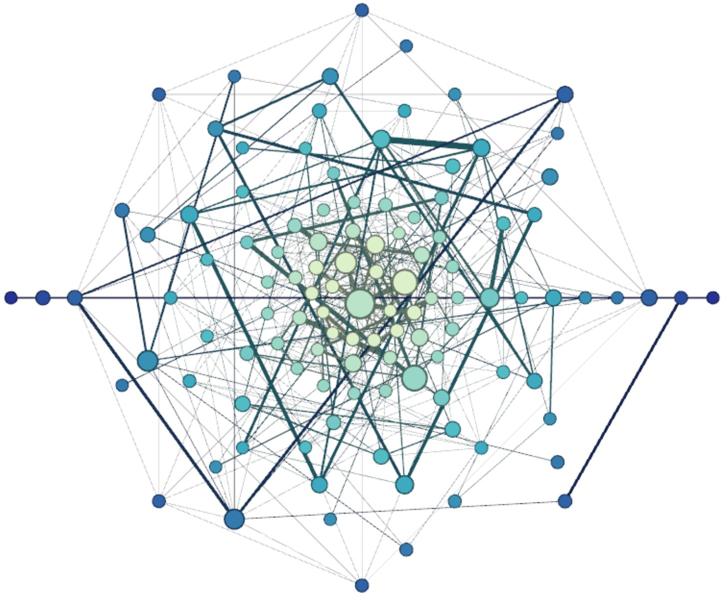
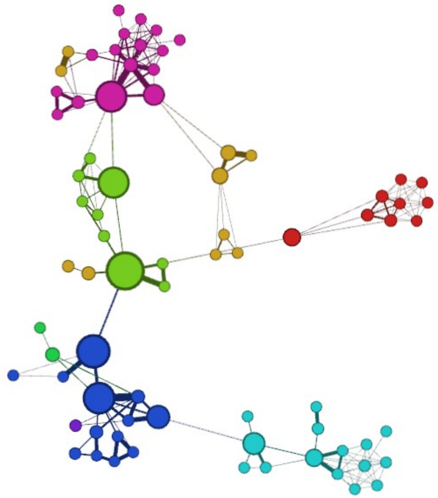
Introduction to the Analysis and Visualisation of Complex Networks

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Pedro Ribeiro
(DCC/FCUP & CRACS/INESC-TEC)

CRACS
INESCTEC
ASSOCIATE LABORATORY
PORTUGAL





CLAD
Associação Portuguesa de
Classificação e Análise de Dados

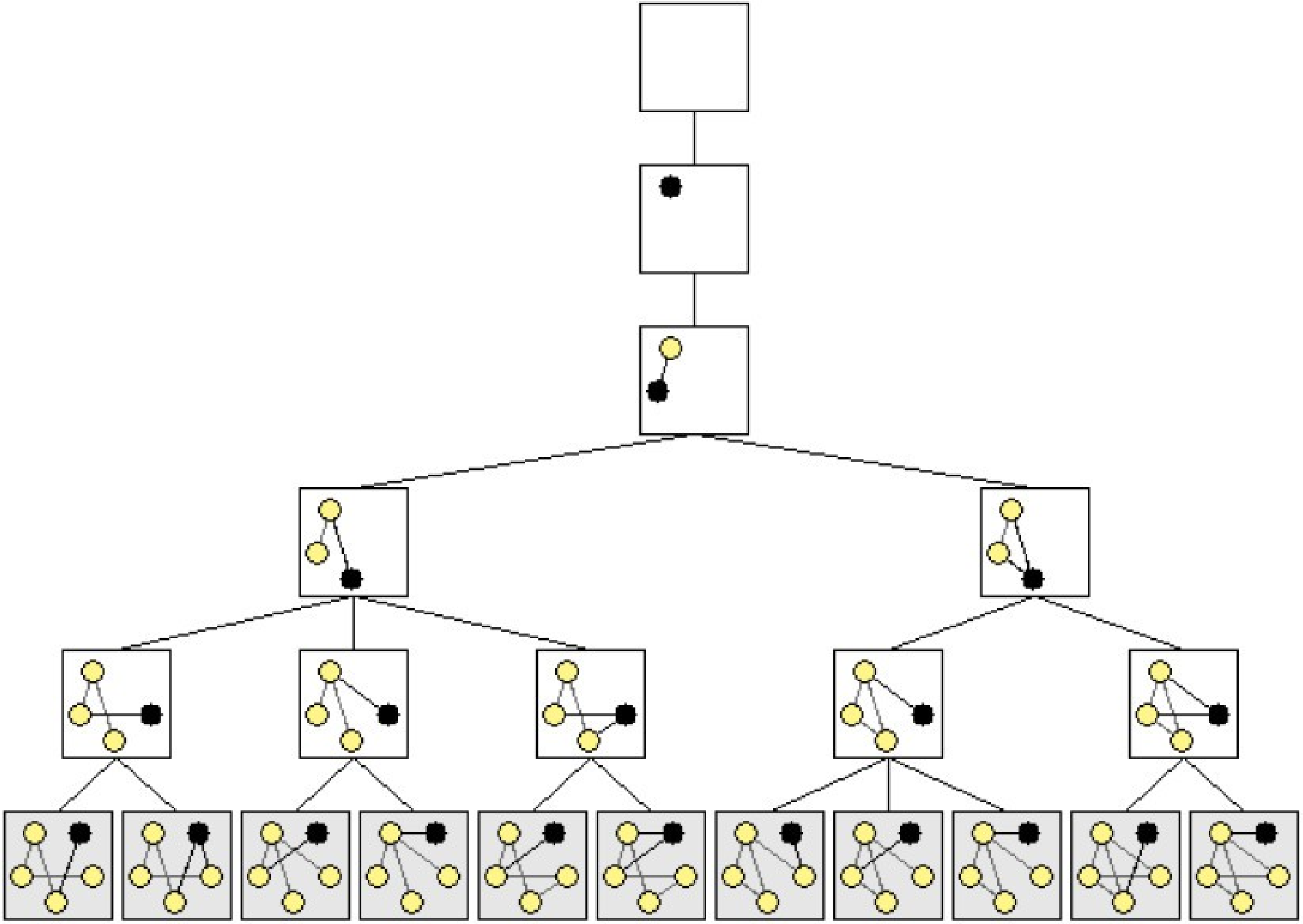
Who am I?

- Name: **Pedro Ribeiro**
- Website: https://www.dcc.fc.up.pt/~pribeiro/
- PhD in Computer Science
- Main research interests:
  - Complex Network Analysis, Network Science, Graph Mining, Data Mining
  - Algorithms and Data Structures, Complexity
  - Parallel and Distributed Computing
- Other research interests:
  - Computer Science Education and Programming Contests
  - Artificial Intelligence, Machine Learning

PhD in Computer Science (2011)

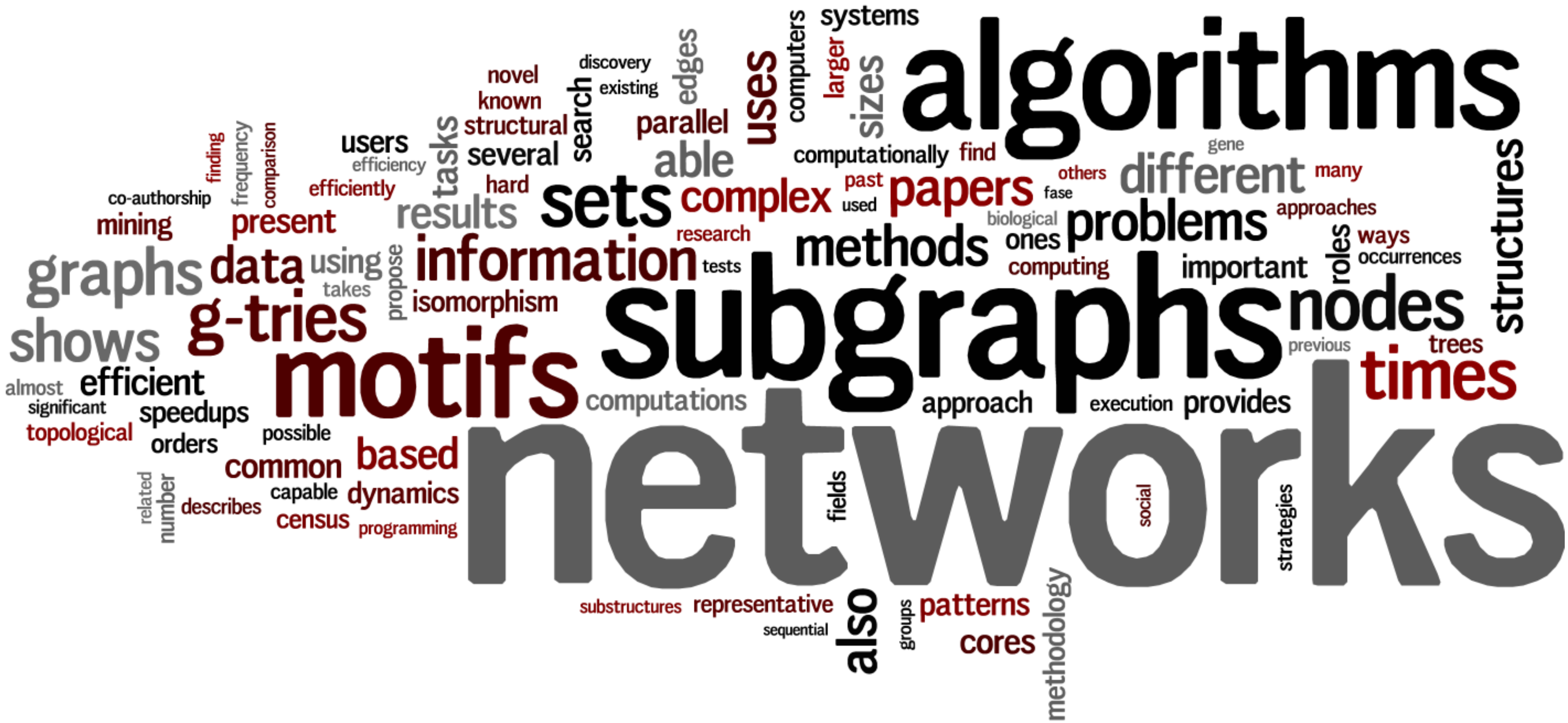
- Thesis: Efficient and Scalable Algorithms for Network Motifs Discovery

PhD: G-Tries



Publications

- A **word cloud** of my publications abstracts *(made in Dec 2015)*



http://www.dcc.fc.up.pt/~pribeiro/pubs_by_year.html

https://scholar.google.com/citations?user=DbR0sO4AAAAJ

Some of my former students

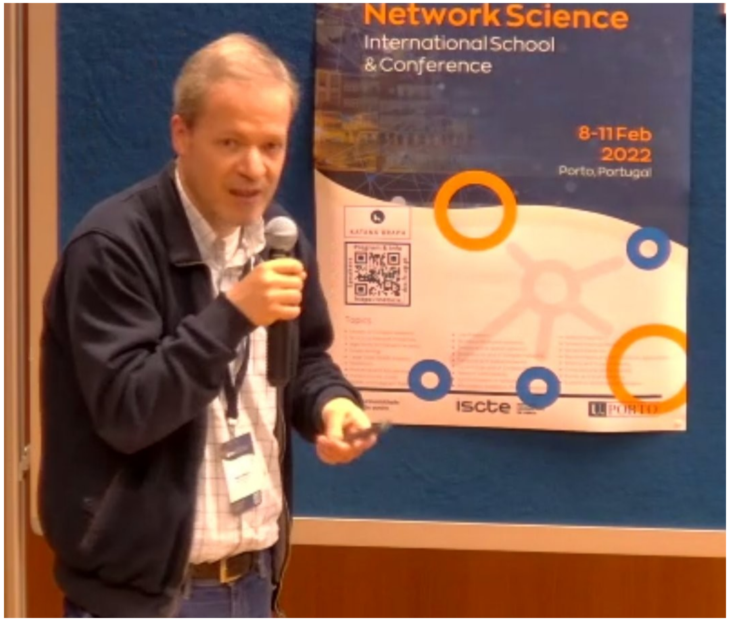
- **PhD Students**
  - **Sarvenaz Choodbdar** (2010-2015):
    On the Characterization and Comparison of Complex Networks
  - **Miguel Araújo** (2012-2017)
    Communities and Anomaly Detection in Large Edge-Labeled Graphs
  - **David Aparício** (2014-2020)
    Network Comparison and Node Ranking in Complex Networks
  - **Jorge Silva** (2016-2021)
    Towards measuring scientific impact using network science
- **MSc Students**
  - F. Justiça (21/22): *Time Series Forecasting via Network Science*
  - J. Ferreira (21/22): *Subgraph Patterns in Spatial Networks*
  - M. Lamas (21/22): *Characterizing Music through Complex Networks*
  - I. Novo (21/22): *On the Summarization of Complex Networks*
  - B. Pinto (20/21): *Subgraph Patterns in Colored Networks*
  - H. Branquinho (19/20): *Counting Subgraphs in Streaming Networks*
  - F. Bento (19/20): *Characterizing the Passing Networks of Football Teams*

Some of my current students

- PhD Students:
  - **Vanessa Silva** (since 2018):
    Analysing Time Series using Complex Networks
  - **Alberto Barbosa** (since 2018)
    Spatio-Temporal Network Patterns
  - **Ahmad Naser Eddin** (since 2019)
    Fraud and Anti-Money Laundering Detection using Network Science
  - **Luciano Grácio** (since 2020)
    Fundamental contributions on Subgraph Counting and Graph Theory
  - **André Meira** (since 2021)
    Multilayer Networks
  - **Hugo Oliveira** (since 2021)
    Transformers for Medical Domains
  - **Miguel Ferreira** (since 2021)
    Spatial Graph Databases

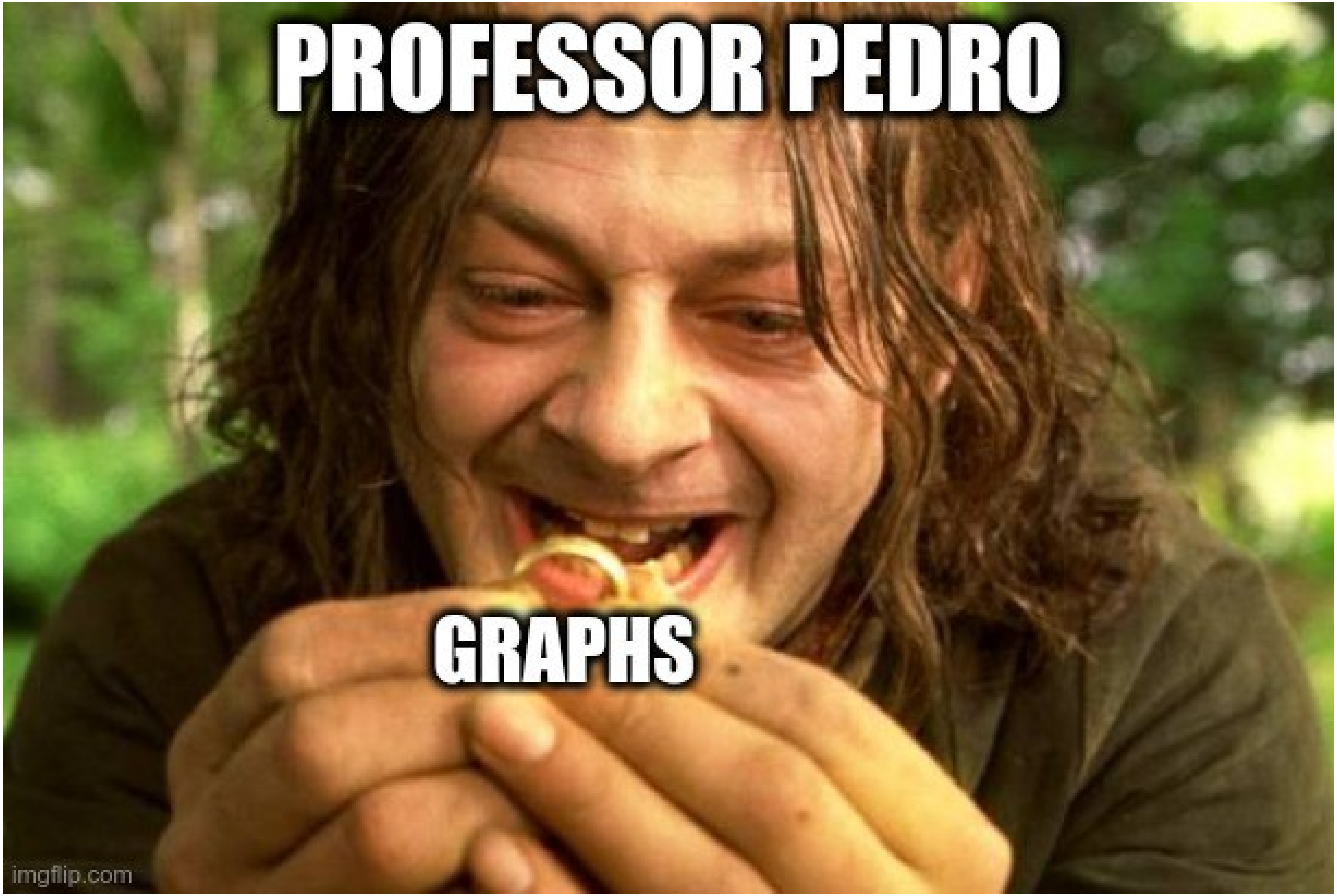
Network Science Events

http://netsci18.dcc.fc.up.pt/



https://netscix.dcc.fc.up.pt/

Graphs



Structure of this Course

Overview

- **Part 1** *(09:45 – 11:15)*:
  - Introduction to the course
  - Motivation and the "small world" phenomenon
  - Emergence of Network Science
  - Brief introduction to Graph Theory and its terminology
- **Part 2** *(11:30 - 13:00)*:
  - Measuring Real Networks and their typical properties
  - Graph generation models (Erdös-Rényi, Watts-Strogatz, Barabasi-Albert)
  - Node Centrality (e.g. closeness, betweenness, eigenvector, PageRank)
- **Part 3** *(14:30 - 16:00)*:
  - Community Structure and Modularity
  - Activity: visualisation and analysis of networks using a graphic platform: Gephi
- **Part 4** *(16:15 - 17:45)*:
  - Network construction
  - Activity: visualization and analysis of networks programmatically: igraph (R)
  - Brief introduction to other topics (e.g. propagation, sub-graphs, link prediction, GNNs)

