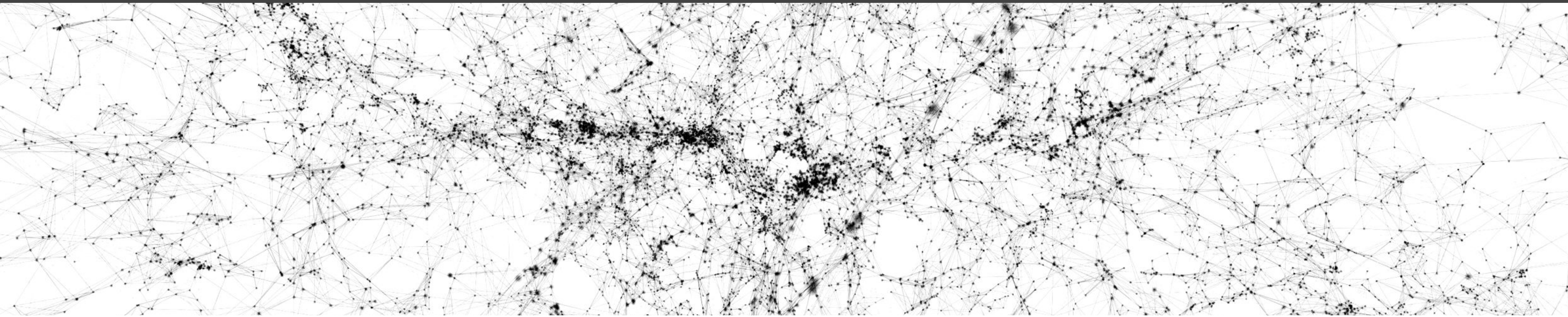
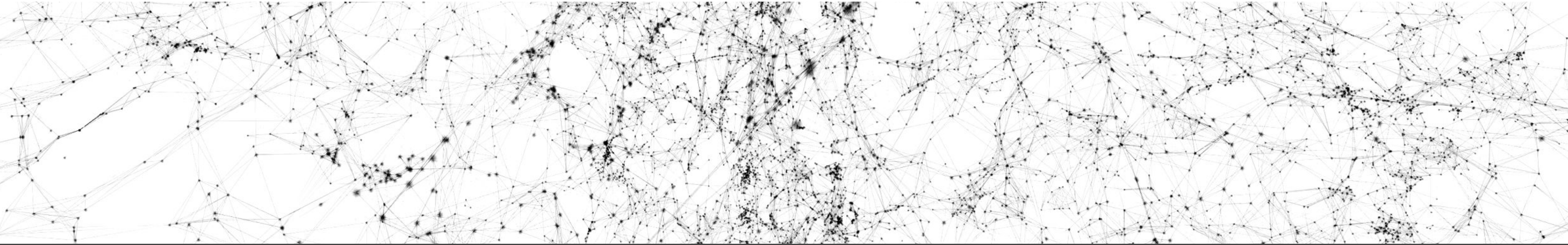


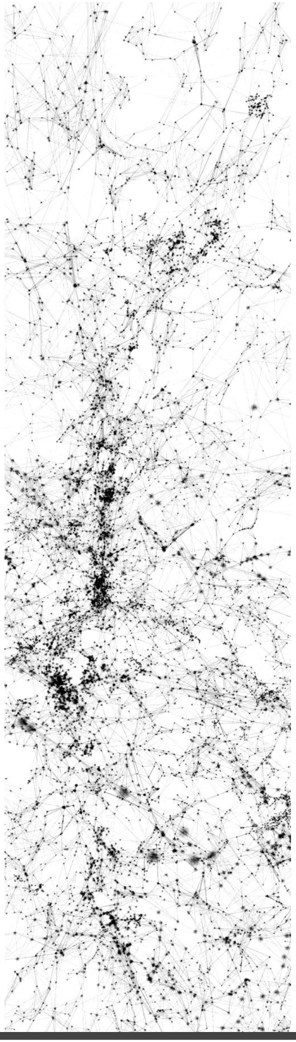
Network Science

Analysis of complex interconnected data



Outline

- Introduction to the course
  - Complex systems is Physics
  - Societies as complex systems
  - Complex data everywhere and at every scale
  - Main tasks in complex data analysis
- Logistics of the course
  - General info
  - Who is in the class
  - What we will learn
  - Grading etc.



