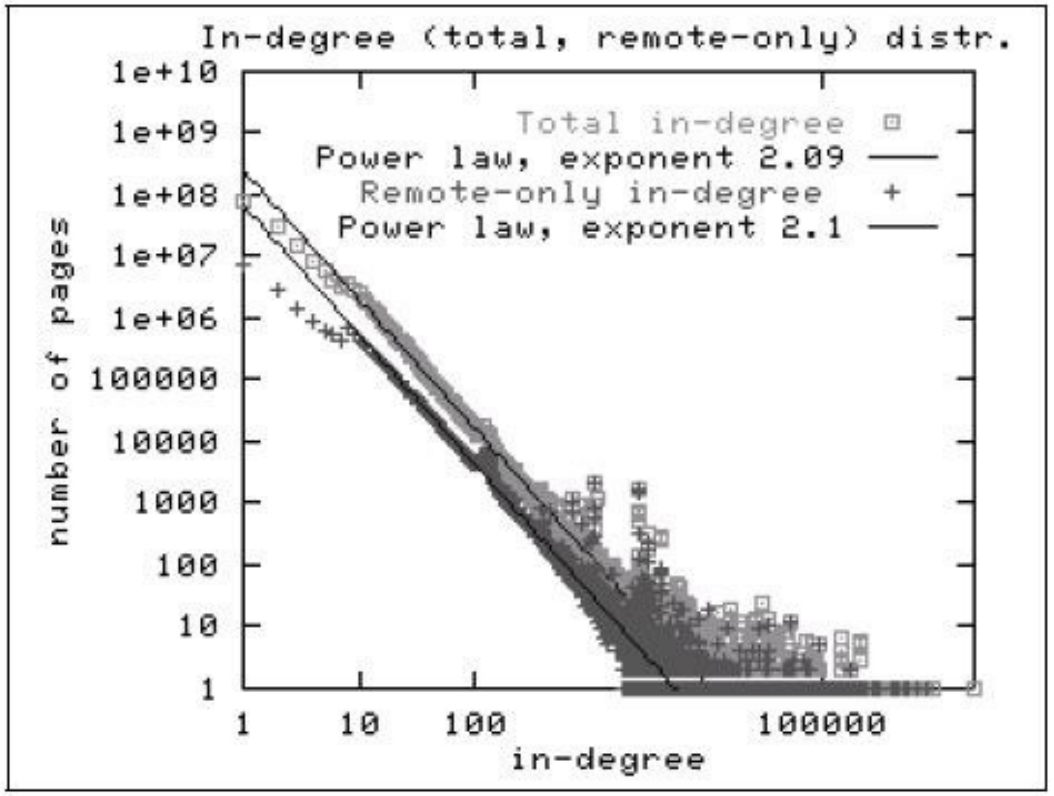
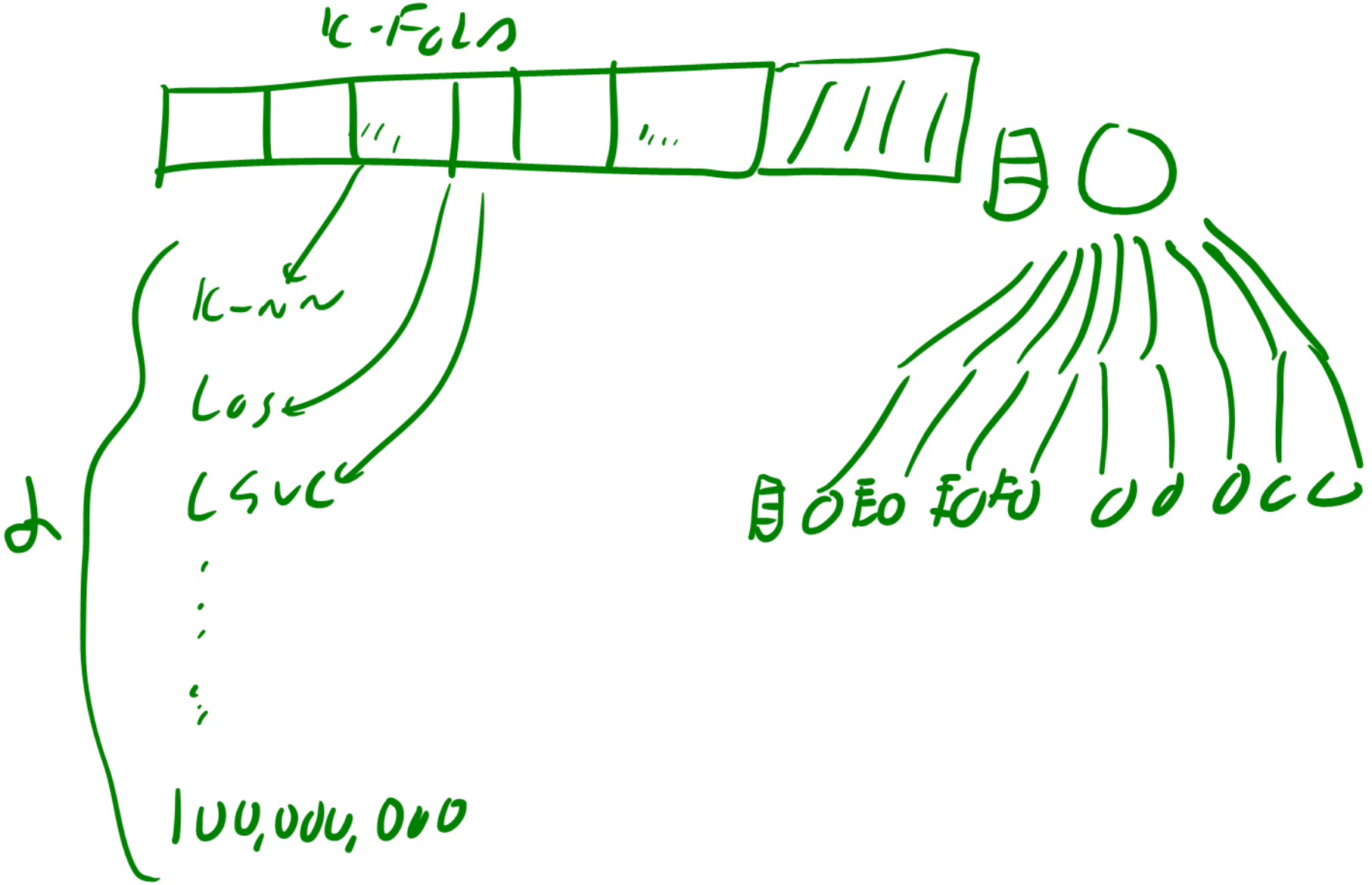


Figure 2: In-degree distribution.

Late 1990's
200M Webpages

Exponential ?

Power Law?



Distribution of Wikipedia