https://www.dcc.fc.up.pt/~pribeiro/clad2022/

Motivation and the “small world” phenomenon

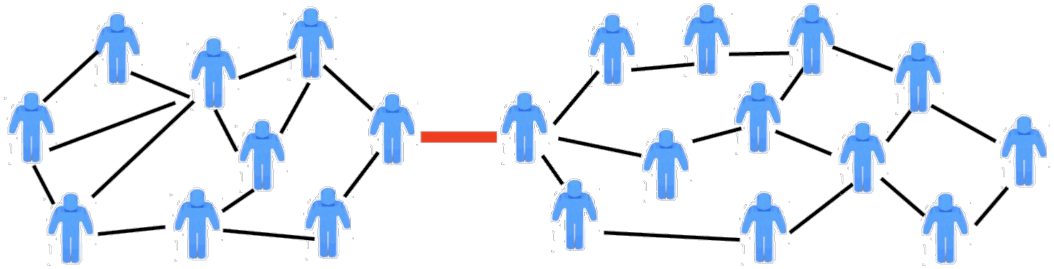
Planet Earth

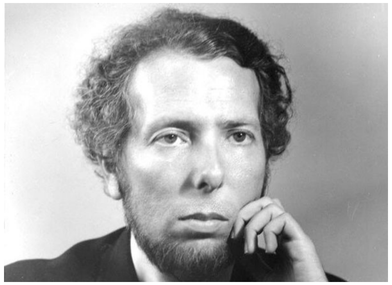

8 Billion
Humans

How many “degrees” of separation?

1929 Frigyes Karinthy

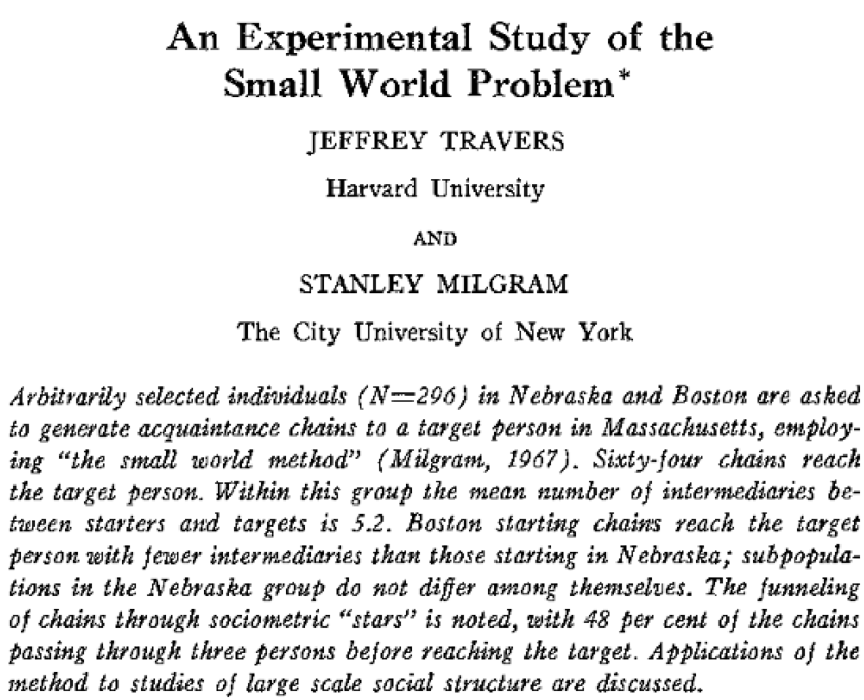
"If you choose a person out of the 1.5 billions of our planet, I bet that using no more than ***five*** individuals, one of them my acquaintance, I could contact the person you chose, using only the list of acquaintances of each one"





1969 Stanley Milgram

- People chosen at random on a US State
- Request to send a letter to a given final person in another state :
  - If you know the final person, send directly to him
  - If not, send to someone you think it is more likely to know him



An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

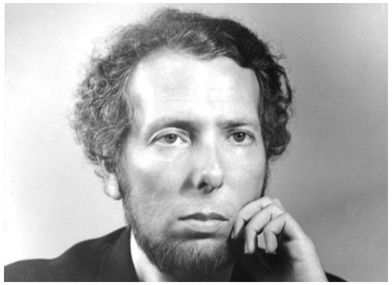
Harvard University

AND

STANLEY MILGRAM

The City University of New York

Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target person with fewer intermediaries than those starting in Nebraska; subpopulations in the Nebraska group do not differ among themselves. The funneling of chains through sociometric "stars" is noted, with 48 per cent of the chains passing through three persons before reaching the target. Applications of the method to studies of large scale social structure are discussed.



1969 Stanley Milgram

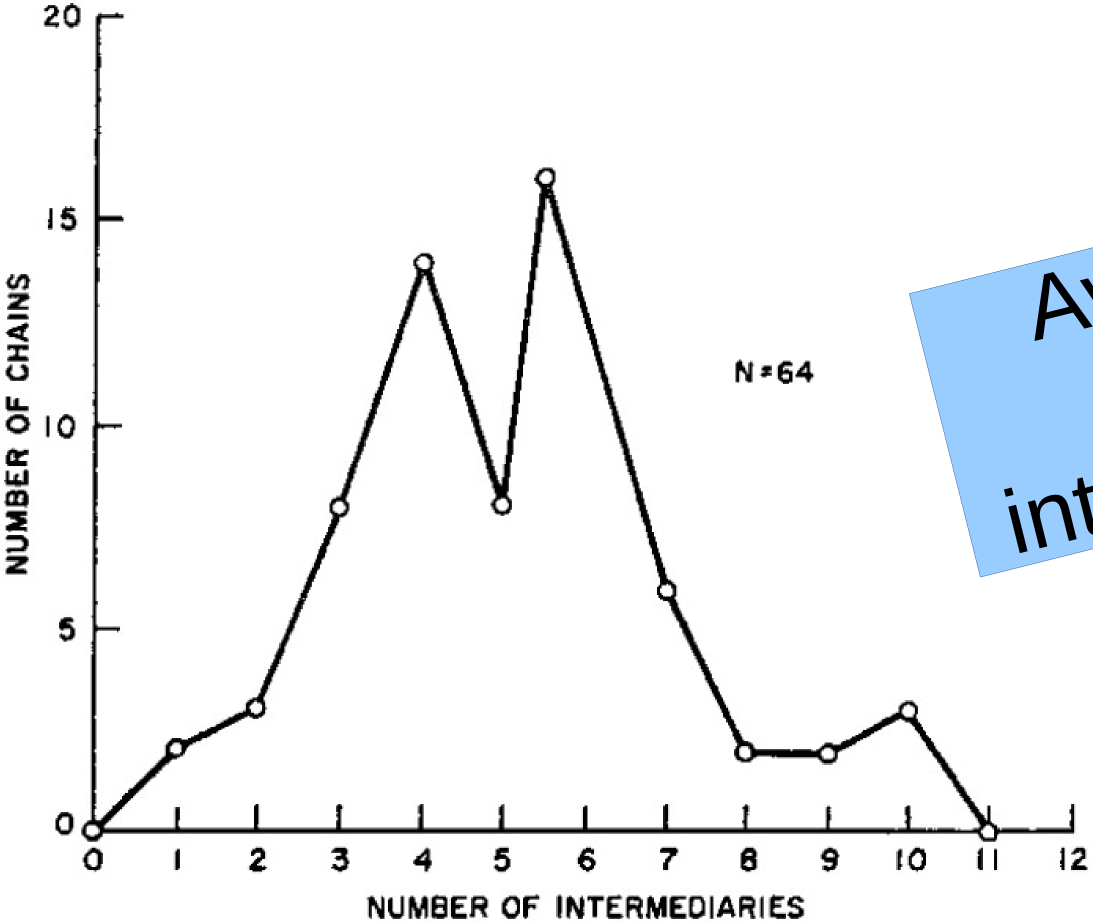


FIGURE 1

Lengths of Completed Chains

Average between 5.5 and 6 intermediate persons

2003 "Small World" Project

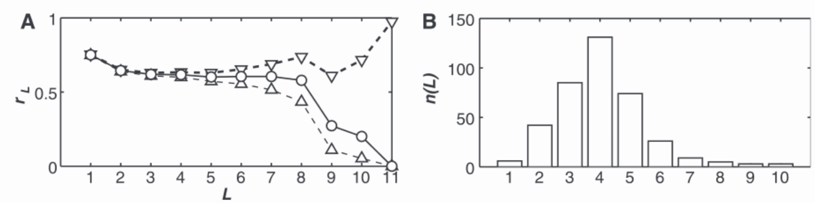
- More than 20.000 chains of emails to 18 persons of 13 countries

Median estimated between 5 and 7

An Experimental Study of Search in Global Social Networks

Peter Sheridan Dodds,[1] Roby Muhamad,[2] Duncan J. Watts[1,2]*

We report on a global social-search experiment in which more than 60,000 e-mail users attempted to reach one of 18 target persons in 13 countries by forwarding messages to acquaintances. We find that successful social search is conducted primarily through intermediate to weak strength ties, does not require highly connected "hubs" to succeed, and, in contrast to unsuccessful social search, disproportionately relies on professional relationships. By accounting for the attrition of message chains, we estimate that social searches can reach their targets in a median of five to seven steps, depending on the separation of source and target, although small variations in chain lengths and participation rates generate large differences in target reachability. We conclude that although global social networks are, in principle, searchable, actual success depends sensitively on individual incentives.



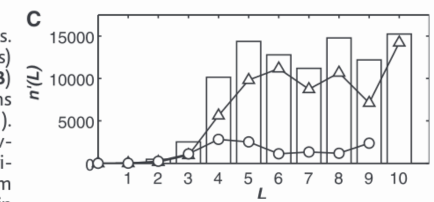


Fig. 1. Distributions of message chain lengths. **(A)** Average per-step attrition rates (circles) and 95% confidence interval (triangles). **(B)** Histogram representing the number of chains that are completed in L steps ($\langle L \rangle = 4.01$). **(C)** "Ideal" histogram of chain lengths recovered from (B) by accounting for message attrition (A). Bars represent the ideal histogram recovered with average values of r [circles in (A)] for the histogram in (B); lines represent a decomposition of the complete data into chains that start in the same country as the target (circles) and those that start in a different country (triangles).

2008 Microsoft Messenger

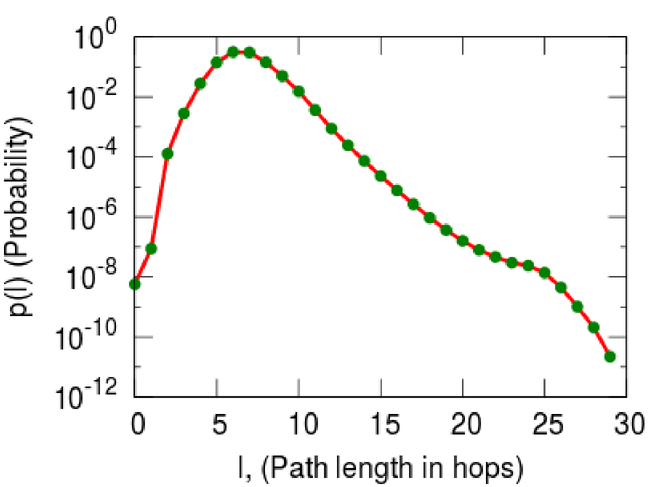
- 30 billion conversations between 240 million persons

Planetary-Scale Views on a Large Instant-Messaging Network

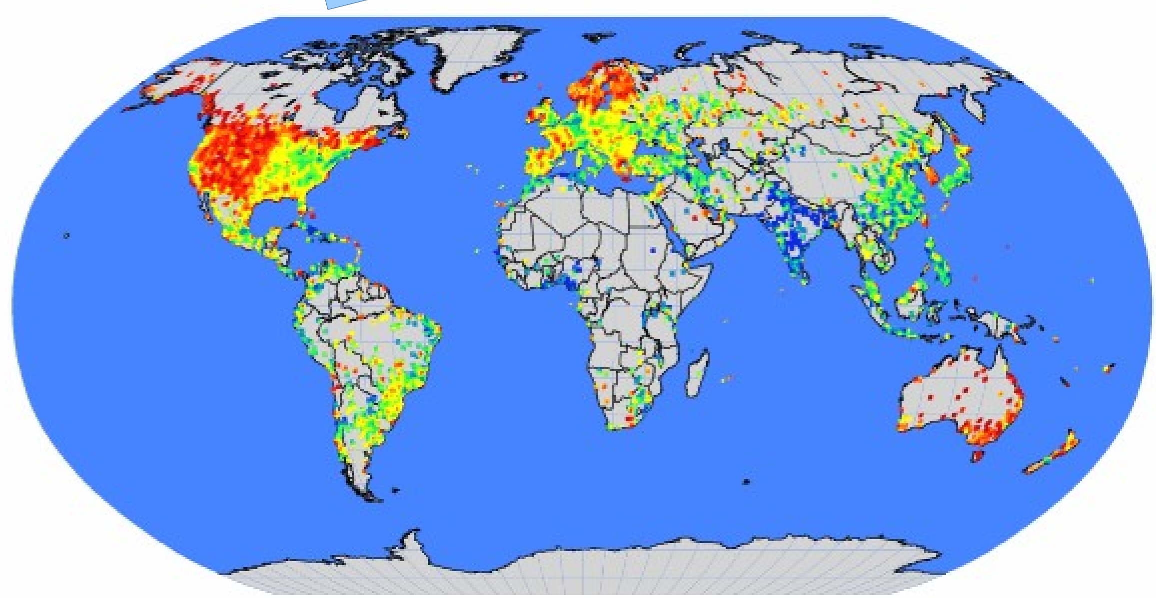
Jure Leskovec*
Carnegie Mellon University
jure@cs.cmu.edu

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

Global Average: 5.6

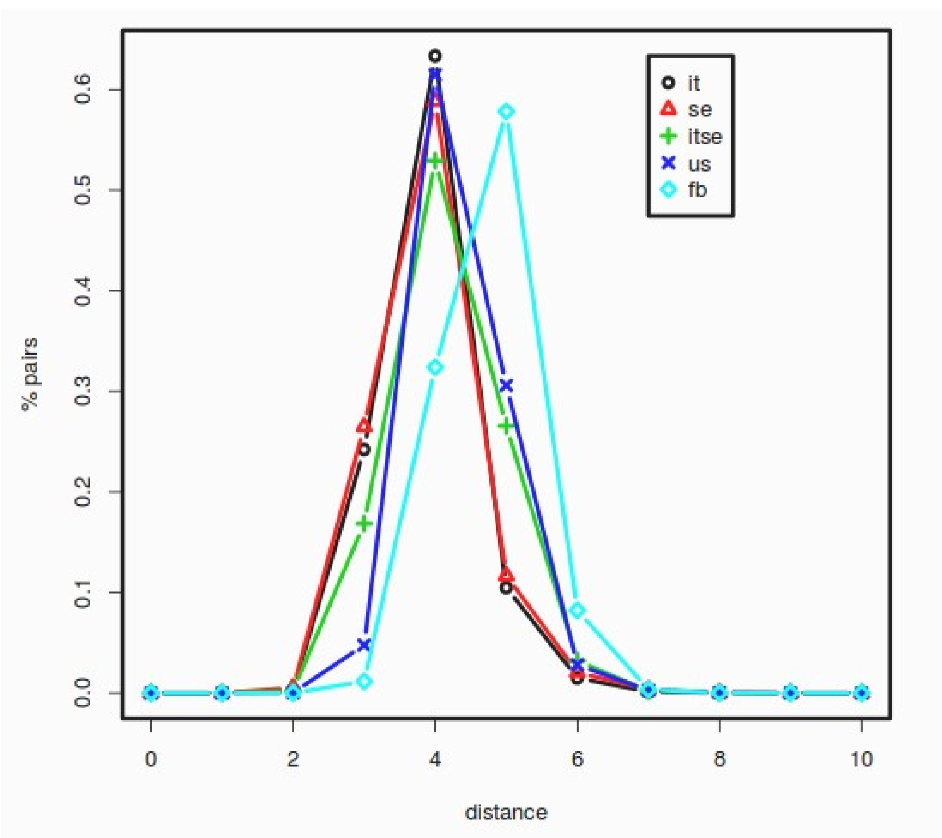


(a) Diameter



2011 Facebook Friendships

- 69 billions of friendships between 721 millions of persons



Global average: 3.74

Computer Science > Social and Information Networks

Four Degrees of Separation

Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, Sebastiano Vigna

(Submitted on 19 Nov 2011 (v1), last revised 5 Jan 2012 (this version, v3))

Frigyes Karinthy, in his 1929 short story "L\'aancszemek" ("Chains") suggested that any two persons are distanc
individuals, one of whom is a personal acquaintance, he could contact the selected individual [...]". It is not comp
graph theory, but the "six degrees of separation" phrase stuck after John Guare's 1990 eponymous play. Followi
one", where "distance" is the usual path length-the number of arcs in the path.) Stanley Milgram in his famous e
average number of intermediaries on the path of the postcards lay between 4.4 and 5.7, depending on the samp
We report the results of the first world-scale social-network graph-distance computations, using the entire Faceb
corresponding to 3.74 intermediaries or "degrees of separation", showing that the world is even smaller than we
interesting geographic subgraphs, looking also at their evolution over time.
The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previ
very accurate.