Why network science?

The world around us is interconnected, and complex systems arise in in different fields.

Connections, interactions, relations are often present in real world data, and in many cases are key to understand the data.

"Learn how to see. Realize that everything connects to everything else."

— Leonardo da Vinci

Research disciplines

Analysis of complex interconnected data is multidisciplinary with researchers from:

- Physics (complex systems)
- Sociology (social networks)
- Mathematics (graph theory)
- Data Mining (graph mining)
- Machine Learning (relational learning, graph neural networks)

And sometimes is considered as its own discipline coined as

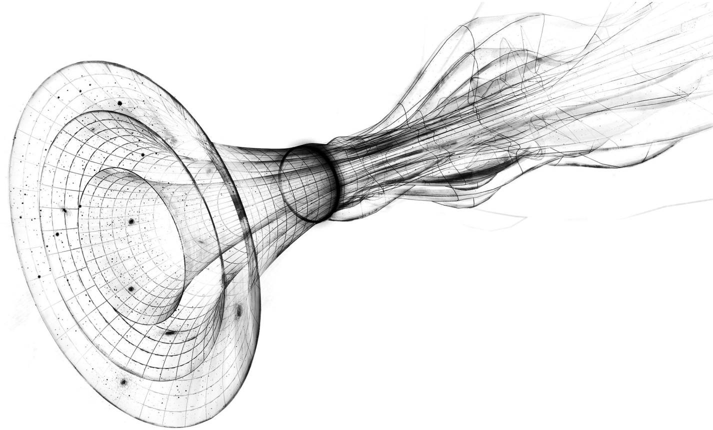
Network Science or Science of Networks

Physics

Examples: deterministic chaos, quantum entanglement, spin glasses

Study of complex systems has a long history in Physics, dating back to Aristotle’s time, and more relevant than ever in this century

“I think the next [21st] century will be the century of complexity”
— Stephen Hawking

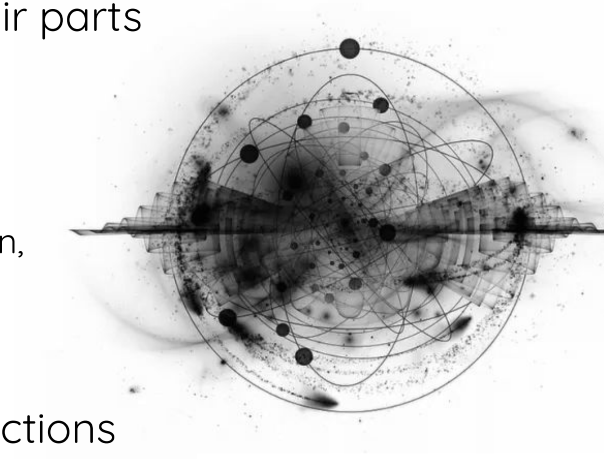


More is different,
— P. Anderson, Science (1972)

philosophy of science and emergent phenomena; limitations of reductionism and the existence of hierarchical levels of science

Complex systems

- consists of many interconnected parts
- characterized by time-dependent interactions among their parts
- are not an aggregation of their separate parts
- when looked at as a whole gives non trivial insights
    - *Emergence*: a property not any of components have on their own, arising during self-organization process
- often interactions change states of parts,
  and the states of the parts change the networks of interactions



com·plex

adjective
/ˌkämˈpleksˌkəmˈpleksˌˈkämˌpleks/

1. consisting of many different and connected parts.
   "a complex network of water channels"
   synonyms: compound, composite, compounded, multiplex
   "a complex structure"

Society as a complex system

Sociology studies the structure of social life, viewing the society as a complex system composed of individuals, whose parts work together through relations, associations, and other forms of connections, and the evolution and dynamics within them affects our life.

Sociology

From early on when the field was being defined as an academic discipline, sociologist emphasized that social science should look at the society as a whole, rather than being limited to the specific actions of individuals.

Social science should be holistic.

— Émile Durkheim (1895)
the principal architect of social science

French sociologist, formally established the academic discipline of sociology, insisted that society was more than the sum of its parts

What is society?

