Web-graphs. Lecture 1.

Books

- D. Easley & J. Kleingerg. Networks, Crowds, and Markets
- M. Newman. Networks. An introduction

Complex networks. Introduction

Types of complex networks:

- WWW
- web-graph (nodes – web-pages, links – hyperlinks)
- host-graph (nodes – web-sites, links – hyperlinks)
- social networks:
  - social interaction/friendship
  - scientific collaboration/ co-authorship
  - ...
- financial and economic networks
- biological networks
- logistics networks
- telecommunication (physical) networks
- ...

Density

$$
\begin{aligned}
G &= (V, E) \\
V &= 1, \dots, n \\
\rho &= \begin{cases} \frac{|E|}{C_n^2}, \text{undirected graph} \\ \frac{|E|}{n(n-1)}, \text{directed graph} \end{cases}
\end{aligned}
$$

Web-graphs are directed **multi**-graphs: any number of edges.

Web-graphs are typically **sparse**: $|E| \in [m_1 n, m_2 n]$, where m_1 and m_2 are some constants.

Density

Examples

- Host-graph (November 2011)
  - 86 818 750 nodes
  - 1 391 401 251 edges
- Social network moikrug.ru (2012)
  - 1 547 516 nodes
  - 7 972 911 edges
- Social network Ya.ru (2012)
  - 4 499 044 nodes
  - 6 563 996 edges

Connectivity. Undirected case

From *Easley & Kleingerg. Networks, Crowds, and Markets*

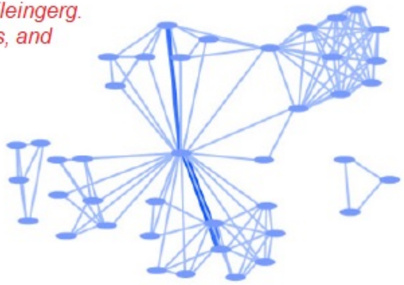

Figure 2.6: The collaboration graph of the biological research center *Structural Genomics of Pathogenic Protozoa (SGPP)* [134], which consists of three distinct connected components. This graph was part of a comparative study of the collaboration patterns graphs of nine research centers supported by NIH's Protein Structure Initiative; SGPP was an intermediate case between centers whose collaboration graph was connected and those for which it was fragmented into many small components.

Disconnected graphs: if a graph is not connected, then it breaks apart naturally into a set of connected "pieces", groups of nodes so that each group is connected when considered as a graph in isolation, and so that no two groups overlap.

Connectivity. Undirected case

Connected component ($=$ component) of a graph:

(i) every node in the subset has a path to every other;

(ii) the subset is not part of some larger set with the property that every node can reach every other.

In complex networks we can typically find one **giant** component (GCC or LCC)

$$\frac{|LCC|}{n} \to \gamma,\ n \to \infty$$

Connectivity. Undirected case

From *Easley & Kleinberg. Networks, Crowds, and Markets*

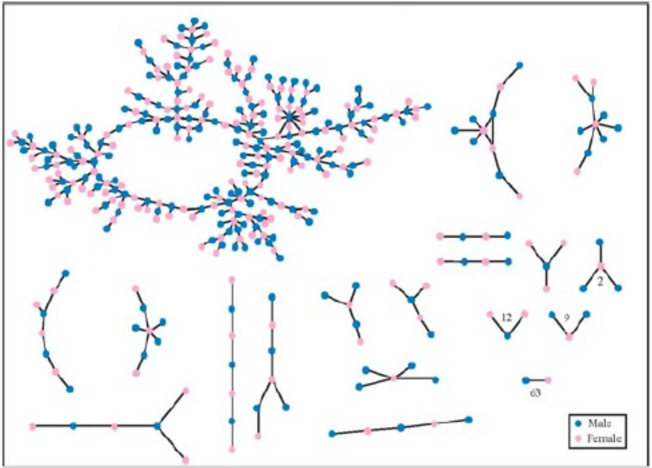

Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [49].

Connectivity. Undirected case

Examples:

- Social network moikrug.ru (2012)
  - 1 547 516 nodes, 7 972 911 edges
  - 145 382 connected components
  - Giant component includes 1 190 249 (**76.9%**) nodes and 4 108 212 edges
- Social network Ya.ru (2012)
  - 4 499 044 nodes, 6 563 996 edges
  - 22 813 connected components
  - Giant component includes 4 445 209 (**98.8%**) nodes and 6 322 600 edges
- Scientific collaboration network based on arXiv preprints (2021):
  - $1.07 \cdot 10^6$ nodes, $4.11 \cdot 10^7$ edges
  - $5.18 \cdot 10^4$ connected components,
  - 90% of nodes in the giant component

Connectivity. Directed case

- Path from a node A to a node B in a directed graph is a sequence of nodes, beginning with A and ending with B, with the property that each consecutive pair of nodes in the sequence is connected by an edge pointing in the forward direction.
- A **weakly connected component (WCC)** in a directed graph = connected component without consideration of edges directions.
- A **strongly connected component (SCC)** in a directed graph is a subset of the nodes such that: (i) every node in the subset has a path to every other; and (ii) the subset is not part of some larger set with the property that every node can reach every other.