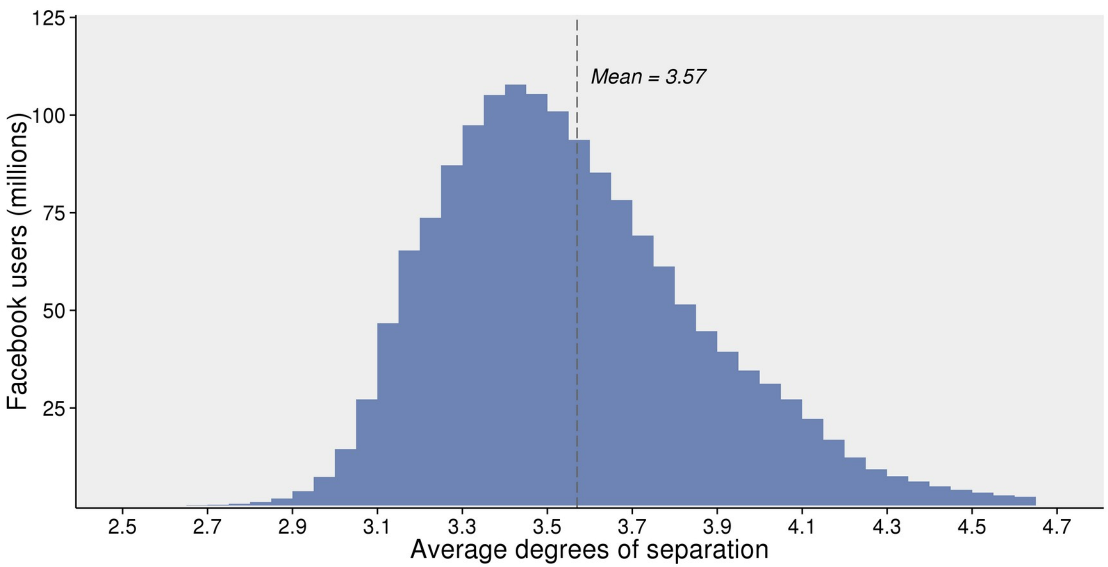
2016 Facebook Friendships

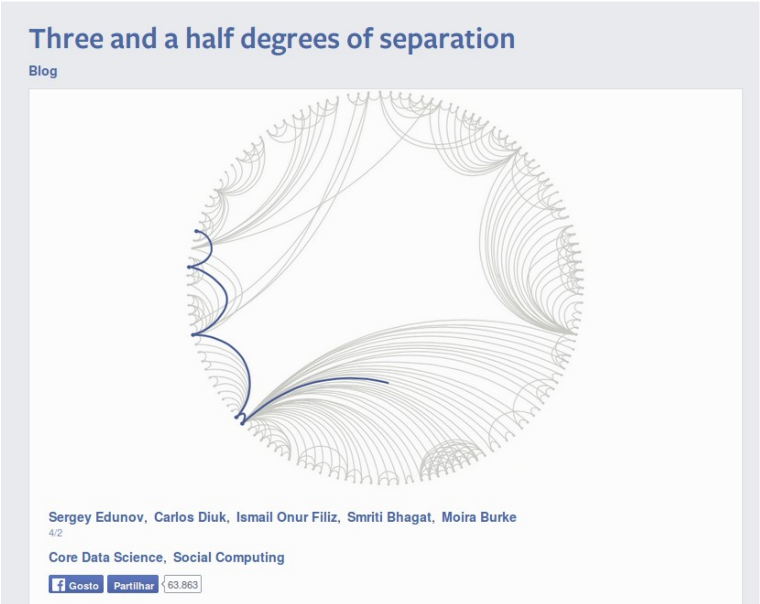
- 1.59 billions of persons

Global average: 3.57

My degrees of separation

Pedro Ribeiro's average degrees of separation from everyone is **3.43**.





How to explain this?

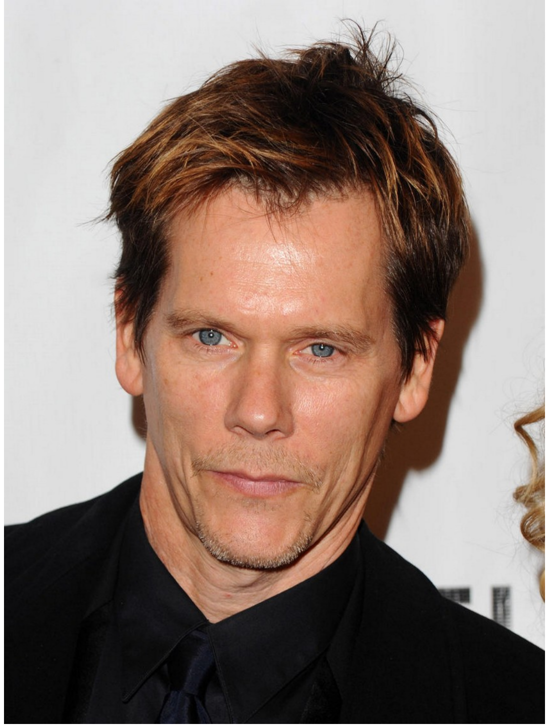
- Imagine that a person has, on average, 100 friends
  - 0 intermediates: 100
  - 1 intermediate: $100^2 = 10.000$
  - 2 intermediates: $100^3 = 1.000.000$
  - 3 intermediates: $100^4 = 100.000.000$
  - 4 intermediates: $100^5 = 10.000.000.000$
  - 5 intermediates: $100^6 = 1.000.000.000.000$

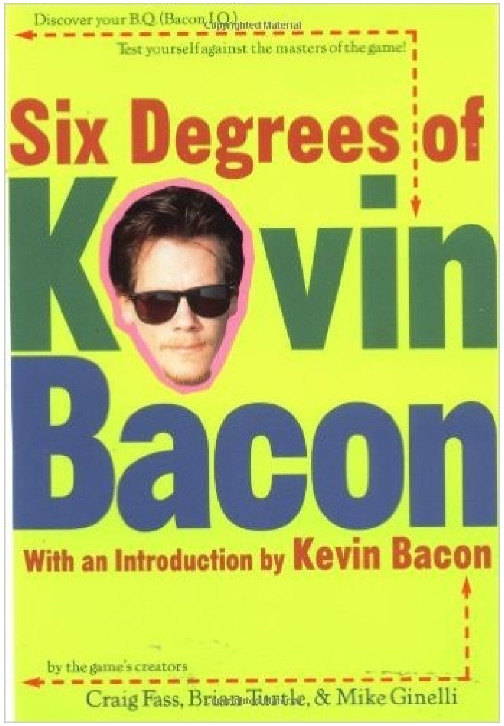
The power of exponentiation

- In practice, not all friends are new, but still there is a very fast growth

More examples of “Small World”

- The six degrees of Kevin Bacon
  - How many connections to link Kevin Bacon to any other actor, director, producer...
  - “Game” initiated in 1994





Joaquim de Almeida

Bret Cullen

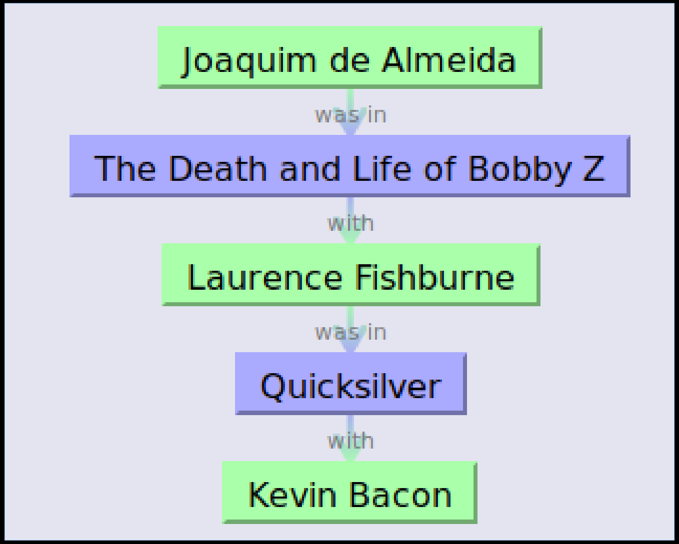
Joaquim de Almeida

(Bacon Number: 2)

Kevin Bacon

Joaquim de Almeida

(Bacon Number: 2)



Nicolau Breyner

Nicolau Breyner

(Bacon Number: 3)

Marilyn Monroe

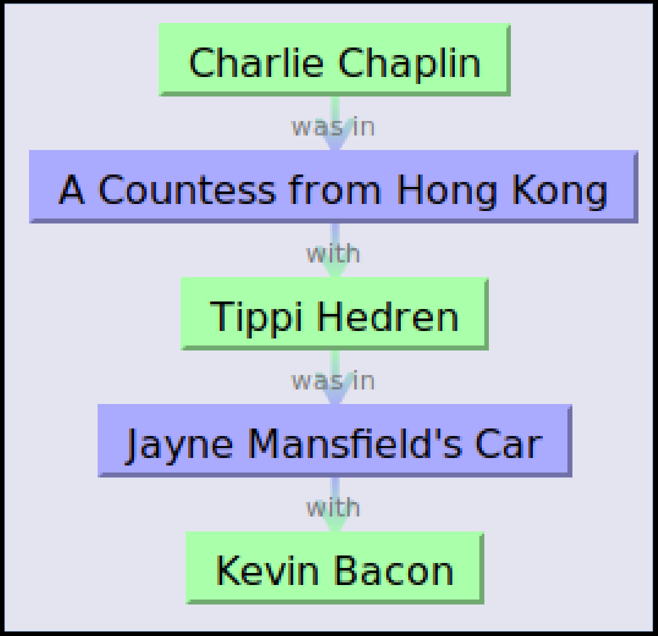
Charlie Chaplin

Marilyn Monroe

(Bacon Number: 2)

Charlie Chaplin

(Bacon Number: 2)



More examples of “Small World”

- The six degrees of Kevin Bacon

https://oracleofbacon.org/

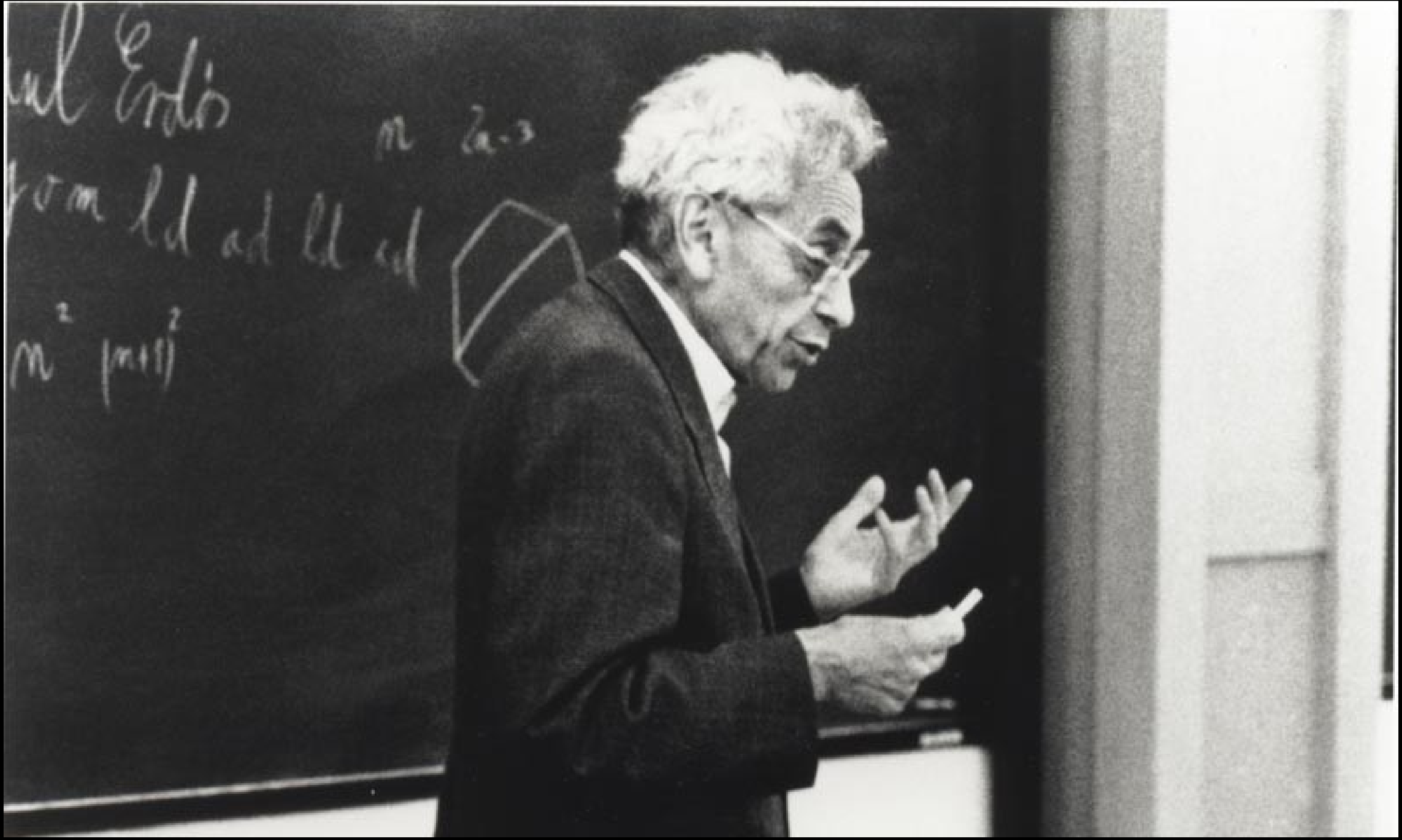
(average number: 3.009)

Kevin Bacon Number	# of persons
0	1
1	3150
2	373876
3	1340703
4	340756
5	28820
6	3383
7	451
8	52
9	8
10	1

“People would start to come up to me in the subway and literally go…”

“Zero! Zero! Zero! Zero!”