— Georg Simmel (1911)
forerunner of Structural functionalism

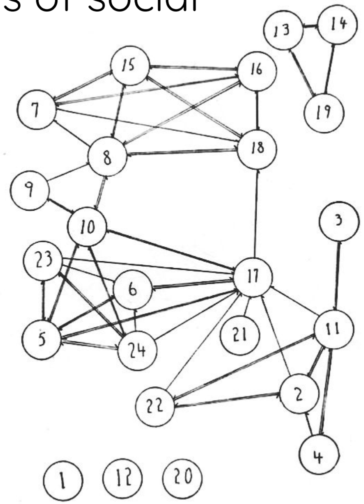
first generation of German sociologists
Sociology is the study of social interaction at the individual and small group level (dyad, triad, ...)

Social networks

a social structure of a set of social actors (e.g. individuals or organizations), and social interactions between the actors (e.g. sets of dyadic ties).

earliest graphical depictions of social networks (sociograms)

Jacob L. Moreno,
Who Shall Survive? (1934)



New York Training School for Girls
(a reformatory school for teenage girls),
within two weeks 14 girls ran away (30x more than the average). Moreno examined 500 girls and their feelings towards each other. He visualized these connections in several sociograms to model channels for the flow of social influence and ideas, and concluded that they behaved based on how they are positioned in their social network.

Read more here

Complex Data

Represents interconnections between the datapoints, and is different from the data representation which considers data as a set of feature vectors (often iid) each a D-dimensional representation for a datapoint

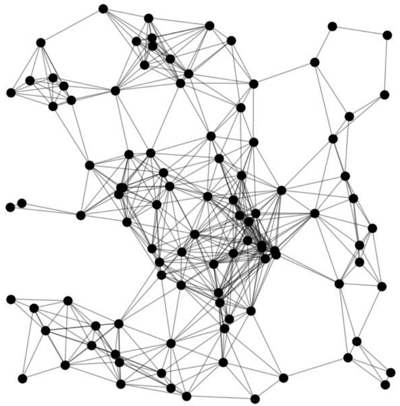
$$\mathcal{D} = \{x^{(n)}\}_{n=1}^{N}$$

feature

instance

$$X = \begin{bmatrix} {x^{(1)}}^T \\ {x^{(2)}}^T \\ \vdots \\ {x^{(N)}}^T \end{bmatrix} = \begin{bmatrix} x_1^{(1)}, & x_2^{(1)}, & \cdots, & x_D^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)}, & x_2^{(N)}, & \cdots, & x_D^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times D}$$

We are used to this



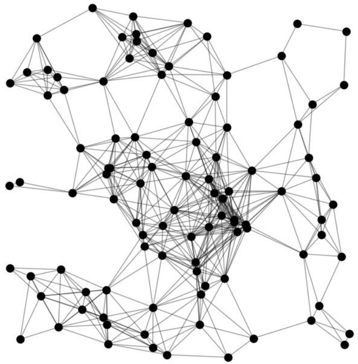
But we are given this

Graph Mining in CS

Analyzing, modelling complex data (not iid, structured)

Comes as flavours of (statistical) relational learning, learning in structured settings, graph convolutional nets, graph representation learning, etc.

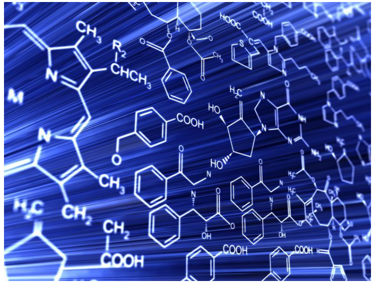
Let's start by seeing where we encounter this type of data and what we can do with this data



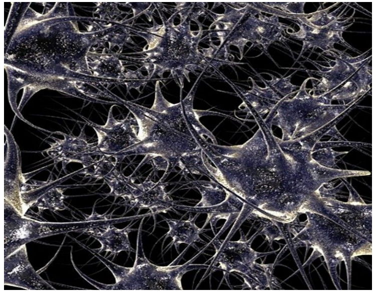
Natural sciences

In natural sciences, we see connections between atoms, molecules, cells, organisms and even we have cosmic web.

