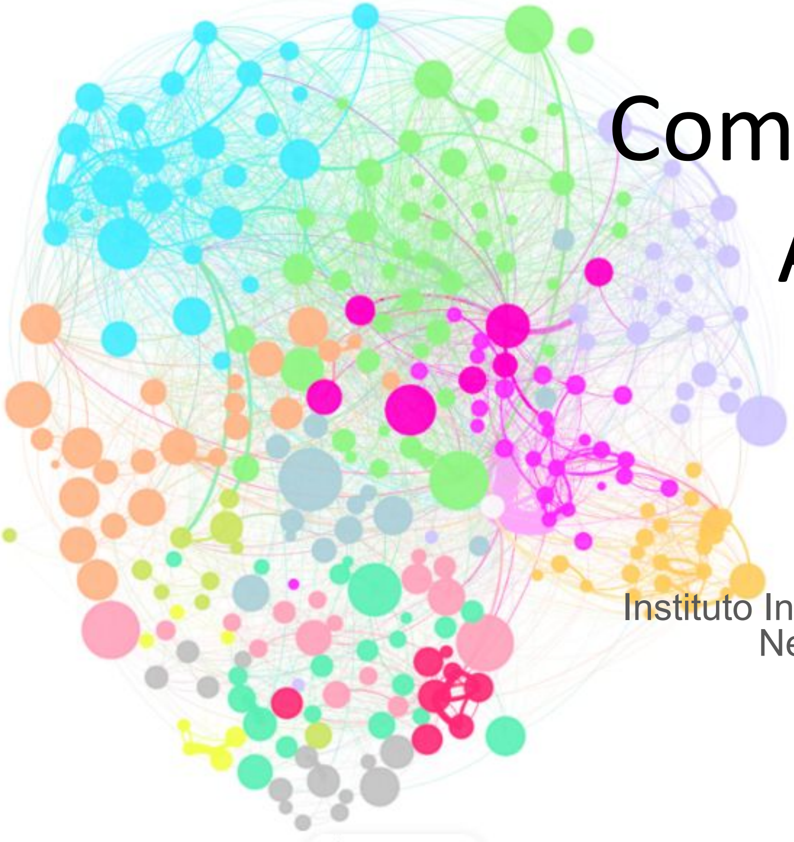


Complex Economic Networks: Analysis, Applications and Data

Viktoriya Semeshenko

Instituto Interdisciplinario de Economía Política (IIEP-UBA-CONICET)
Networks and Complex Systems Laboratory (**Netlab**)

vika.semeshenko@gmail.com
@vsemesh @Netlab_IIEP

Class 1: Introduction to Networks

Outline of classes

- **Class 1**:
  - General Introduction to Networks. Basics and Fundamental metrics
- **Class 2**:
  - Economic case of study. Analogy and differences with social networks. Appropriate metrics. Descriptive statistics.
- **Class 3**:
  - Different networks reduction techniques (Thresholds model, Industry space). Centrality measures. Core-decomposition. Distances.
- **Class 4**:
  - Communities. Concepts and definitions. Hierarchical clustering. Algorithms. Analysis.

Networks

Imagine a world where people have no friends. Where roads never intersect. Where computers are not interconnected. This world without networks would be a very sad and boring place, where nothing happens — and even if something happened, nobody would know. Such a world is unimaginable, because our life is completely defined by networks: relationships, interactions, communications, and the Web. Biological networks governing the interactions between genes in our cells determine our development, neural networks in our brain make us think, information networks guide our knowledge and culture, transportation networks allow us to move, and social networks sustain our life.

Ref: [Filippo Menczer, Santo Fortunato, Clayton A. Davis, A First Course in Network Science. Cambridge University Press 2020]

Networks

- Before enter to the definitions and basic graph concepts about networks, let's look at a few examples, and what's more, let's talk about **stories** that come with each example
    - Social
    - Infrastructure (Power Grid)
    - Financial