More examples of “Small World”



Paul Erdös

Erdös Number

More examples of “Small World”

- Erdös number:
  - Scientific articles and very prolific mathematician

http://wwwp.oakland.edu/enp/

Home | Preferences | Free Tools

AMERICAN MATHEMATICAL SOCIETY
MATHSCINET®
MATHEMATICAL REVIEWS

Search MSC | Collaboration Distance | Current Journals | Current Publications

MR Erdos Number = 4

Pedro Ribeiro	coauthored with	Srinivasan Parthasarathy[1]	MR3385657
Srinivasan Parthasarathy[1]	coauthored with	Yusu Wang	MR3685725
Yusu Wang	coauthored with	Boris Aronov	MR2347131
Boris Aronov	coauthored with	Paul Erdős[1]	MR1289067

Change First Author **Change Second Author** **New Search**

Emergence of Network Science

Complexity

"I think the next century will be the century of **complexity**"

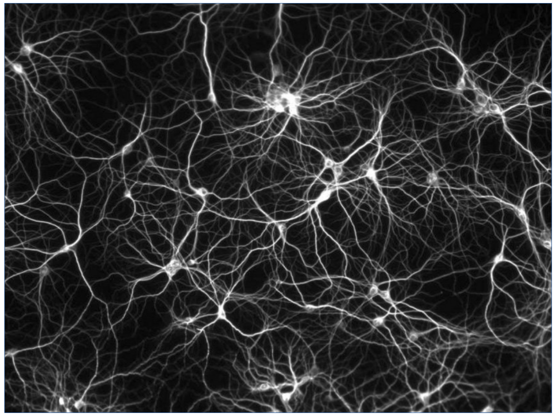
Stephen Hawking (Jan, 2000)

The Real World is Complex

World Population: 8 billions

The Real World is Complex

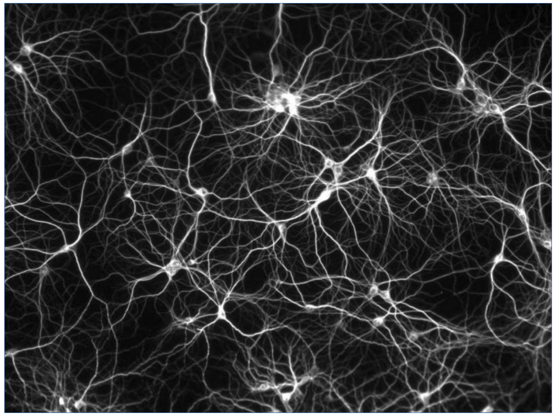
World Population: **8 billions**



Human Brain Neurons:
100 billions

The Real World is Complex

World Population: 8 billions

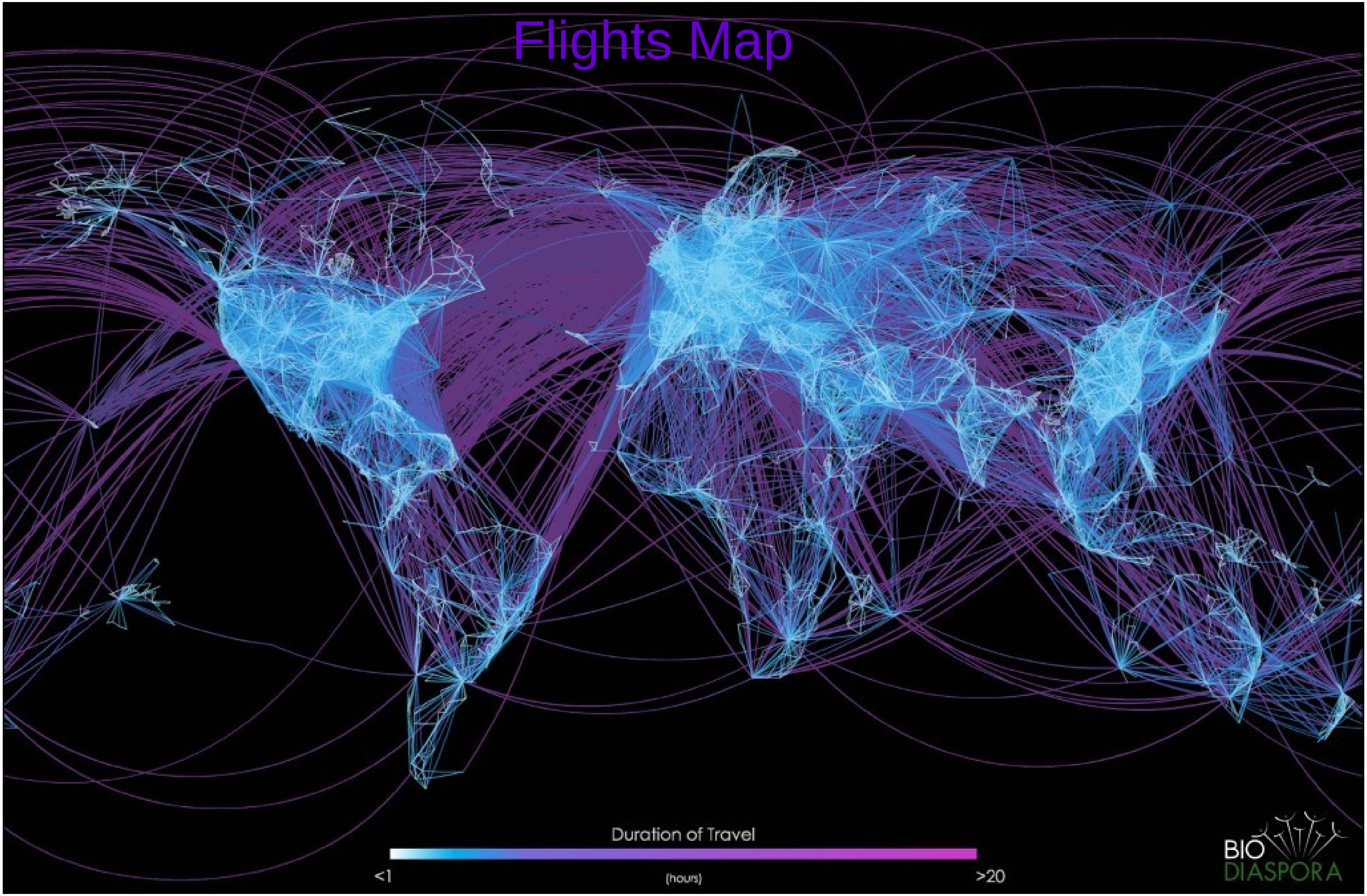


Human Brain Neurons: 100 billions

Internet Devices: >10 billions

Complex Systems → Complex Networks

Flights Map

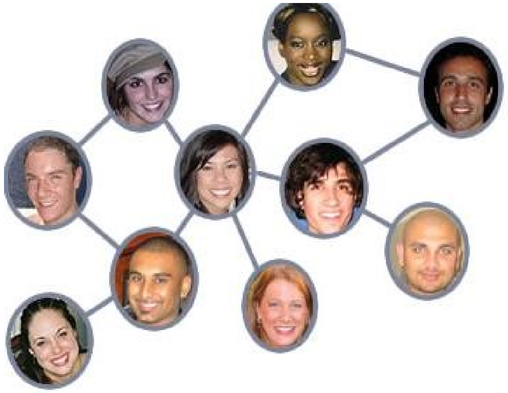


Complex Networks are Ubiquitous

Social

Complex Networks are Ubiquitous

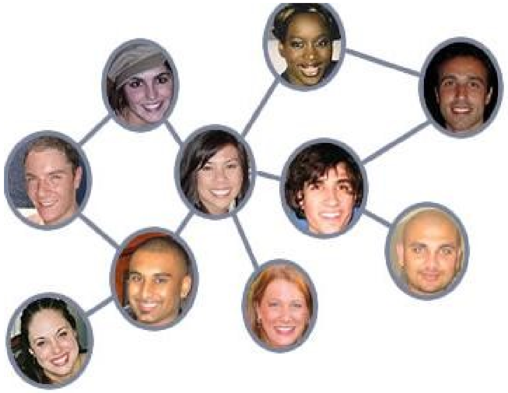
Social



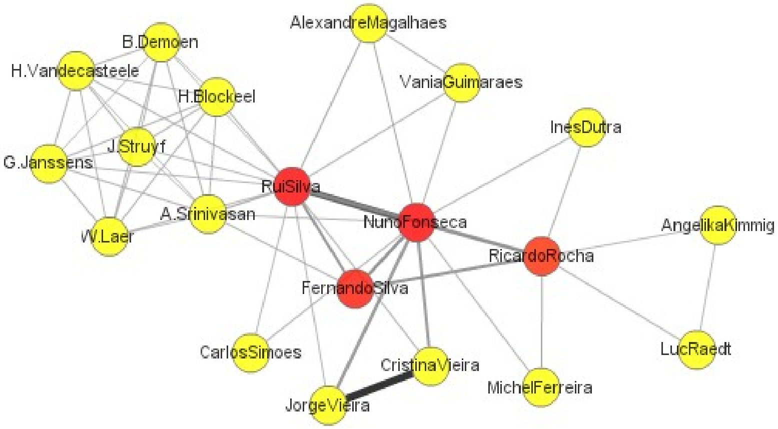
Facebook

Complex Networks are Ubiquitous

Social



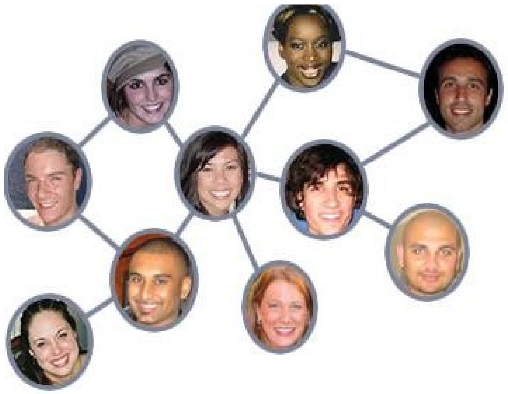
Facebook



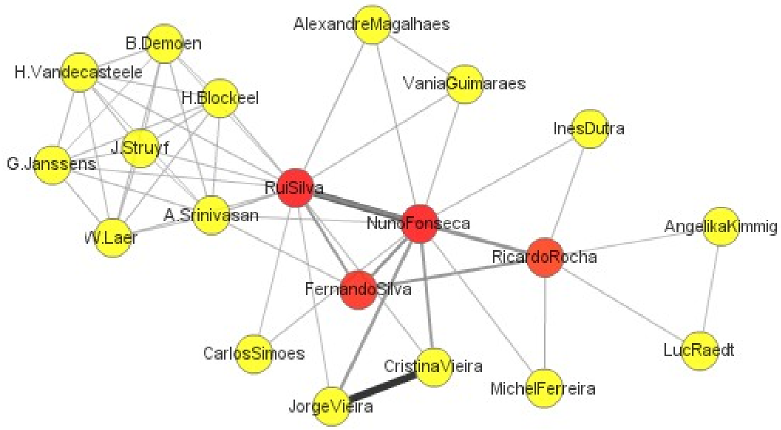
Co-authorship

Complex Networks are Ubiquitous

Social



Facebook

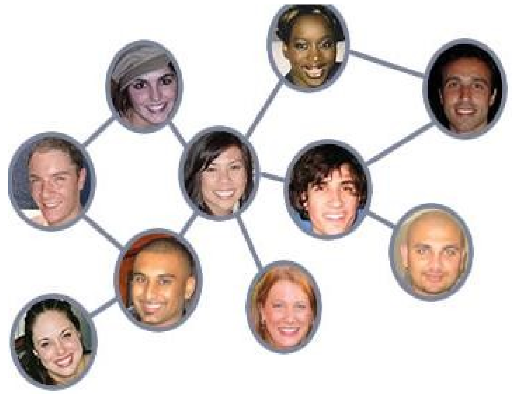


Co-authorship

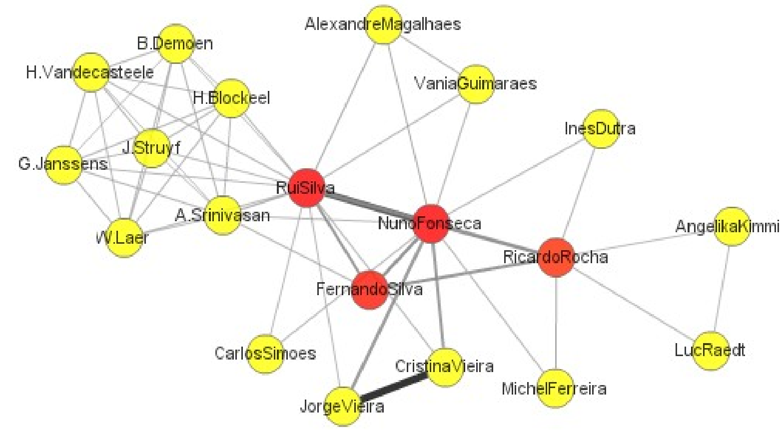
Biological

Complex Networks are Ubiquitous

Social

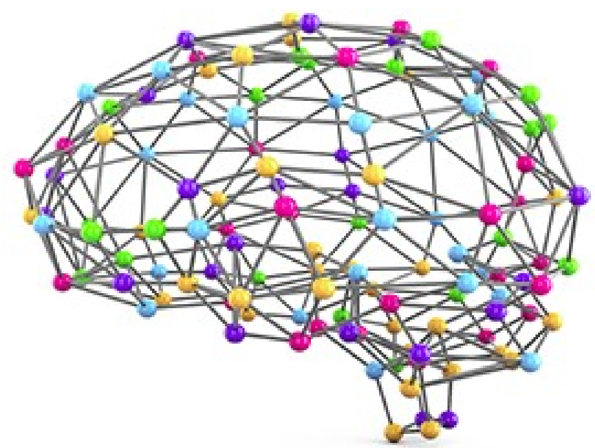


Facebook



Co-authorship

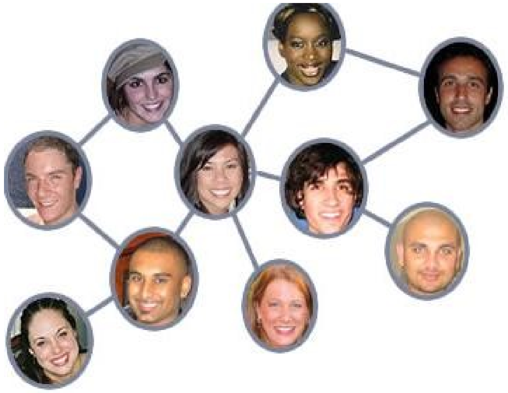
Biological



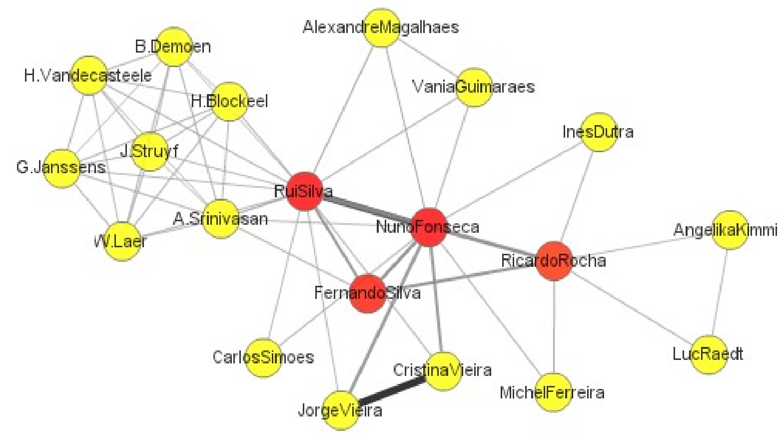
Brain

Complex Networks are Ubiquitous

Social

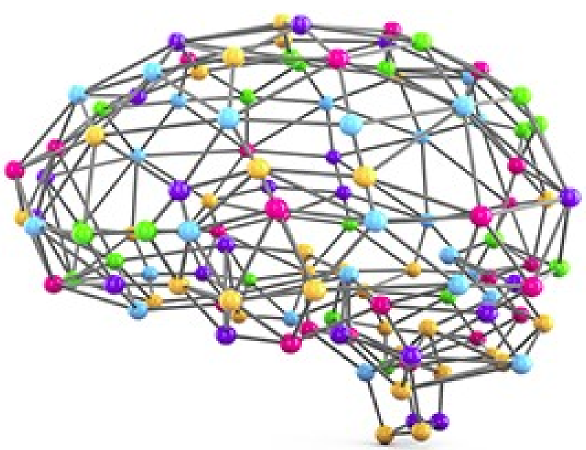


Facebook

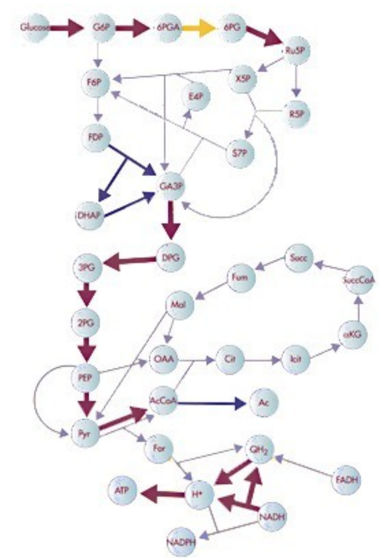


Co-authorship

Biological



Brain



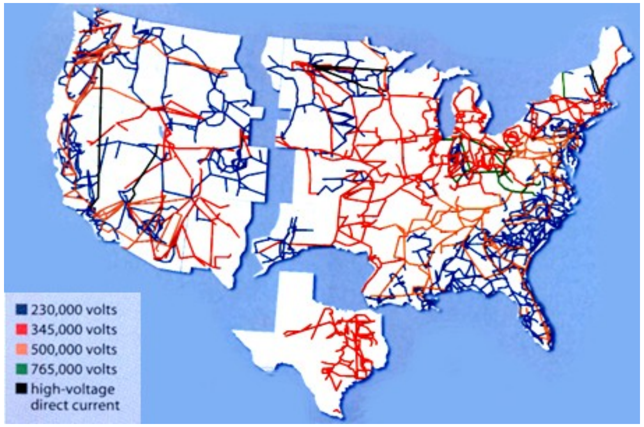
Metabolism (proteins)