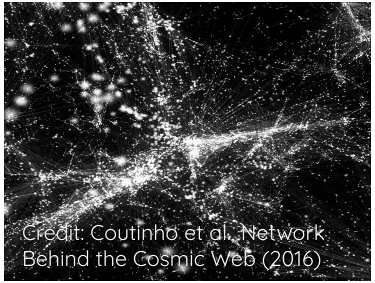
Chemistry



Biology



Physics



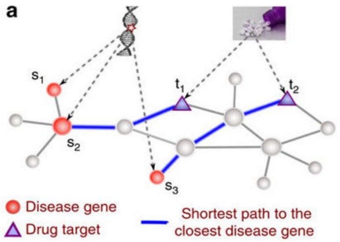
Credit: Coutinho et al. Network Behind the Cosmic Web (2016)

Check the interactive demo of galaxy networks here: https://cosmicweb.kimalbrecht.com/

Applied sciences

Interconnected systems exist in many applied sciences and other fields. There are numerous studies which show looking at these compex system, as a whole, gives us non trivial insights and is necessary to understand these systems.

Medicine

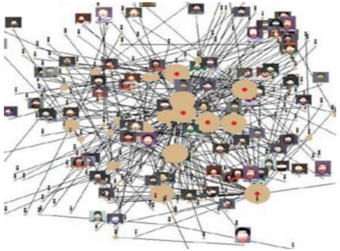


Disease Gene Network

Credit: Guney et al. (2016)

“the emergence of most diseases cannot be explained by single-gene defects, but involve the breakdown of the coordinated function of distinct gene groups”

Law



Criminal Network

Credit: Xu et al. (2005)

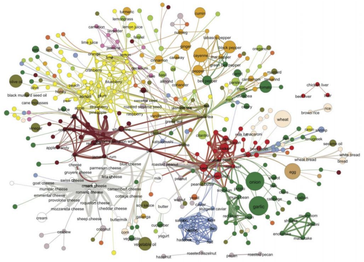
Economics

Trading Network

Credit: Adamic et al. (2017)

“strong feedback between the trading behaviour in buyers and sellers networks and the market conditions”

Culinary



Flavor Network

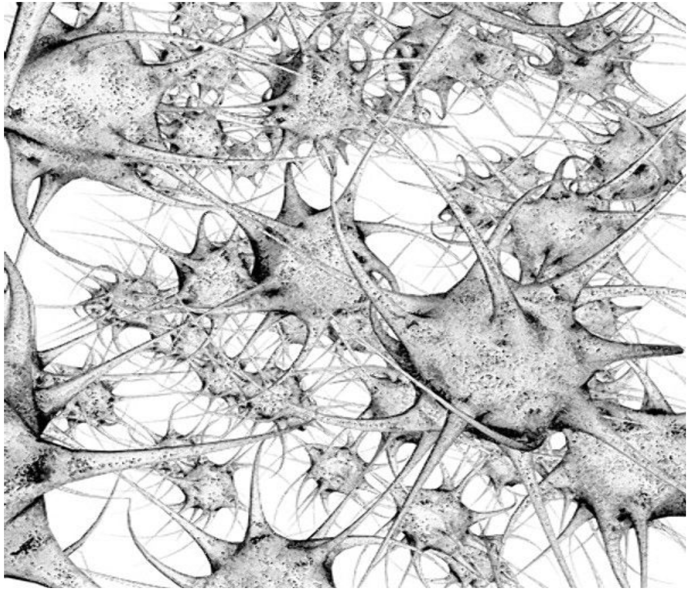
Credit: Ahn et al. (2011)

Read on food pairing theories and check out the interactive demo: https://foodgalaxy.jp/

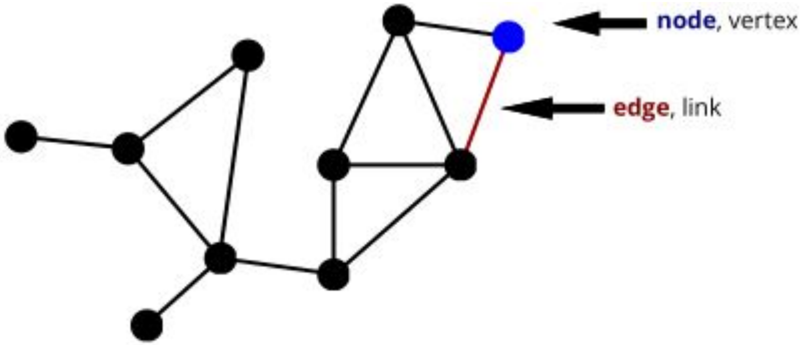
Different scales

Interconnected systems exist at different scales, for instance in biology we have networks

- Within Cells
  - Protein-Protein Interaction Networks
  - Gene Interaction Networks
  - Metabolic Networks
- Between Cells
  - Cell Signaling Networks
  - Neural Networks
- Between Organisms
  - Food Webs
- Between Species
  - Species Interaction Networks



Graphs: the default data representation



Extension: weighted, directed, signed, multi-edges and multi-type nodes (heterogenous), attributed (nodes and or edges have feature vectors), dynamic (sequence of graphs), multilayer networks (multi-view), hypergraphs (beyond pairwise relations), etc.