Connectivity. Directed case. Bow-tie structure

From *Easley & Kleingerg. Networks, Crowds, and Markets*

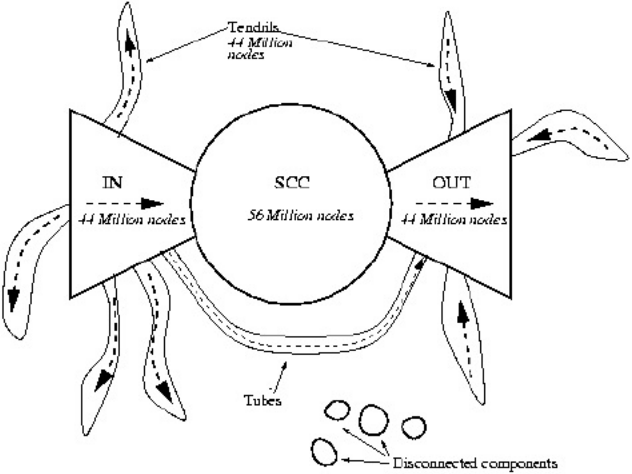

Figure 13.7: A schematic picture of the bow-structure of the Web (image from [80]). Although the numbers are now outdated, the structure has persisted.

Connectivity. Directed case. Bow-tie structure

Bech, Morten L.; Atalay, Enghin (2008) : The topology of the federal funds market, ECB Working Paper, No. 986, European Central Bank (ECB), Frankfurt a. M.

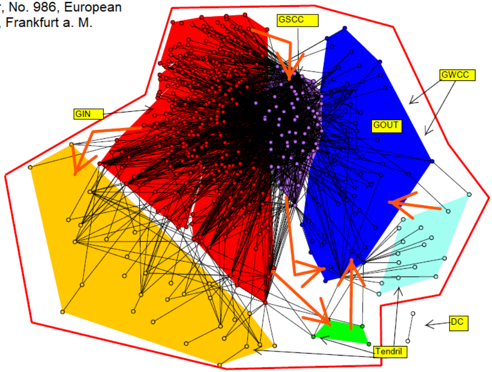

Figure 9: Federal funds network for September 29, 2006. GWCC = giant weakly connected component, DC = disconnected component, GSCC = giant strongly connected component, GIN = giant in-component, GOUT = giant out- component. On this day there were 57 nodes in the GSCC, 303 nodes in the GIN, 67 nodes in GOUT, 50 nodes in the tendrils and 2 nodes in a disconnected component.

Small-World Phenomenon

- **Diameter** is the maximum of shortest paths between nodes.
- Typically in GSCC of web-graphs with consideration of directions diameter is about **10-20**. Without consideration of directions (in GWCC) – ≈ 6
- Scientific collaboration network based on arXiv preprints (2021):
  - all papers: $d = 21$
  - math: $d = 25$
  - CS: $d = 26$
  - Phys: $d = 21$

Small-World Phenomenon

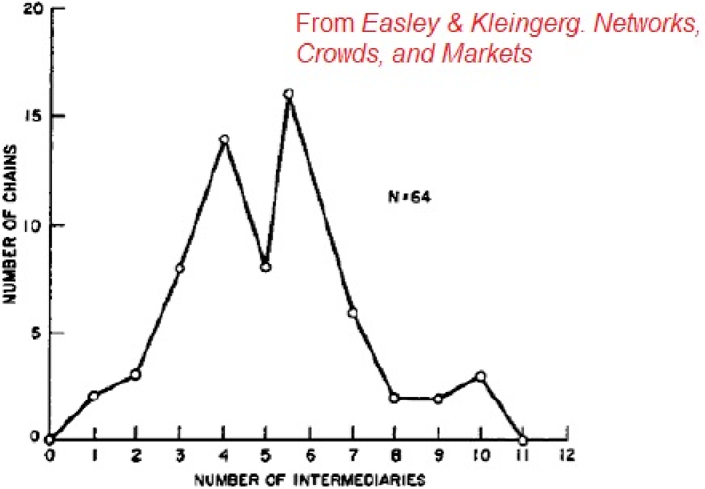

Figure 2.10: A histogram from Travers and Milgram's paper on their small-world experiment [391]. For each possible length (labeled "number of intermediaries" on the x-axis), the plot shows the number of successfully completed chains of that length. In total, 64 chains reached the target person, with a median length of six.

Giant component. Stability

- Stability to attacks on random nodes
- Let
  - G_n be a sequence of growing in time host-graphs,
  - $p \in (0, 1)$ some constant.
- $G_{n,p}$ is a graph obtained from G_n by removal of each node with probability p.
- With probability tending to 1 ($n \to \infty$) there exists GCC in $G_{n,p}$.

Giant component. Vulnerability

- Vulnerability to attacks on hubs.
- Sort nodes by degree in descending order.
- Remove $[c|V|]$ first nodes.
- Let
  - G_n be a sequence of growing in time host-graphs,
  - $c \in (0, 1)$ some constant.
- $G_{n,c}$ is a graph obtained from G_n by removal of first $[c|V|]$ nodes by degree.
- $\exists c^* : \ \forall c \leq c^*$ there exists GCC in $G_{n,c}$, while for $c \geq c^*$ – it does not.

Degree distribution

Let $G_n = (V_n, E_n)$ is a sequence of real growing networks (n = time). Let $\mathrm{adeg} \in \{\mathrm{in-deg}, \mathrm{out-deg}, \mathrm{tot-deg}\}$. There exist constants γ and c such as

$$\frac{|\{v \in V_n : \mathrm{adeg} v = d_n\}|}{|V_n|} \sim \frac{c}{d_n^\gamma}$$

Let

- x is a node's degree
- $\mathrm{PDF}(x) \sim \frac{1}{x^\gamma}$
- $\mathrm{CCDF}(x) = 1 - \mathrm{CDF}(x) \sim \frac{1}{x^{\gamma-1}}$

Degree distribution

Vázquez, A., Pastor-Satorras, R., & Vespignani, A. (2002). Large-scale topological and dynamical properties of the Internet. Physical Review E, 65(6), 066130.