Complex Networks are Ubiquitous

Spatial

Complex Networks are Ubiquitous

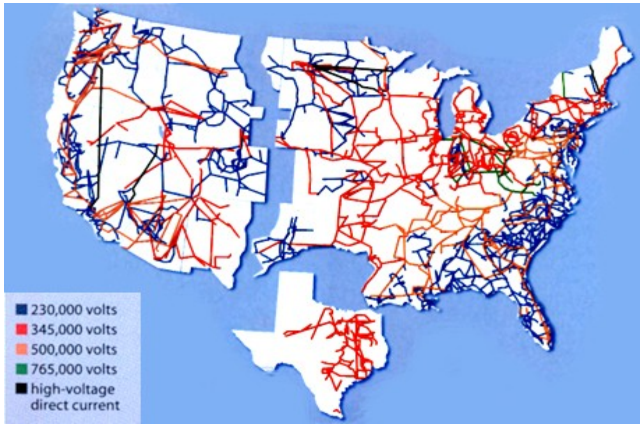
Spatial



Power

Complex Networks are Ubiquitous

Spatial

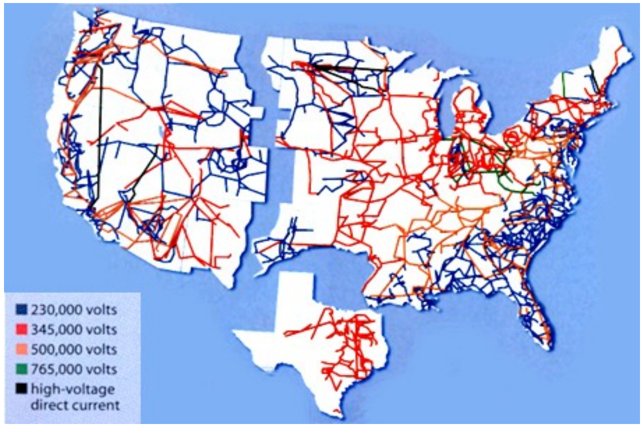


Power

Roads

Complex Networks are Ubiquitous

Spatial



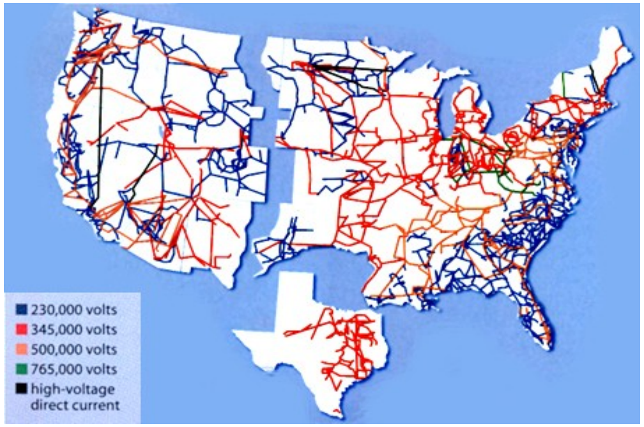
Power

Roads

Software

Complex Networks are Ubiquitous

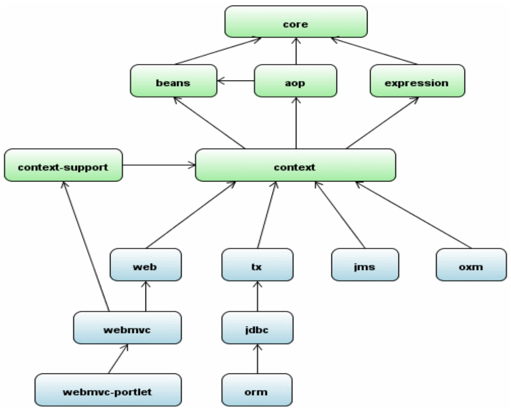
Spatial



Power

Roads

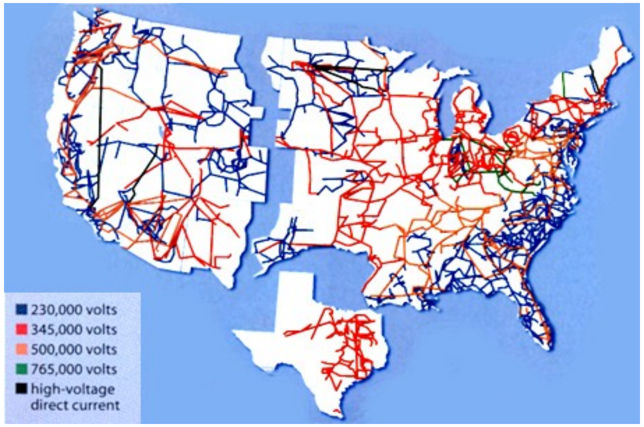
Software



Module Dependency

Complex Networks are Ubiquitous

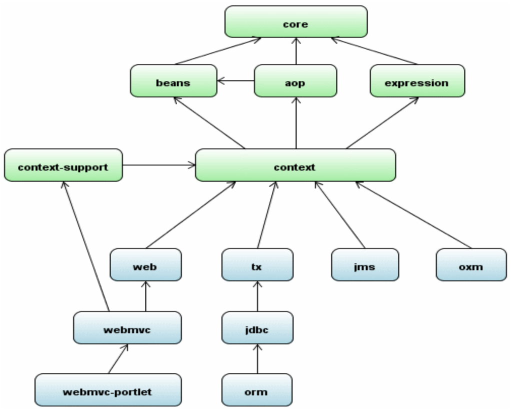
Spatial



Power

Roads

Software

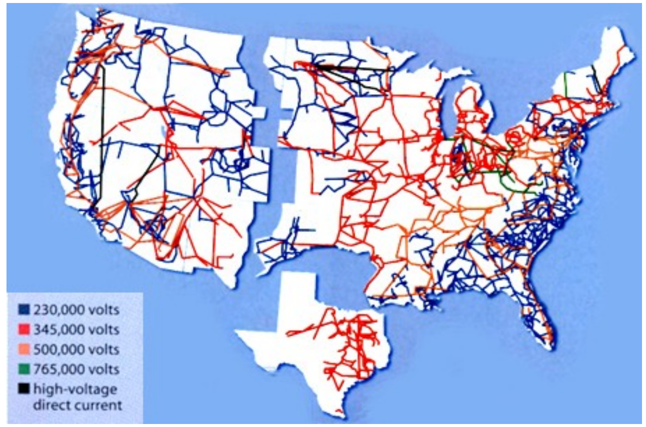


Module Dependency

Text

Complex Networks are Ubiquitous

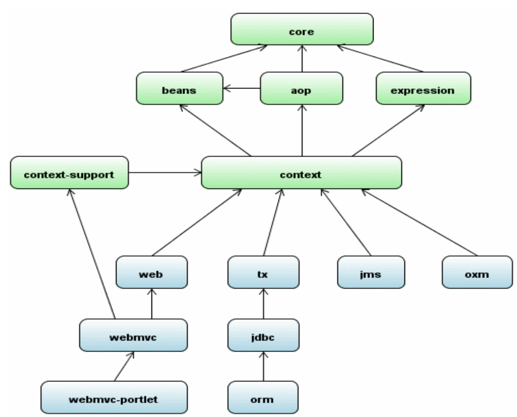
Spatial



Power

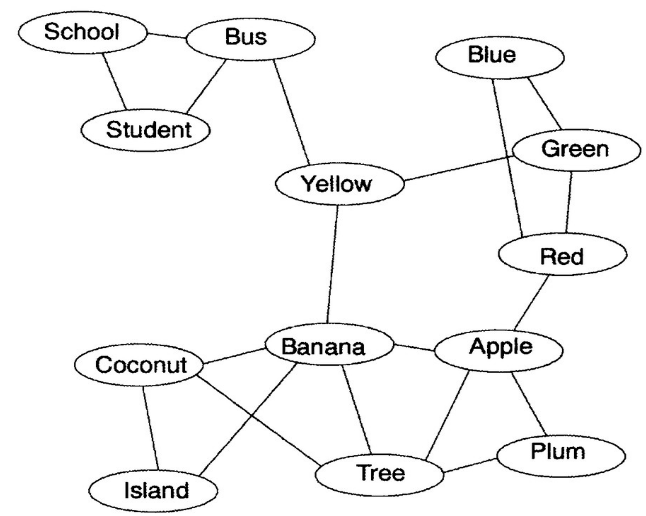
Roads

Software



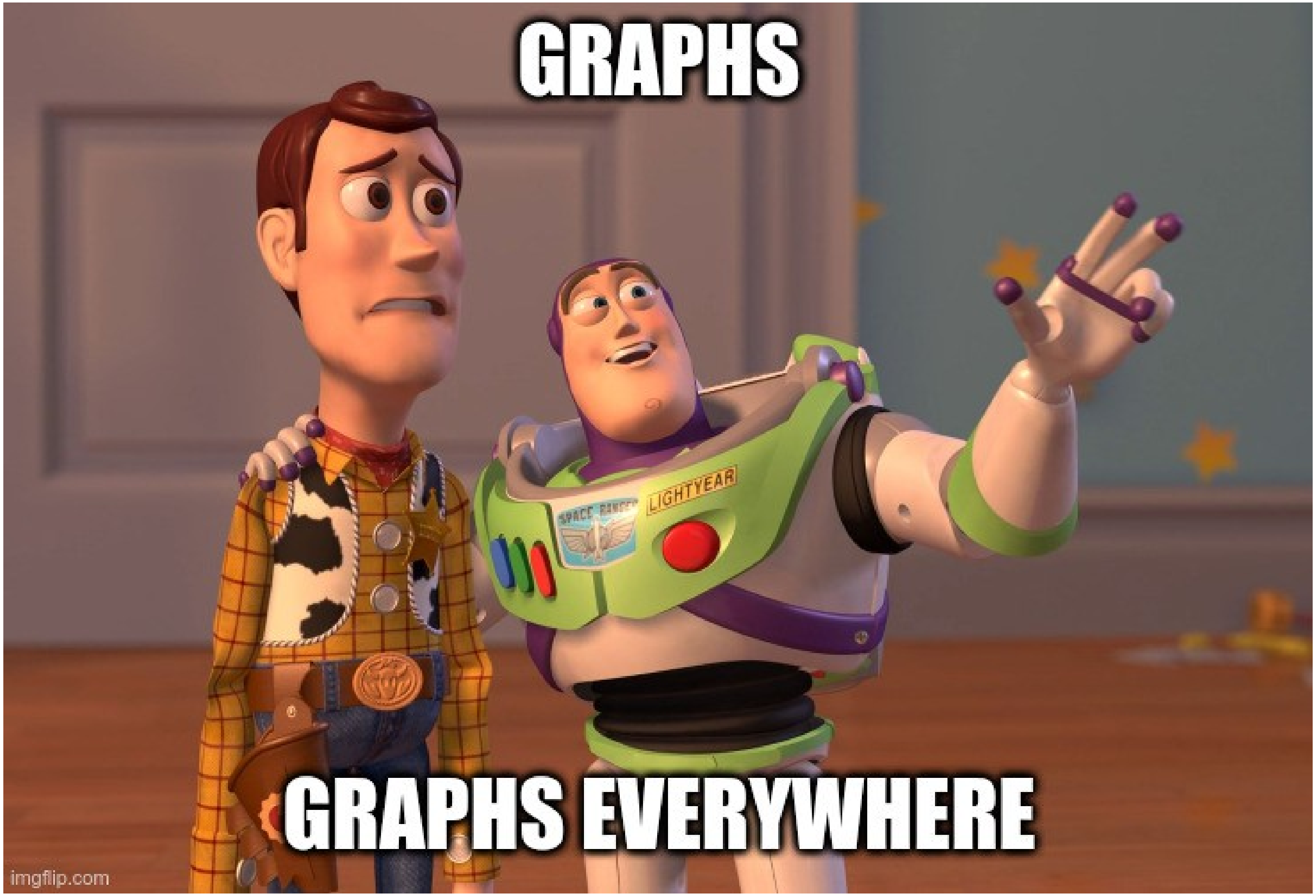
Module Dependency

Text



Semantic

Complex Networks are Ubiquitous



Network Science

Behind many complex systems there is a **network** that defines the **interactions** between the components

In order to understand the systems... we need to understand the **networks**!

Network Science

- **Network Science** has been emerging on this century as a new discipline:
  - Origins on **graph theory** and **social** network research

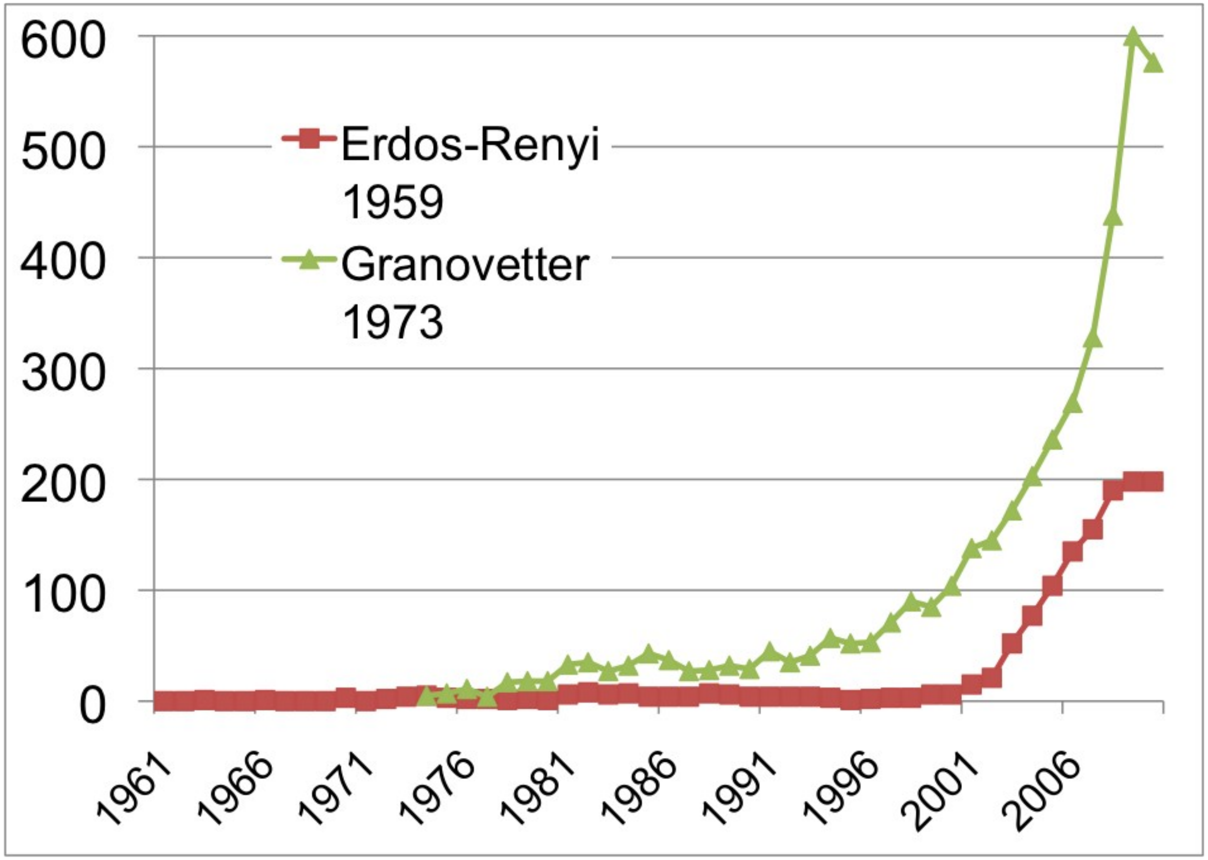


Image: Adapted from (Barabasi, 2015)

Why now?

- Two main contributing factors:

Why now?

- Two main contributing factors:

1) The emergence of **network maps**

Why now?

- Two main contributing factors:

1) The emergence of **network maps**

- Movie actor network: 1998
- World Wide Web: 1999
- Citation Network: 1998
- Metabolic Network: 2000
- PPI Network: 2001

Why now?