Adjacency: the default data structure

Adjacency Matrix

	0	1	2	3	4	5	6	7	8	9	10	11
0	0	1	1	0	0	0	0	0	0	0	0	1
1	1	0	1	1	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	1	1	0	0	0	0	0	0
4	0	0	0	1	0	1	1	0	0	0	0	0
5	0	0	0	1	1	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	1	1	0	0	0
7	0	0	0	0	0	0	1	0	1	0	0	0
8	0	0	0	0	0	0	1	1	0	0	1	0
9	0	0	0	0	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	0	1	1	0	1
11	1	0	0	0	0	0	0	0	0	1	1	0

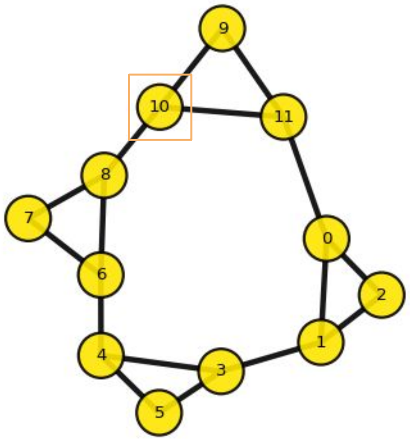
Adjacency List

0 : { 1 , 2 , 11 }
1 : { 0 , 2 , 3 }
2 : { 0 , 1 }
3 : { 1 , 4 , 5 }
4 : { 3 , 5 , 6 }
5 : { 3 , 4 }
6 : { 4 , 7 , 8 }
7 : { 6 , 8 }
8 : { 6 , 7, 10 }
9 : { 10 , 11 }
10 : { 8 , 9 , 11 }
11 : { 0 , 9 , 10 }

Edge List

{ (0, 1), (0, 2), (0, 11),
(1, 0), (1, 2), (1, 3),
(2, 0), (2, 1),
(3, 1), (3, 4), (3, 5),
(4, 3), (4, 5), (4, 6),
(5, 3), (5, 4),
(6, 4), (6, 7), (6, 8),
(7, 8), (7, 6),
(8, 6), (8, 7), (8, 10)
(9, 10), (9,11),
(10, 8), (10, 9), (10, 11),
(11, 0), (11, 9), (11, 10) }