- Availability of Data

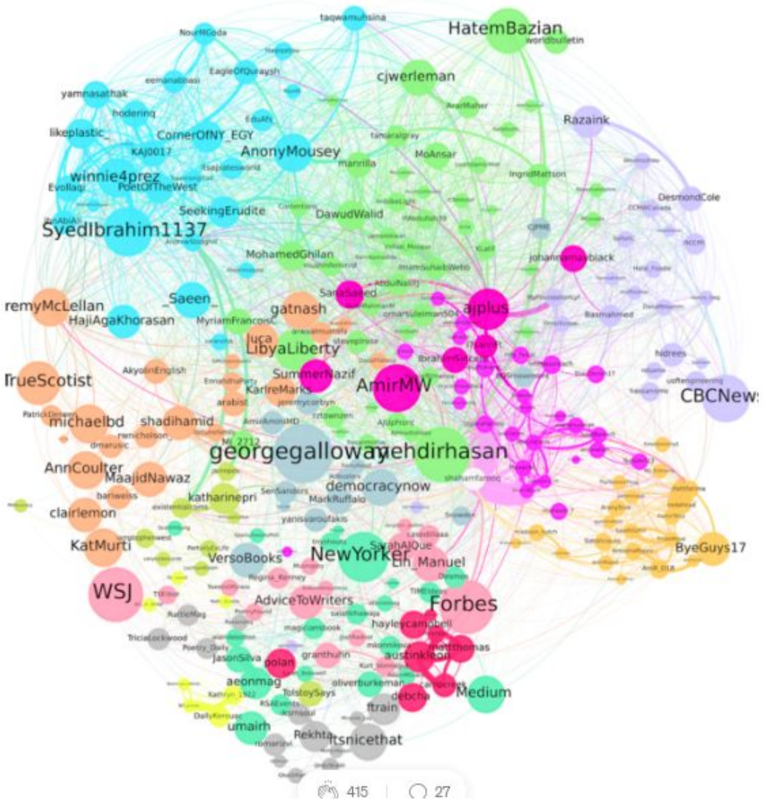
Social Networks

- A social network is a group of persons connected by some type of relationship (state sharing interests)
- Many different types of social networks
    - small size groups: coordination problems/game theory
    - Facebook
    - Twitter
    - Instagram
    - LinkedIn

Story 1: A Twitter Ego-Network

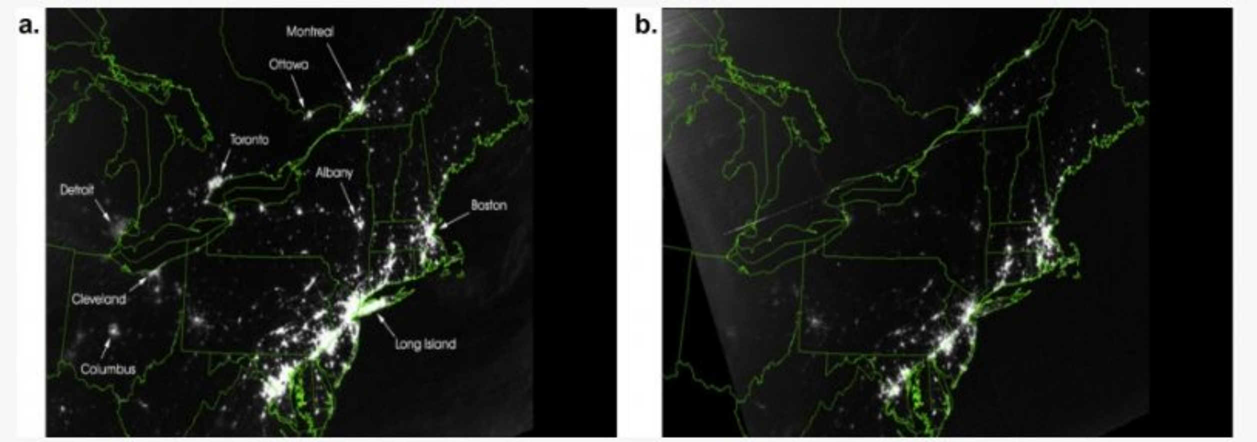
Source: A Graph-based approach to community detection in Twitter Networks

1. Data Mining (API)
2. Building the network
   a. 1-step Neighbourhood
3. Community detection
   a. Find communities based on ‘information flow’ — Infomap
   b. “Users” within interconnected clusters tend to share similar characteristics – Homophily
   c. Classifying Twitter Topic-Networks Using Social Network Analysis



Story 2: Power Grid Network

- At first glance, images (a) and (b) are indistinguishable, showing lights shining in highly populated areas and dark spaces marking vast forests and uninhabited oceans.
- On closer inspection, we notice differences: Toronto, Detroit, Cleveland, Columbus, and Long Island, bright in (a), have darkened in (b).
- Actual image of the northeastern US on August 14, 2003, before and after the blackout that left approximately 45 million people in eight US states, and another 10 million in Ontario without power. The 2003 blackout is a typical example of a cascading failure. **Vulnerability due to interconnectivity**



Source: Network Science. AL Barabási

- Satellite image on Northeast United States on August 13th, 2003,at 9:29pm (EDT), 20 hours *before* the 2003 blackout.
- The same as above, but 5 hours *after* the blackout.