- Two main contributing factors:

1) The emergence of **network maps**

- Movie actor network: 1998
- World Wide Web: 1999
- Citation Network: 1998
- Metabolic Network: 2000
- PPI Network: 2001

- **436 nodes** – 2003
(email exchange, Adamic-Adar, SocNets)
- **43,553 nodes** – 2006
(email exchange, Kossinets-Watts, Science)
- **4.4 million nodes** – 2005
(friendships, Liben-Nowell, PNAS)
- **800 million nodes** – 2011
(Facebook, Backstrom et al.)

Size matters!

Why now?

- Two main contributing factors:

2) Universality of network **characteristics**

Why now?

- Two main contributing factors:

2) Universality of network **characteristics**

- The architecture and topology of networks from different domains exhibit more similarities that what one would expect

Why now?

- Two main contributing factors:

2) Universality of network **characteristics**

- The architecture and topology of networks from different domains exhibit more similarities that what one would

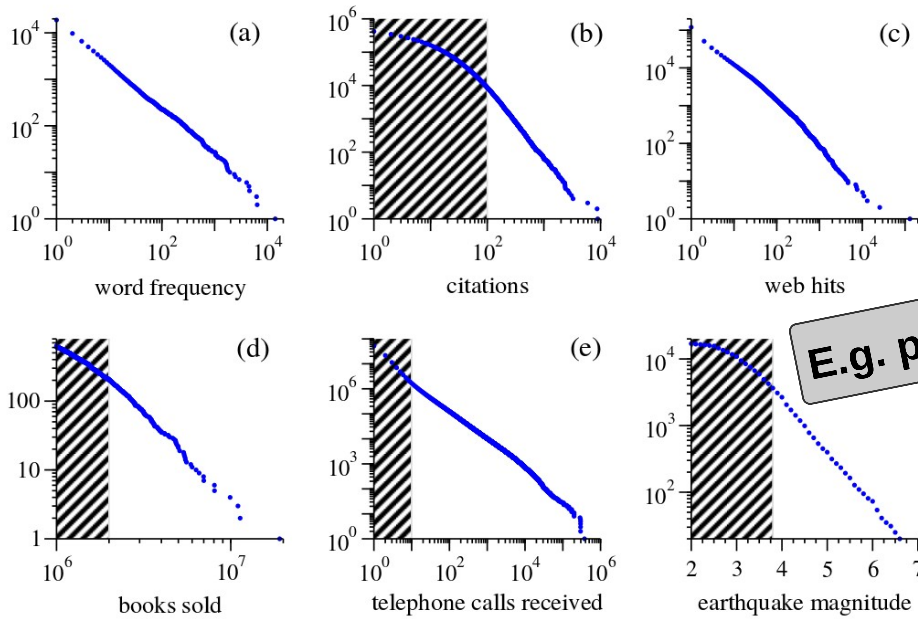


Image: Adapted from (Newman, 2005)

E.g. power laws

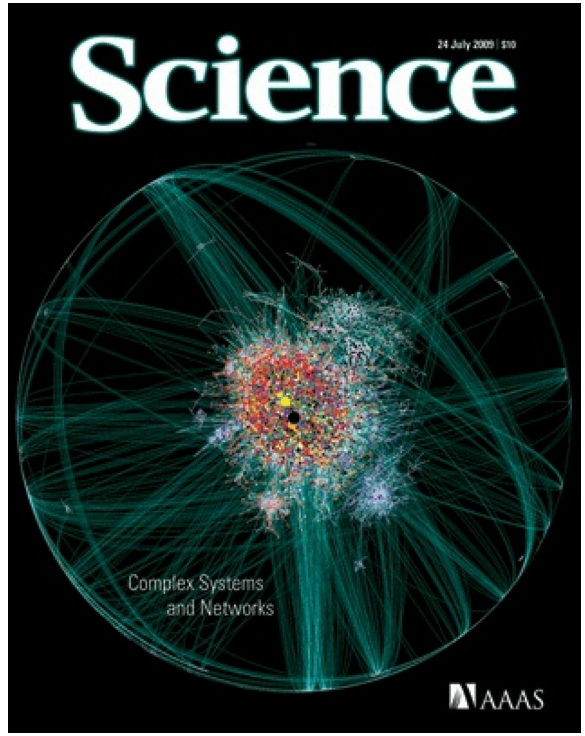
Many real world networks are power law

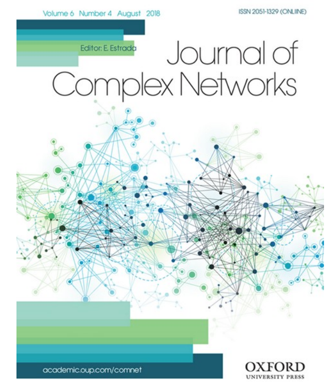
	exponent α (in/out degree)
film actors	2.3
telephone call graph	2.1
email networks	1.5/2.0
sexual contacts	3.2
WWW	2.3/2.7
internet	2.5
peer-to-peer	2.1
metabolic network	2.2
protein interactions	2.4

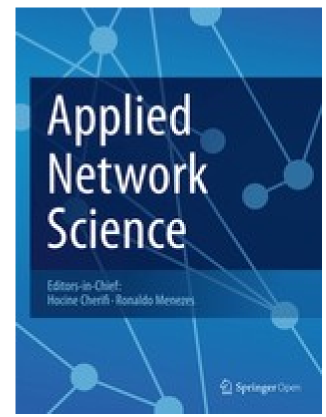
Image: Adapted from Leskovec, 2015

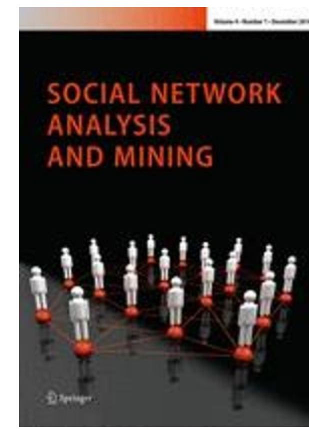
Impact of Network Science: Economic

Impact of Network Science: Scientific









netsi

Northeastern University
Network Science Institute

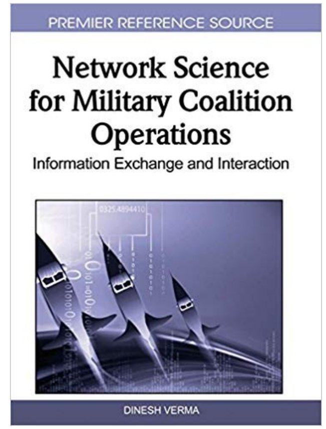
GEMS LAB
Graph Exploration & Mining at Scale

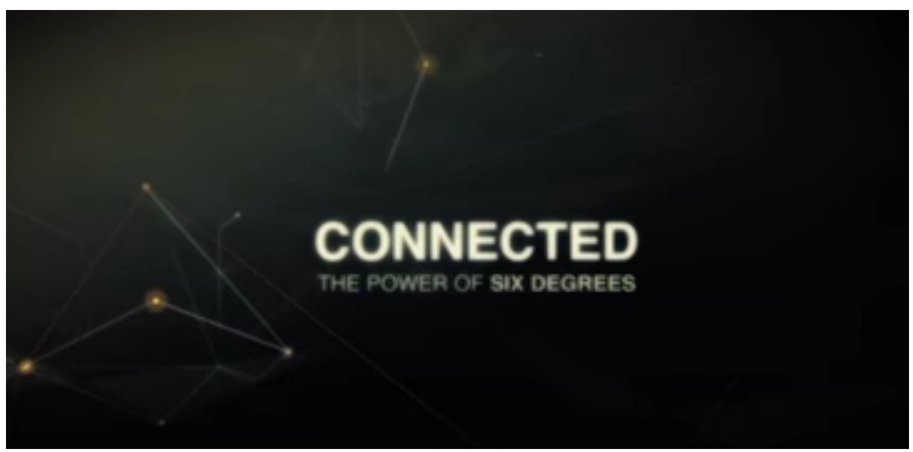
Porto
Winter School on
Network Science

Indiana University
Network Science Institute

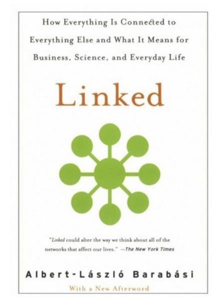
Impact of Network Science: Societal

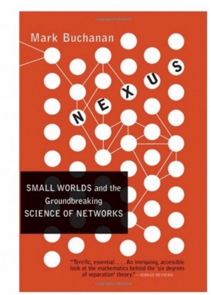
Network Science Center West Point

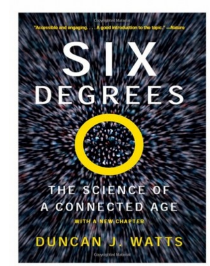


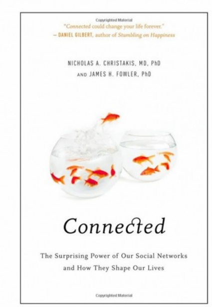


https://youtu.be/2rzxAyY7D7k









Reasoning about Networks

- **What do we hope to achieve from studying networks?**
  - Patterns and statistical **properties** of network data
  - **Design principles** and **models**
  - **Algorithms** and **predictive** models to answer questions and make predictions

Mining and Learning with Graphs

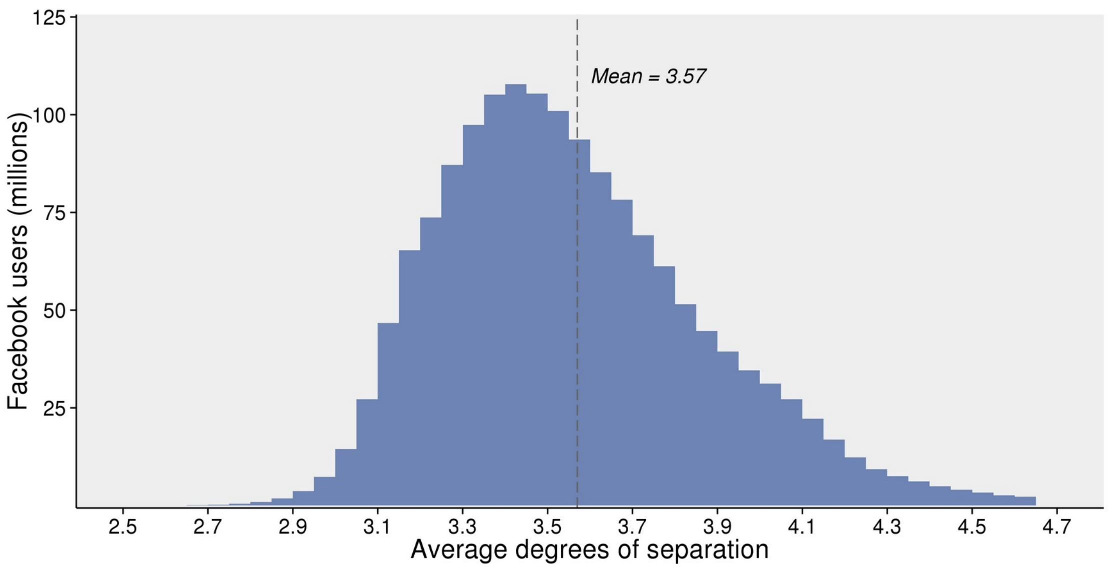
- **How do we mine networks?**
  - **Empirically:** Study network data to find organizational principles
    - How do we measure and quantify networks?
  - **Mathematical models:** Graph theory and statistical models
    - Models allow us to understand behaviors and distinguish surprising from expected phenomena
  - **Algorithms** for analyzing graphs
    - Hard computational challenges

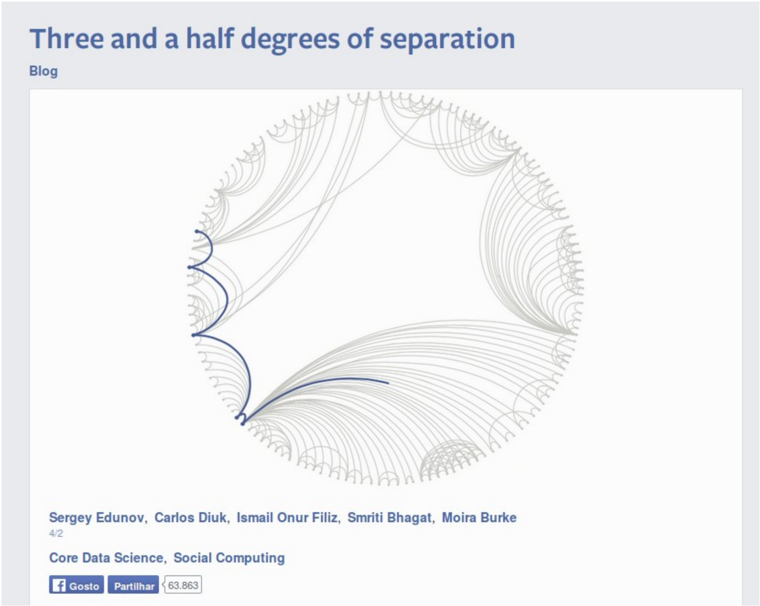
Network Science Topics

- Some possible tasks:

Network Science Topics

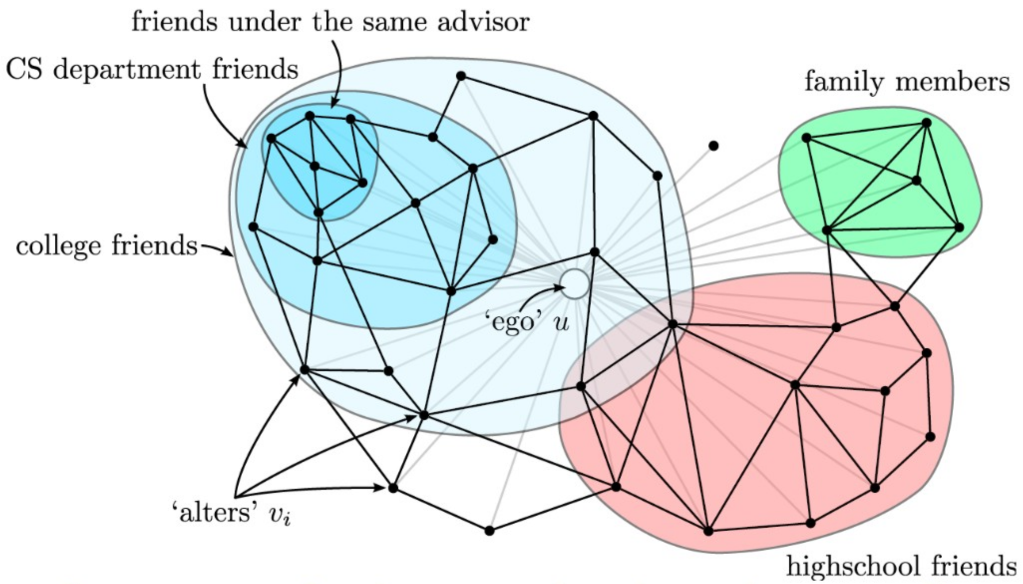
- Some possible tasks:
  - General Patterns
    - Ex: “scale-free”, “small-world”





Network Science Topics

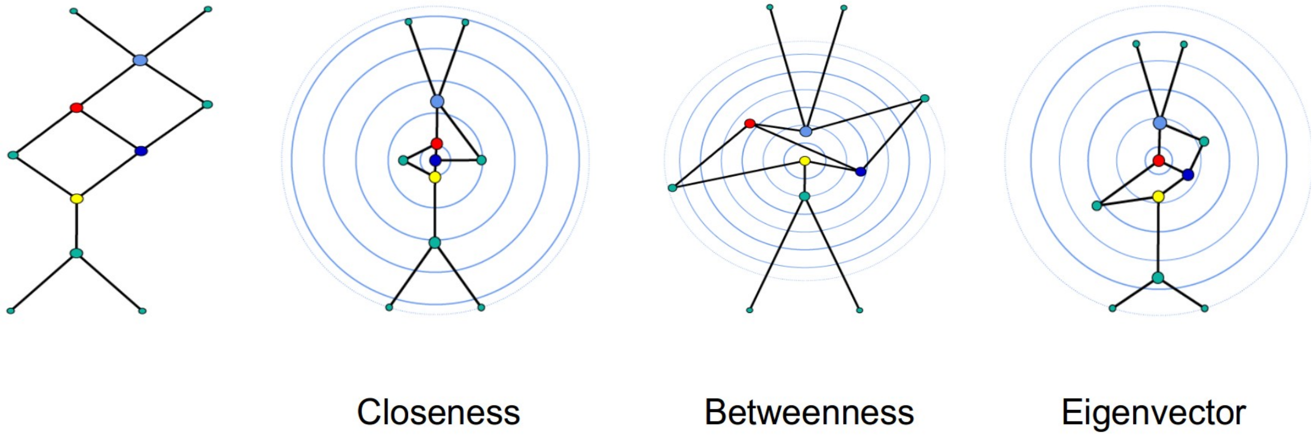
- Some possible tasks:
  - General Patterns
    - Ex: “scale-free”, “small-world”
  - Community Detection
    - What groups of nodes are “related”?