Simple Graph

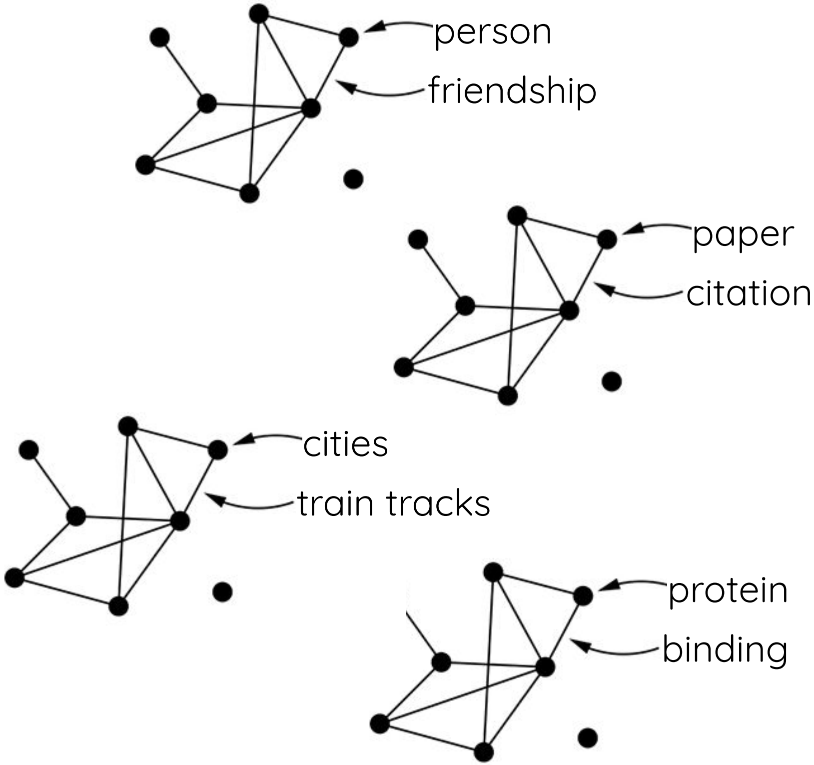


$$G(V,E), V = \{1\dots n\}, E = \{(i,j) | i,j \in [1\dots n]) \wedge A_{ij} = 1\}$$

Real world graphs are sparse (lots of zeros) and we use sparse matrix representations which in practice are similar to adjacency (LIL format)/edge list (COO format) and only store non-zero values.

Historic graph datasets

	Network	Type	n	m
Social	film actors	undirected	449 913	25 516 482
	company directors	undirected	7 673	55 392
	math coauthorship	undirected	253 339	496 489
	physics coauthorship	undirected	52 909	245 300
	biology coauthorship	undirected	1 520 251	11 803 064
	telephone call graph	undirected	47 000 000	80 000 000
	email messages	directed	59 912	86 300
	email address books	directed	16 881	57 029
	student relationships	undirected	573	477
	sexual contacts	undirected	2 810	
Information	WWW `nd.edu`	directed	269 504	1 497 135
	WWW Altavista	directed	203 549 046	2 130 000 000
	citation network	directed	783 339	6 716 198
	Roget's Thesaurus	directed	1 022	5 103
	word co-occurrence	undirected	460 902	17 000 000
Technological	Internet	undirected	10 697	31 992
	power grid	undirected	4 941	6 594
	train routes	undirected	587	19 603
	software packages	directed	1 439	1 723
	software classes	directed	1 377	2 213
	electronic circuits	undirected	24 097	53 248
	peer-to-peer network	undirected	880	1 296
Biological	metabolic network	undirected	765	3 686
	protein interactions	undirected	2 115	2 240
	marine food web	directed	135	598
	freshwater food web	directed	92	997
	neural network	directed	307	2 359

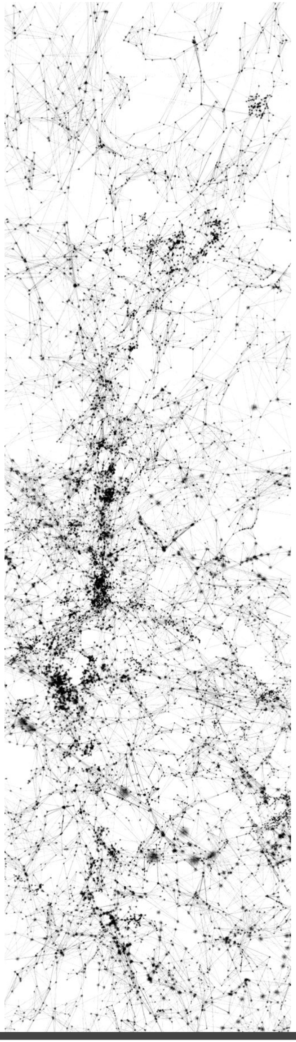


From: Newman ME. The structure and function of complex networks. SIAM review. 2003;45(2):167-256.

If interested, read part one of Newman's book on different types of network

Common tasks in network science

- Pattern & Anomaly Detection
- Modelling of Structure, Evolution, & Dynamics
- Measurements of Ranking & Similarity
- Clustering & Community Detection
- Prediction of Missing Link & Attributes
- Summarization, Visualization, & Layouts
- Temporal analysis of Evolution & Diffusion



Measurements of ranking & similarity

- **Ranking**: who is more important, or influential?
  - Degree Centrality, Betweenness Centrality, PageRank

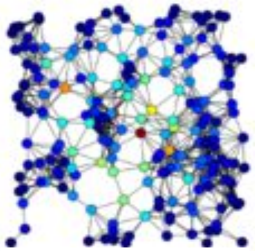
$$R(i) \text{ for } i \text{ in } [1..n]$$

- **Similarity**: how close are two nodes?
  - Shortest Path, Information Flow, common neighbors