Vulnerability due to interconnectivity: avalanches

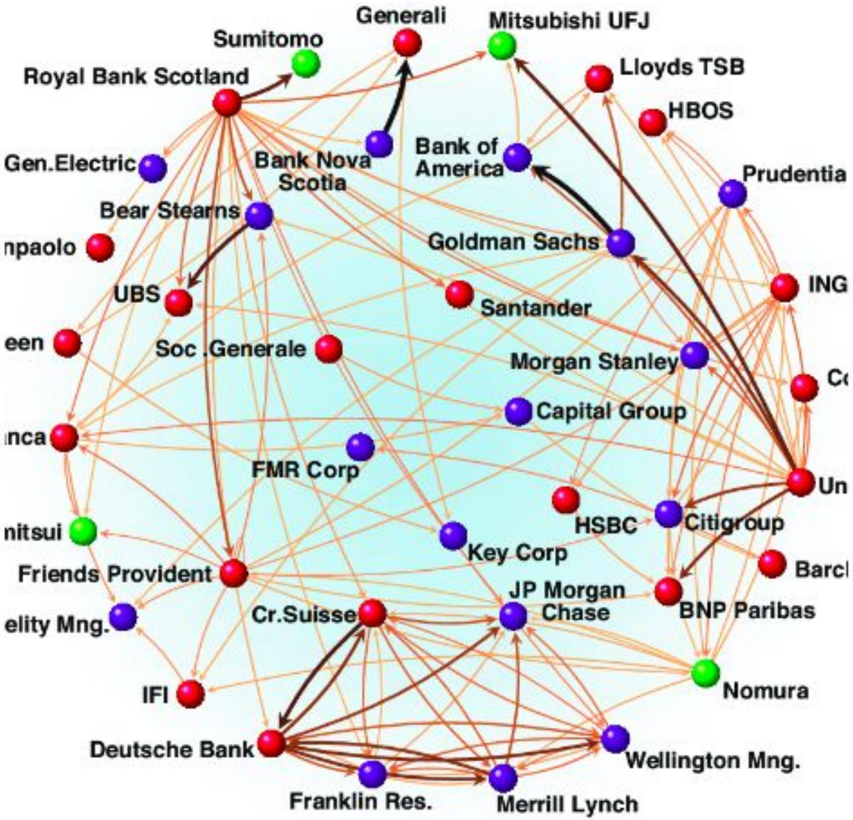
- When a network acts as a transportation system, a local failure spreads loads to other nodes. If the additional load is negligible, the system can absorb it without problems, and the failure goes unnoticed. However, if the additional load is too much for neighboring nodes, they will also fail and redistribute the load to their neighbors.
- We are faced with a **cascading** event, the magnitude of which depends on the <u>position</u> and <u>capacity</u> of the nodes that initially failed.
- The Northeast blackout example illustrates several important issues:
  - to avoid damage from cascades, we need to understand the <u>structure</u> of the network in which the cascade propagates
  - we need to model the dynamic processes that take place in these networks, such as the flow of electricity
  - we need to find out how the interaction between network structure and dynamics affects the robustness of the whole system. Although cascade failures may seem random and unpredictable, they follow reproducible laws that can be quantified and even be predicted using network science tools

Story 3: Financial Networks

- Cascading failures have been observed in many complex systems
  - financial systems

➢ **Propagation**
➢ **Fragility** and **robustness**
➢ **Too big** to fail
➢ **Too central** to fail

¿ Prediction?

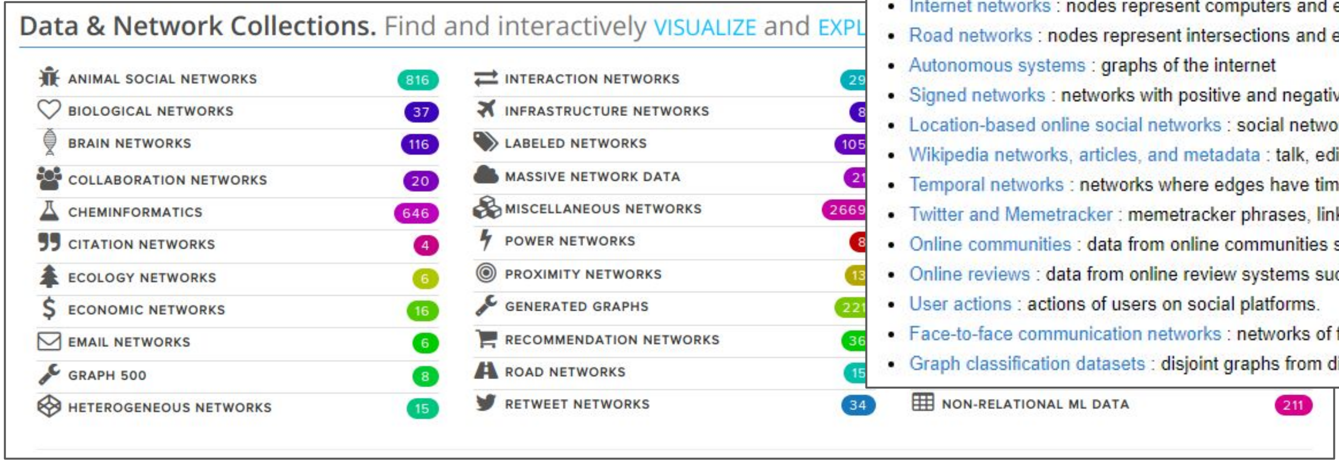


Source: [Economic Networks: The New Challenges, Frank Schweitzer et al, 2009]

Availability of Data: Fuentes

- Stanford Large Network Dataset Collection
- mobility patterns:
  - Air transportation US network
- Internet Movie Database
- An Interactive Scientific Network