Discover circles and why they exist

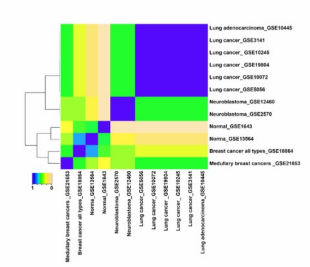
Network Science Topics

- Some possible tasks:
  - General Patterns
    - Ex: “scale-free”, “small-world”
  - Community Detection
    - What groups of nodes are “related”?
  - Node Classification
    - Importance and function of a certain node?



Network Science Topics

- Some possible tasks:
  - General Patterns
    - Ex: “scale-free”, “small-world”
  - Community Detection
    - What groups of nodes are “related”?
  - Node Classification
    - Importance and function of a certain node?
  - Network Comparison
    - What is the type of the network?



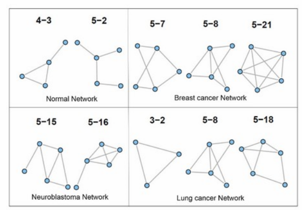
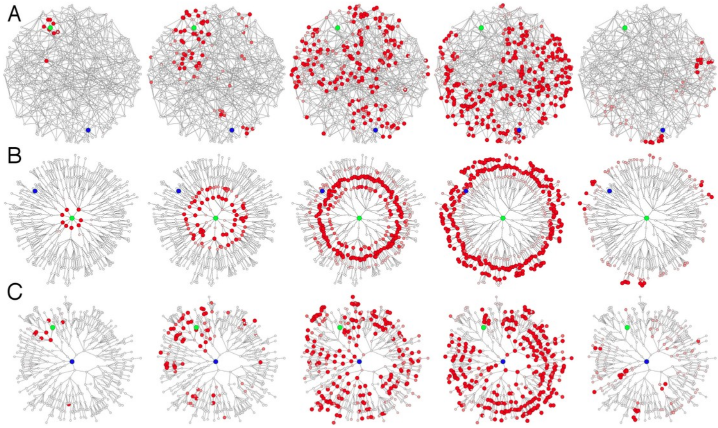


Figure 5: Discriminating subgraphs for each type of networks.

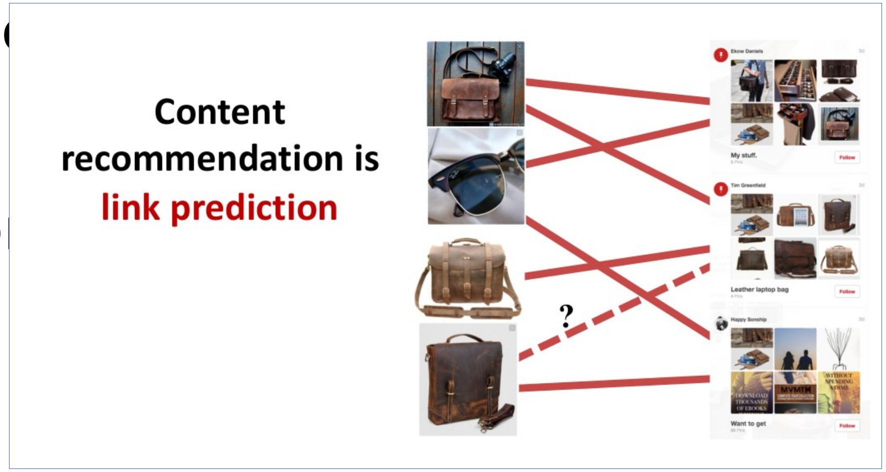
Network Science Topics

- Some possible tasks:
  - General Patterns
    - Ex: “scale-free”, “small-world”
  - Community Detection
    - What groups of nodes are “related”?
  - Node Classification
    - Importance and function of a certain node?
  - Network Comparison
    - What is the type of the network?
  - Information Propagation
    - Epidemics? Robustness?



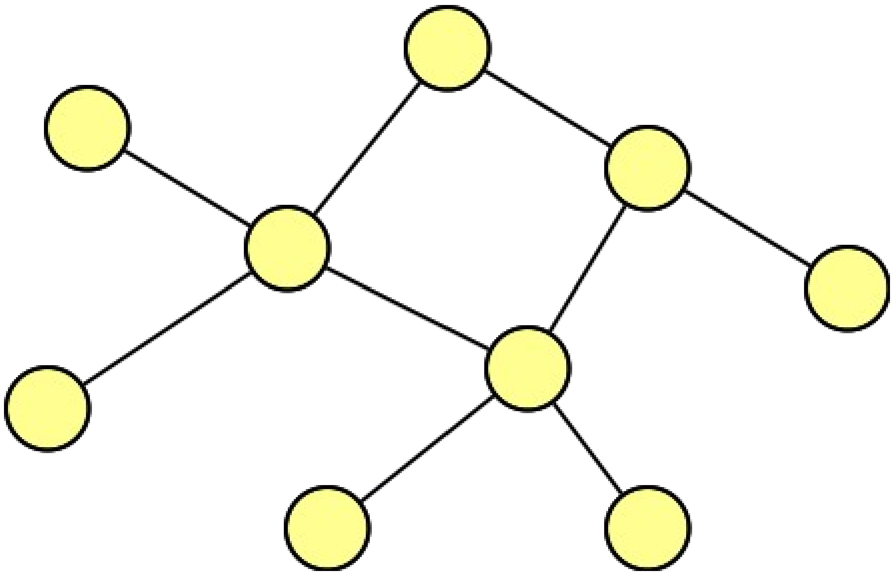
Network Science Topics

- Some possible tasks:
  - General Patterns
    - Ex: “scale-free”, “small-world”
  - Community Detection
    - What groups of nodes are “related”?
  - Node Classification
    - Importance and function of a
  - Network Comparison
    - What is the type of the netwo
  - Information Propagation
    - Epidemics? Robustness?
  - Link prediction
    - Future connections? Errors in graph constructions?



Brief Introduction to Graph Theory and Network Vocabulary

Terminology



- **Objects:** nodes, vertices ***V***
- **Interactions:** links, edges ***E***
- **System:** network, graph ***G(V,E)***

Networks or Graphs?

- **Network** often refers to real systems
  - Web, Social network, Metabolic network
  - Language: Network, node, link

- **Graph** is a mathematical representation of a network
  - Web graph, Social graph (a Facebook term)
  - Language: Graph, vertex, edge

We will try to make this distinction whenever it is appropriate, but **in most cases we will use the two terms interchangeably**

Choosing the Network

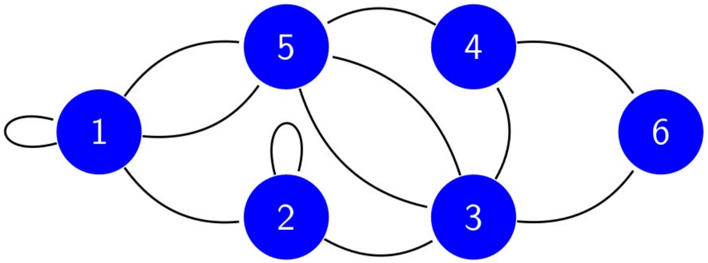
- If you connect individuals that work with each other, you will explore a **professional network**
- If you connect those that are friends, you will be exploring a **friendship network**
- If you connect scientific papers that cite each other, you will be studying the **citation network**

- **Another example:** if you connect all papers with the same word in the title, what will you be exploring?
- There might be **several possible representations**

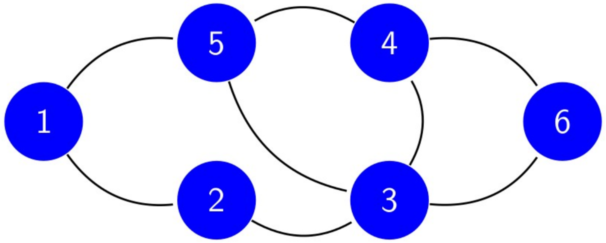
The choice of the network representation of a given domain determines our ability to use it successfully

Simple and multi-graphs

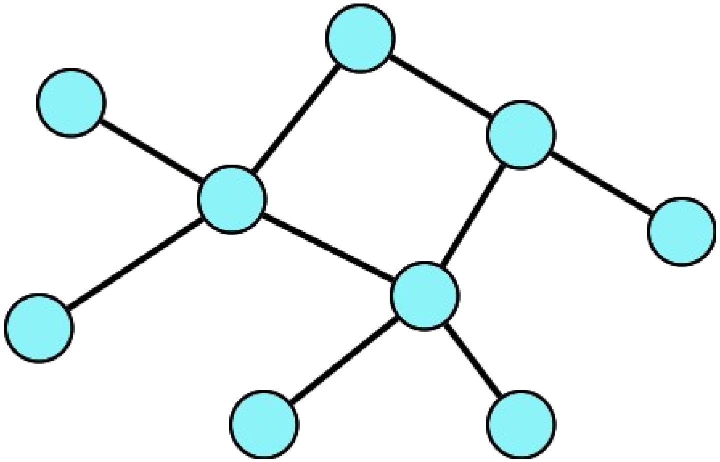
- **In general, graphs may have self-loops and multi-edges**
  - A graph with either is called a **multi-graph**

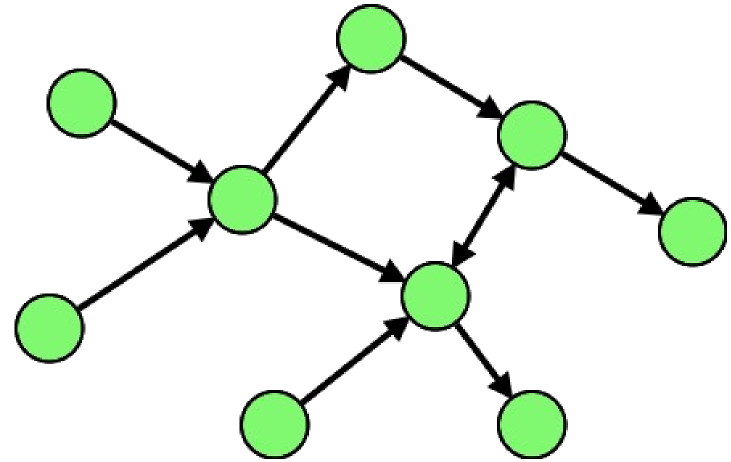


  - Today we will mostly work with **simple graphs**, with no self-loops or multi-edges



Network Types





Undirected	Directed
• co-authorship networks	• www hyperlinks
• actor networks	• phone calls
• facebook friendships	• roads network

Network Types

Edge Attributes

- Examples:
  - **Weight** (duration call, distance road, ...)
  - **Ranking** (best friend, second best friend, ...)
  - **Type** (friend, relative, co-worker, ...) [colored edges]
  - We can have a set of **multiple** attributes

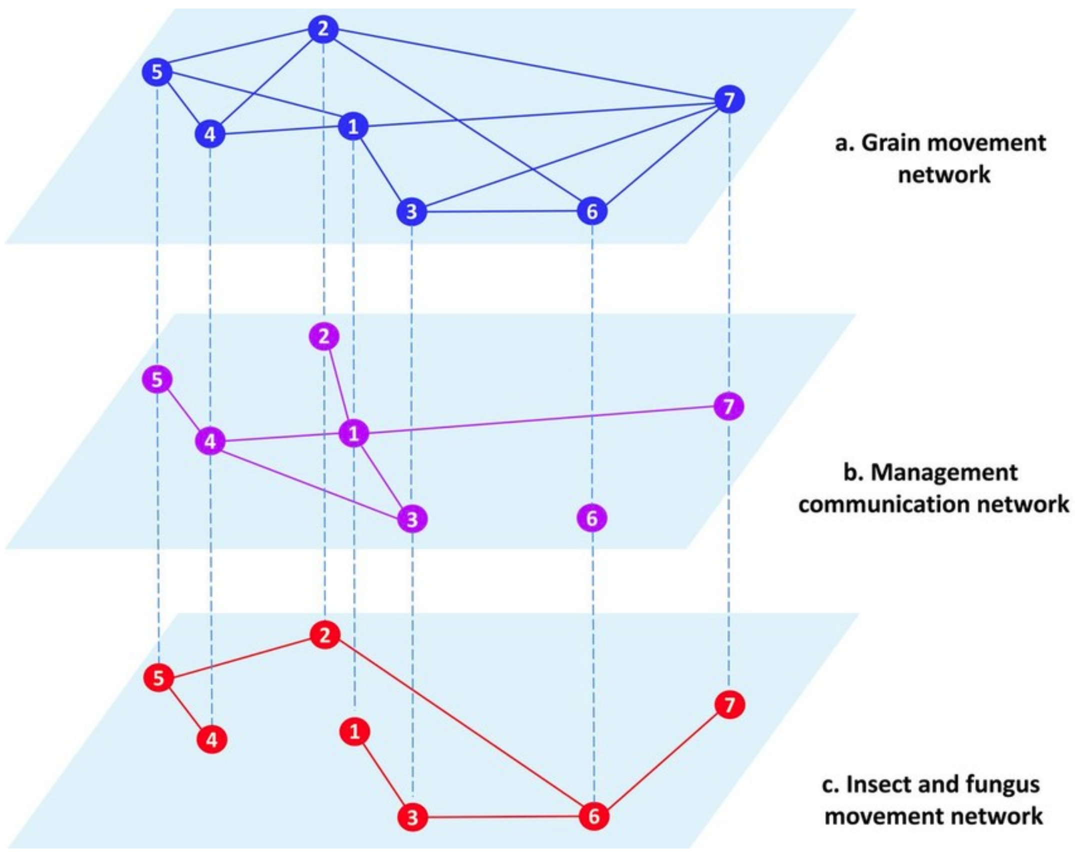
Node Attributes

- Examples:
  - **Type** (nationality, sex, age, ...) [colored nodes]
  - We can have a set of **multiple** attributes

Network Types

Multiplex Networks

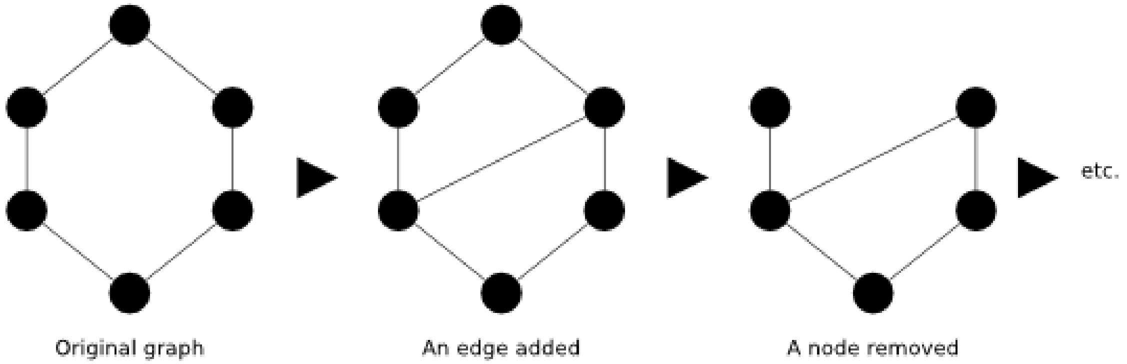
- Different layers (types) of connections



Network Types

Temporal Networks

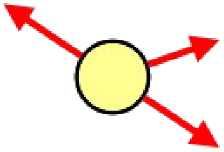
- Evolution over time



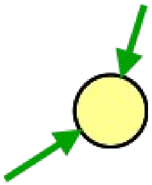
Node Properties

- **From immediate connections**

  - **Outdegree**
    how many directed edges originate at node

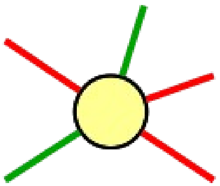


  - **Indegree**
    how many directed edges are incident on a node



  - **Degree** (in or out)

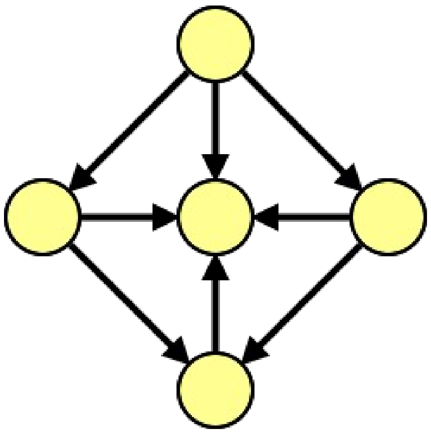
    number of outgoing and incoming edges