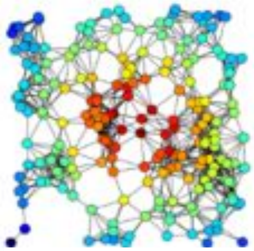
$$S(i,j) \text{ for } i,j \text{ in } [1..n]$$

Ranking nodes

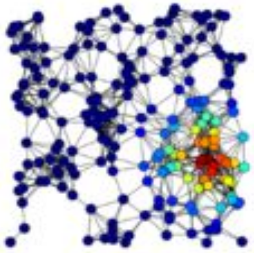
- Degree Centrality
- Closeness Centrality
  - average length of the shortest paths
- Betweenness Centrality
  - number of shortest paths
- Eigenvector Centrality
  - connections to high-scoring nodes contribute more
  - e.g. Katz & PageRank



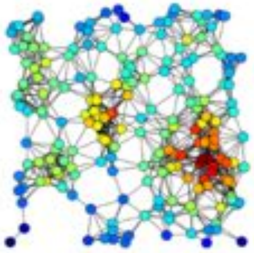
Betweenness



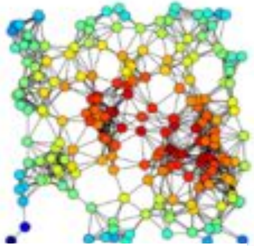
Closeness



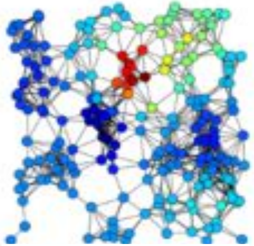
Eigenvector



Degree

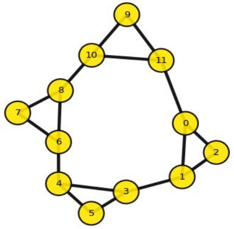


Harmonic

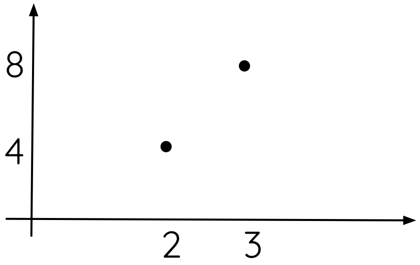


Katz

Degree, the basic measurement

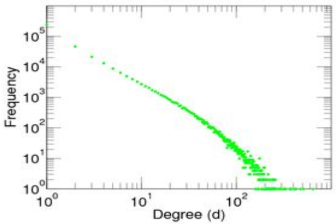


- Marginals of $\mathbf{A}$
  - $\mathbf{d_i = \Sigma_j A_{ij}}$
- Degree Distribution



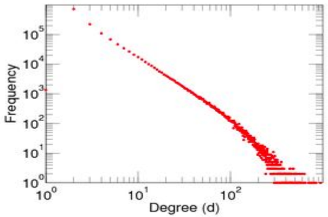
	0	1	2	3	4	5	6	7	8	9	10	11	
0	0	1	1	0	0	0	0	0	0	0	0	1	3
1	1	0	1	1	0	0	0	0	0	0	0	0	3
2	1	1	0	0	0	0	0	0	0	0	0	0	2
3	0	1	0	0	1	1	0	0	0	0	0	0	3
4	0	0	0	1	0	1	1	0	0	0	0	0	3
5	0	0	0	1	1	0	0	0	0	0	0	0	2
6	0	0	0	0	1	0	0	1	1	0	0	0	3
7	0	0	0	0	0	0	1	0	1	0	0	0	2
8	0	0	0	0	0	0	1	1	0	0	1	0	3
9	0	0	0	0	0	0	0	0	0	0	1	1	2
10	0	0	0	0	0	0	0	0	1	1	0	1	3
11	1	0	0	0	0	0	0	0	0	1	1	0	3
	3	3	2	3	3	2	3	2	3	2	3	3	**32**

Degree distribution



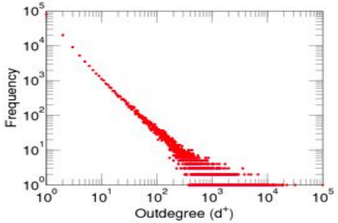
Actor degree distribution

[Actor-Movies](#)