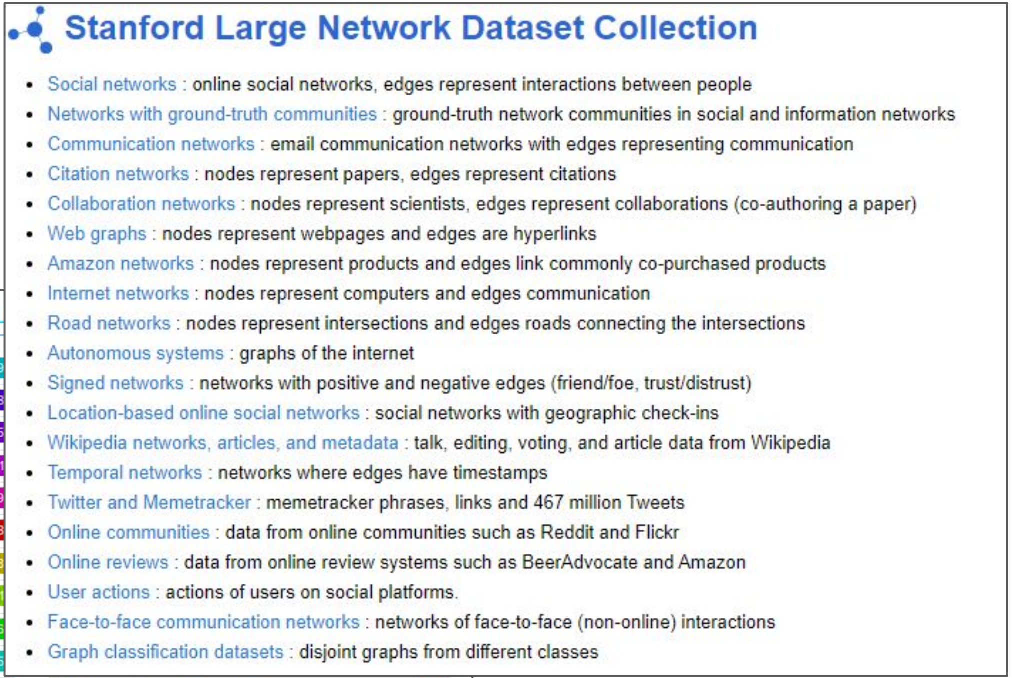
Data Repository https://networkrepository.com/



Data & Network Collections. Find and interactively VISUALIZE and EXPL

ANIMAL SOCIAL NETWORKS	816	INTERACTION NETWORKS	29
BIOLOGICAL NETWORKS	37	INFRASTRUCTURE NETWORKS	8
BRAIN NETWORKS	116	LABELED NETWORKS	105
COLLABORATION NETWORKS	20	MASSIVE NETWORK DATA	21
CHEMINFORMATICS	646	MISCELLANEOUS NETWORKS	2669
CITATION NETWORKS	4	POWER NETWORKS	8
ECOLOGY NETWORKS	6	PROXIMITY NETWORKS	13
ECONOMIC NETWORKS	16	GENERATED GRAPHS	221
EMAIL NETWORKS	6	RECOMMENDATION NETWORKS	36
GRAPH 500	8	ROAD NETWORKS	15
HETEROGENEOUS NETWORKS	15	RETWEET NETWORKS	34

NON-RELATIONAL ML DATA 211



Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : social networks with geographic check-ins
- Wikipedia networks, articles, and metadata : talk, editing, voting, and article data from Wikipedia
- Temporal networks : networks where edges have timestamps
- Twitter and Memetracker : memetracker phrases, links and 467 million Tweets
- Online communities : data from online communities such as Reddit and Flickr
- Online reviews : data from online review systems such as BeerAdvocate and Amazon
- User actions : actions of users on social platforms.
- Face-to-face communication networks : networks of face-to-face (non-online) interactions
- Graph classification datasets : disjoint graphs from different classes

Availability of Data

- Social networks
  - any scale and resolution
  - this data provides us with an unprecedented opportunity to discover, track, mine and model what people do
  - problem of ethics: expose of private personal information

- Economic networks
  - usually private data
  - public surveys (ex: *EPH*)
  - alternative mechanisms: web scraping/ complex network analysis/ semantic analysis

Definitions and the Basic Graph Concepts

Sources used in the course:

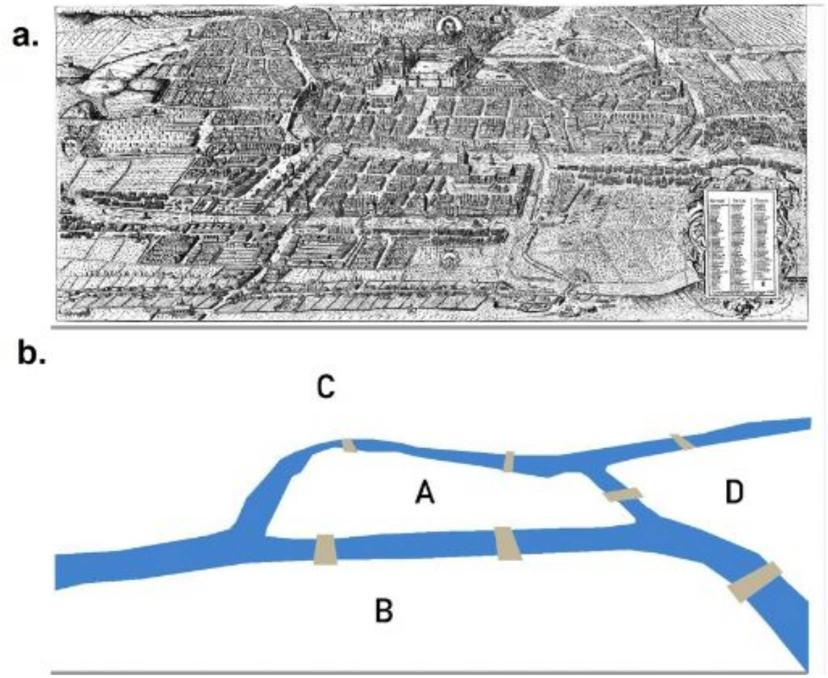
- Barabási: Albert-László Barabási. Network Science
- Menczer et al: Filippo Menczer, Santo Fortunato, Clayton A. Davis, A First Course in Network Science. Cambridge University Press 2020