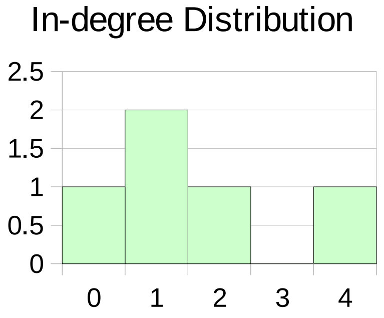
Node Properties

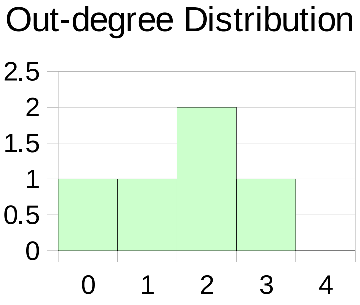
- **Degree related metrics:**

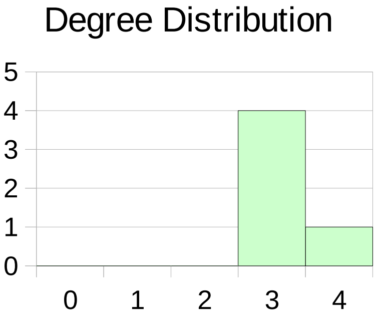
  - **Degree sequence**
    an ordered list of the (in,out) degree of each node
    - In-degree sequence: [4, 2, 1 , 1, 0]
    - Out-degree sequence: [3, 2, 2, 1, 0]
    - Degree sequence: [4, 3, 3, 3, 3]



  - **Degree Distribution**
    a frequency count of the occurrences of each degree
    [usually plotted as probability → normalization]







Sparsity of Networks

- **Real Networks are usually very Sparse!**

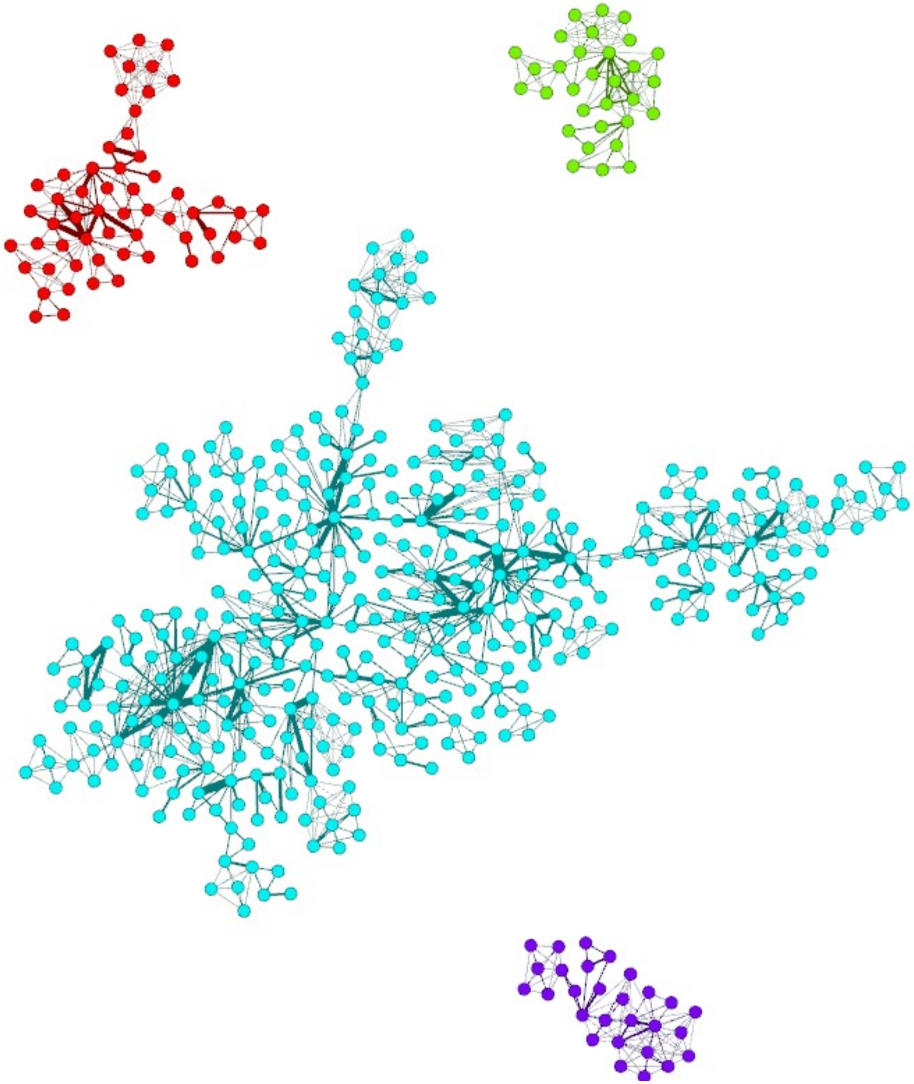
Network	Dir/Undir	Nodes	Edges	Avg. Degree
Internet	Undirected	192,244	609,066	**6.33**
WWW	Directed	325,729	1,479,134	**4.60**
Power Grid	Undirected	4,941	6,594	**2.67**
Mobile Phone Calls	Directed	36,595	91,826	**2.51**
Email	Directed	57,194	103,731	**1.81**
Science Collaboration	Undirected	23,133	93,439	**8.08**
Actor Network	Undirected	702,388	29,397,908	**83.71**
Citation Network	Directed	449,673	4,689,479	**10.43**
E. Coli Metabolism	Directed	1,039	5,082	**5.58**
Protein Interactions	Undirected	2,018	2,930	**2.90**

  – A graph where every pair of nodes is connected is called a **complete graph** (or a **clique**)

Table: Adapted from (Barabasi, 2015)

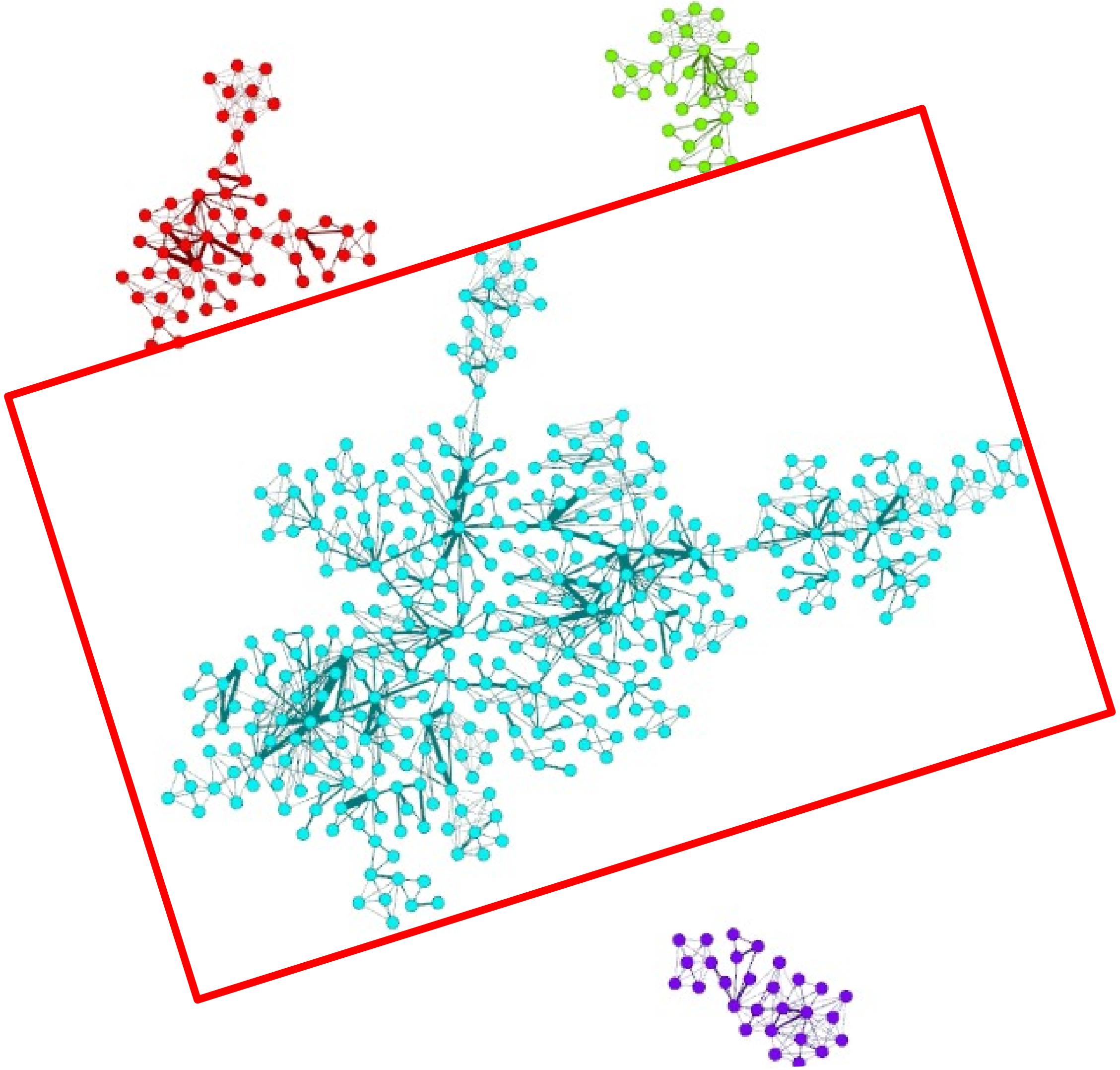
Connectivity

- **Not everything is connected**



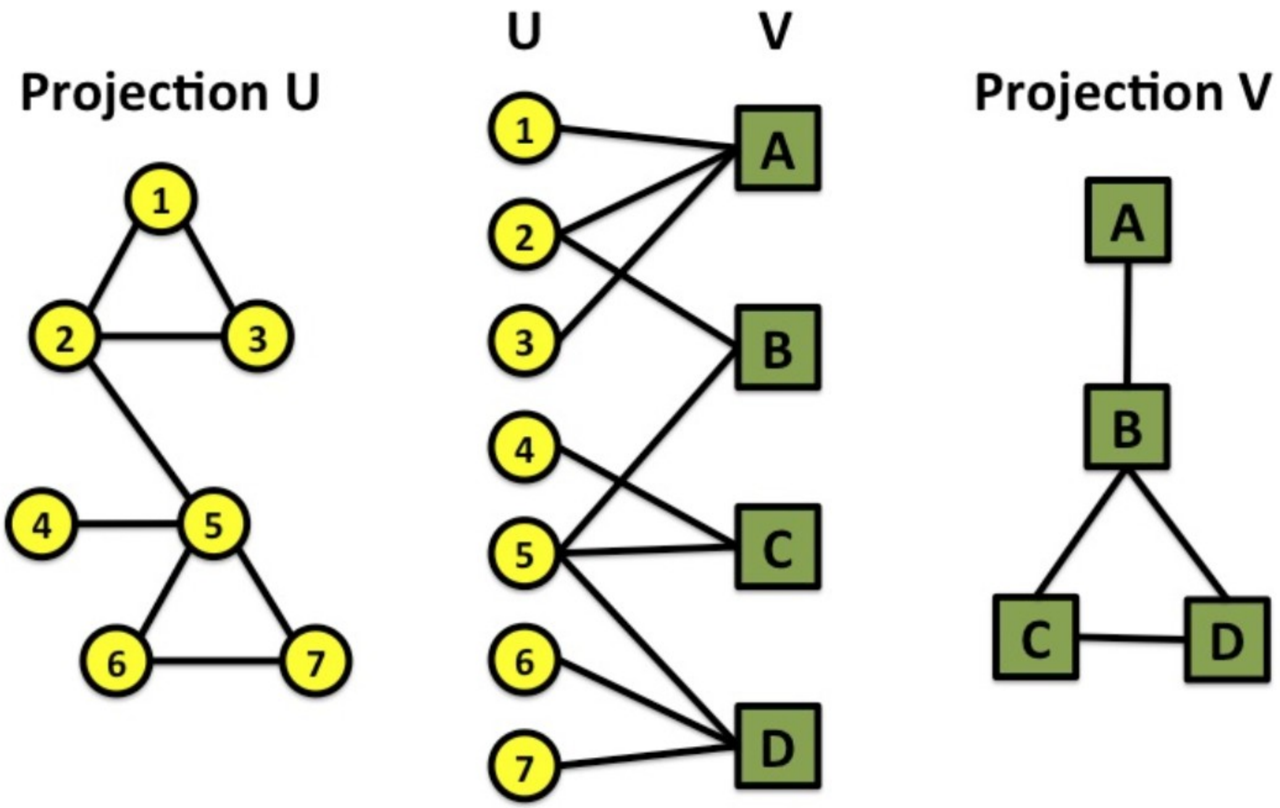
Connectivity

- **If the largest component has a large fraction of the nodes we call it the giant component**



Bipartite

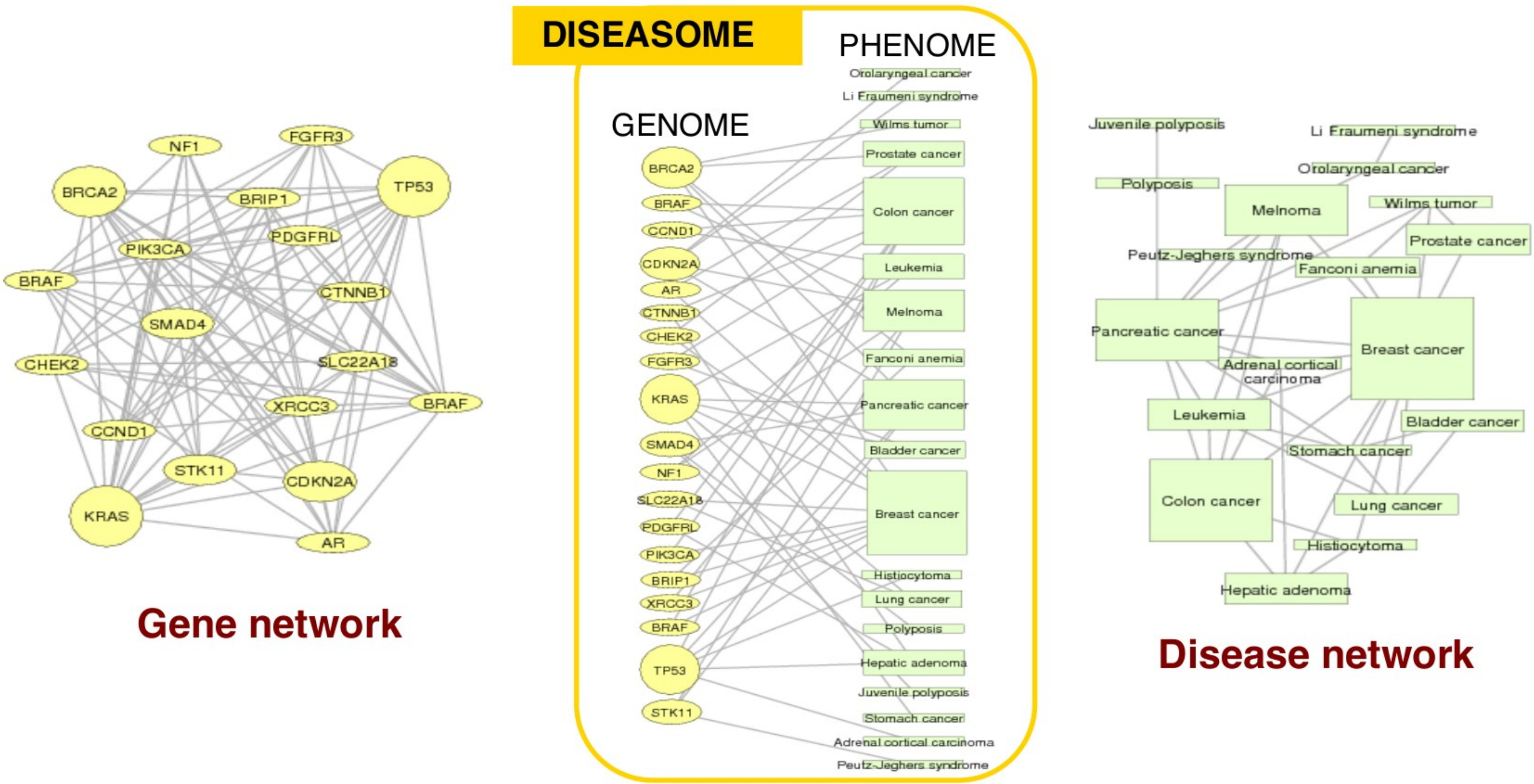
- **A bipartite graph is a graph whose nodes can be divided into two disjoint sets U and V such that every edge connects a node in U to one in V.**



Example:
- Actor Network
  . U = Actor
  . V = Movies

Image: Adapted from Leskovec, 2015

Bipartite Network Projections



Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

Bipartite – Human Disease Network