Networks: Definitions and Basics

A network $\boldsymbol{G}$ has two parts, a set of $\boldsymbol{N}$ elements, called *nodes* or *vertices*, and a set of $\boldsymbol{L}$ pairs of nodes, called *links* or *edges*. The link joins the nodes $\boldsymbol{i}$ and $\boldsymbol{j}$. A network can be *undirected* or *directed*. A directed network is also called a *digraph*. In directed networks, links are called *directed* links and the order of the nodes in a link reflects the direction: the link goes from the source node $\boldsymbol{i}$ to the target node $\boldsymbol{j}$. In undirected networks, all links are *bi-directional* and the order of the two nodes in a link does not matter. A network can be *unweighted* or *weighted*. In a weighted network, links have associated *weights*: the weighted link between nodes i and j has weight $\boldsymbol{w}$. A network can be both *directed* and *weighted*, in which case it has directed weighted links.

Networks: Definitions and Basics

The Bridges of Königsberg ([Euler](#) video)



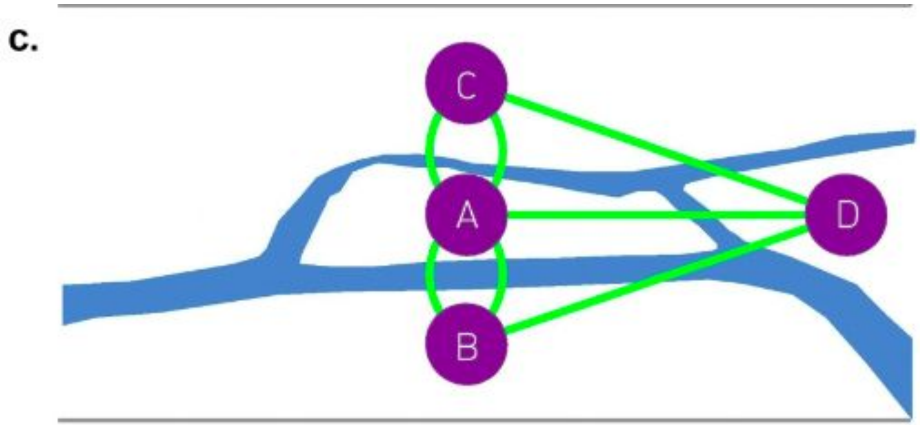
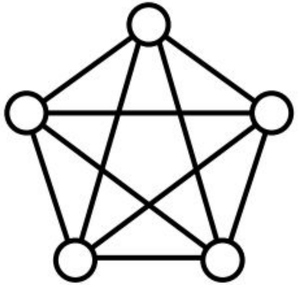


Image 2.1

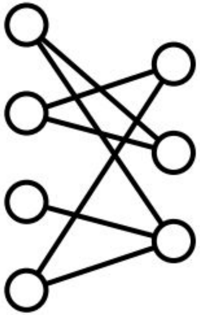
The Bridges of Königsberg

- A contemporary map of Königsberg (now Kaliningrad, Russia) during Euler's time.
- A schematic illustration of Königsberg's four land pieces and the seven bridges across them.
- Euler constructed a graph that has four nodes (A, B, C, D), each corresponding to a patch of land, and seven links, each corresponding to a bridge. He then showed that there is no continuous path that would cross the seven bridges while never crossing the same bridge twice. The people of Königsberg gave up their fruitless search and in 1875 built a new bridge between B and C, increasing the number of links of these two nodes to four. Now only one node was left with an odd number of links. Consequently we should be able to find the desired path. Can you find one yourself?

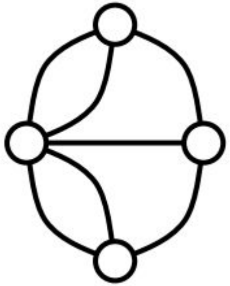
Networks: types



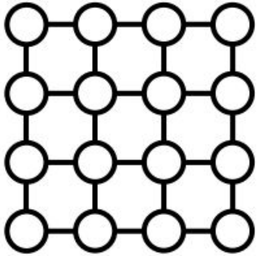
complete



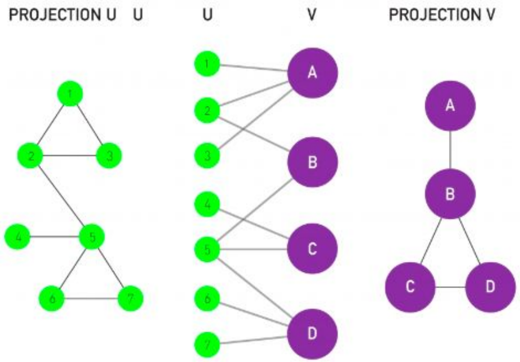
bipartite



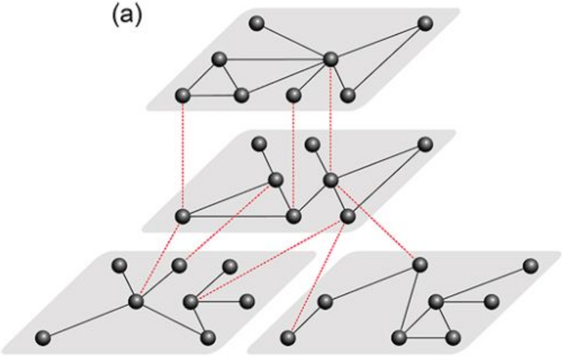
random



regular



bipartite:projections



multilayer

Networks: Density and Sparsity

The maximum number of links is bounded by the possible number of distinct connections among the nodes of the system

- if all possible pairs of nodes are connected by links → a **complete** network

- Undirected: is the number of distinct pairs of nodes $L_{\max} = \binom{N}{2} = N(N-1)/2$
- Directed: each pair of nodes counts, one for each direction $L_{max} = N(N-1)$
- **Density** $d = L/L_{\max}$
  - Undirected: $d = L/L_{\max} = \dfrac{2L}{N(N-1)}$
  - Directed: $d = L/L_{\max} = \dfrac{L}{N(N-1)}$

This is an important feature that helps in dealing with network structure → **Sparsity**
The fewer edges are in a network, the sparser it is

Networks: Density and Sparsity

- *Complete* network, $L=L_{max}$, so $d=1$
- *Sparse* network, $L \ll L_{max}$, so $d \ll 1$
- The network is sparse if the number of links grows proportionally to the number of nodes ($L \sim N$), or even slower
- If the number of links grows faster, e.g. quadratically with network size ($L \sim N^2$), then we say that the network is dense

To illustrate the importance of network sparsity, let us consider the example of Facebook. At the time of writing, Facebook has around 2 billion users ($N \approx 2 \times 10^9$). If this was a complete network, there would be $L \approx 10^{18}$ links — that is a number with 18 zeros, and there is no way to store so much data! But fortunately, social networks are very sparse and Facebook is no exception. Each user has on average 1000 friends or less, so that the density is approximately $d \approx 10^{-6}$. That is still a lot of data, but Facebook can manage it.

Examples of sparse networks

Source: Menczer, Fortunato, Davis. (2020) A First Course In Network Science)

Network	Type	Nodes (N)	Links (L)	Density (d)	Average degree ($\langle k \rangle$)
Facebook Northwestern Univ.		10,567	488,337	0.009	92.4
IMDB movies and stars		563,443	921,160	0.000006	3.3
IMDB co-stars	W	252,999	1,015,187	0.00003	8.0
Twitter US politics	DW	18,470	48,365	0.0001	2.6
Enron email	DW	87,273	321,918	0.00004	3.7
Wikipedia math	D	15,220	194,103	0.0008	12.8
Internet routers		190,914	607,610	0.00003	6.4
US air transportation		546	2,781	0.02	10.2
World air transportation		3,179	18,617	0.004	11.7
Yeast protein interactions		1,870	2,277	0.001	2.4
C. elegans brain	DW	297	2,345	0.03	7.9
Everglades ecological food web	DW	69	916	0.2	13.3

Degree, Average Degree and Degree Distribution

- k_i is the **degree** of the i^{th} node in the network: represents the number of links to other nodes.
- Undirected network: the total number of links, L, is the sum of the node degrees $L = \frac{1}{2}\sum_{i=1}^{N} k_i$
- The **average degree**
  - Undirected networks: $\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$

- Directed networks, we distinguish between incoming degree, k_i^{in}, representing the number of links that point to node *i*, and outgoing degree, k_i^{out}, representing the number of links that point from node *i* to other nodes. A node's total degree, is given by $k_i = k_i^{in} + k_i^{out}$

- Directed networks, the total number of links is $L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$
- The average degree of a directed network is

$$\langle k^{in} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{out} = \frac{L}{N}$$

Networks: Definitions and Basics

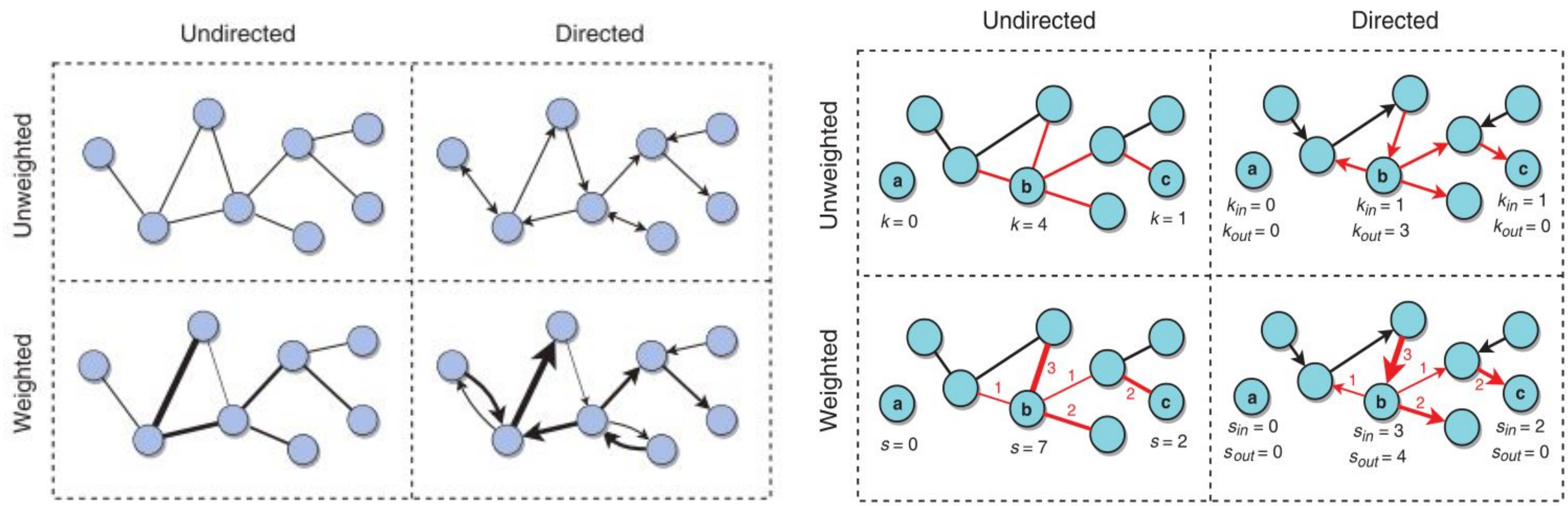


Image 1. Illustrations of degree and strength in directed, undirected, weighted, and unweighted networks. *Source*: Barabási.

Network representation

- **Adjacency Matrix**
  - of a network with N nodes has N rows and N columns, and each element A_{ij} takes value 1 if there is a link pointing from node *i* to node *j*, and 0 otherwise.
  - The degree k_i of node *i* can be directly obtained from the elements of the matrix:
    - Undirected: is a sum over either the rows or the columns of the matrix (symmetric)

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{j=1}^{N} A_{ji}$$

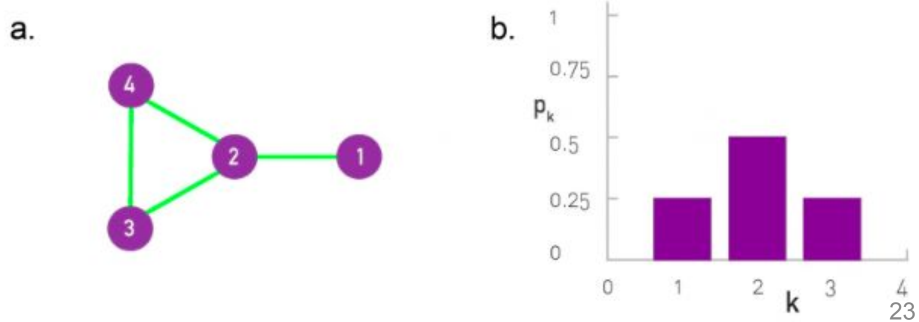
    - Directed: the sums over the adjacency matrix' rows and columns provide the incoming and outgoing degrees, respectively

$$k_i^{in} = \sum_{j=1}^{N} A_{ij} \qquad k_i^{out} = \sum_{j=1}^{N} A_{ji}$$

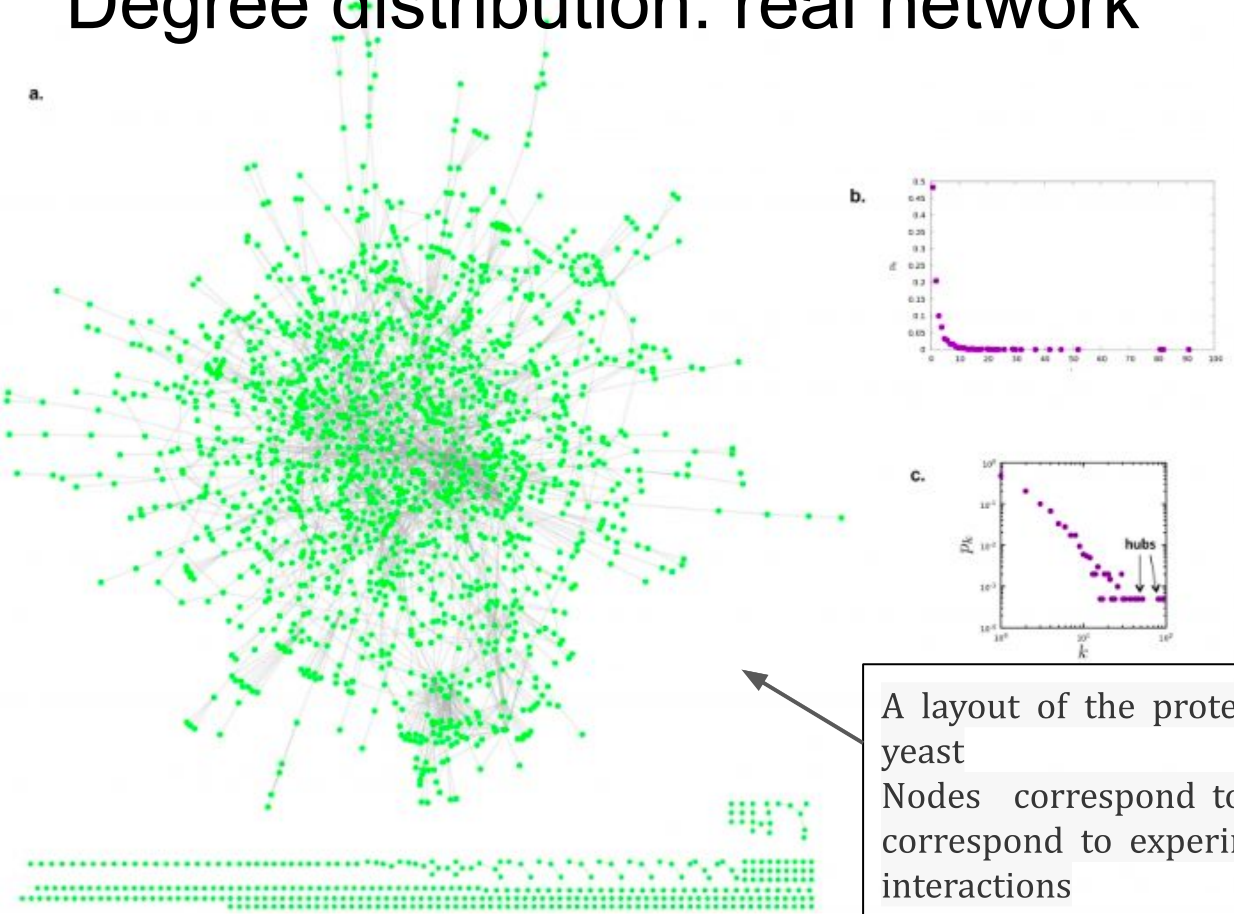
- **Adjacency List**
  - A list of links whose element "i→j" shows a link going from node i to node j
  - Also represented as "i→{j1,j2,j3,...}"

Degree Distribution

- The degree distribution, p_k, provides the probability that a randomly selected node in the network has degree k, ($\sum p_k = 1$)
- For a network with N nodes the degree distribution is the normalized histogram is given by $p_k = N_k/N$, where N_k is the number of degree-k nodes
- The degree distribution plays a central role in network theory after the discovery of scale-free networks
- Most of network properties requires to know p_k
- The precise functional form of p_k determines many network phenomena, from network robustness to the propagation dynamics



Degree distribution: real network



The degree distribution of this network.The observed degrees vary between k=0 (isolated nodes) and k=92, which is the degree of the most connected node, called a hub. Almost half of the nodes have degree one (i.e. p1=0.48).

A layout of the protein interaction network of yeast
Nodes correspond to yeast proteins and links correspond to experimentally detected binding interactions

Clustering Coefficient

- The clustering coefficient captures the degree to which the neighbors of a given node link to each other
- For a node *i* with degree k_i the local clustering coefficient is defined as $C_i = \frac{2L_i}{k_i(k_i - 1)}$
  - where L_i represents the number of links between the k_i neighbors of node *i*
  - Note that C_i is between 0 and 1

- $C_i = 0$ if none of the neighbors of node *i* link to each other
- $C_i = 1$ if the neighbors of node *i* form a complete graph, i.e. they all link to each other
- C_i is the probability that two neighbors of a node link to each other. Consequently $C = 0.5$ implies that there is a 50% chance that two neighbors of a node are linked
- C_i measures the network's local link density: The more densely interconnected the neighborhood of node i, the higher is its local clustering coefficient.

Clustering Coefficient

- Average Clustering Coefficient $\langle C \rangle$, represents the average of C_i of all nodes in the network

$$\langle C \rangle = \frac{1}{N} \sum C_i$$

- $\langle C \rangle$ is the probability that two neighbours of a randomly selected node link to each other

Further readings

There are several other excellent textbooks on network science

- Data science of several case studies, Caldarelli and Chessa (2016)
- To explore the connections to economics and sociology, Easley and Kleinberg (2010)
- Complex Webs in Nature and Technology Caldarelli, 2007
- Dynamical processes on complex networks, Barrat et al., 2008;
- Multilayer Networks Kivelä et al. (2014) and Boccaletti et al. (2014)

See you next class!