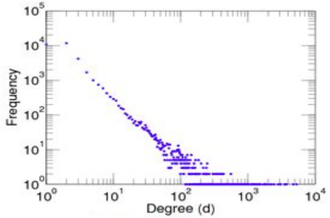
Author degree distribution

[Researcher-Publications](#)



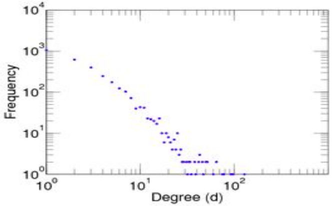
Outdegree distribution

[Wiki communications](#)



Degree distribution

[Internet Topology](#)

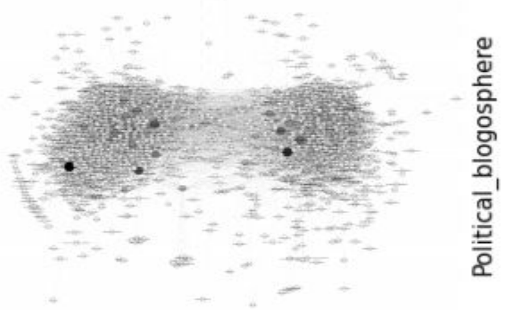


Degree distribution

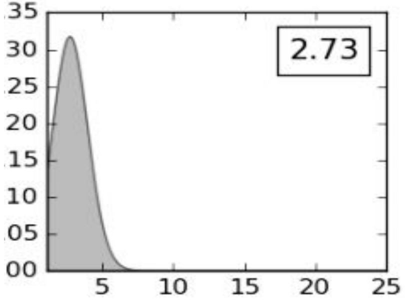
[Protein Interactions](#)

Explore different datasets with precomputed statistics here: http://konect.cc/

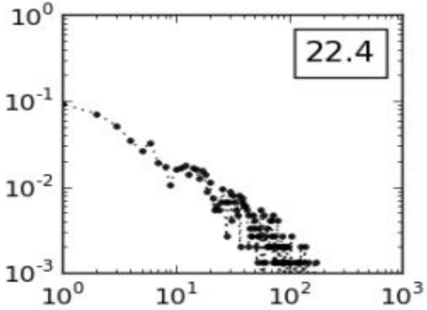
Patterns in networks



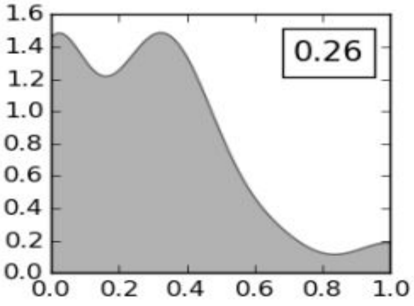
small-world



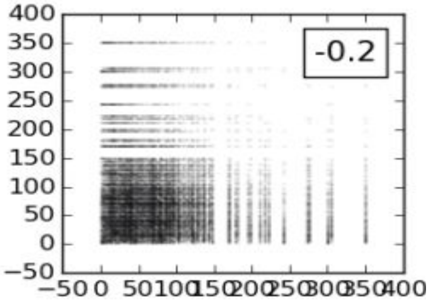
scale-free



transitive

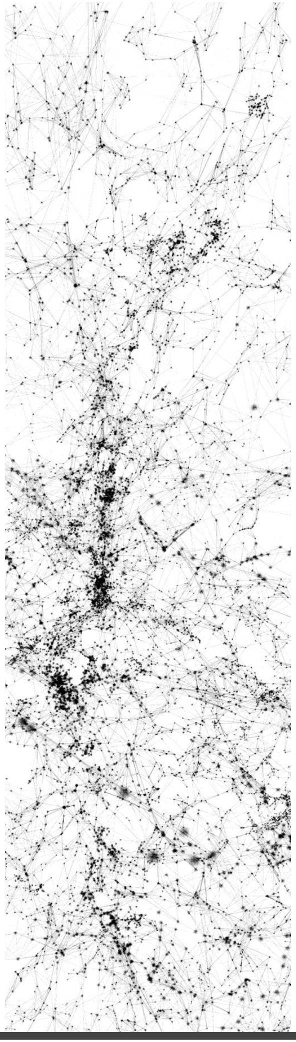


assortative mixing



Outline

- Introduction to the course
  - Complex systems is Physics
  - Societies as complex systems
  - Complex data everywhere and at every scale
  - Main tasks in complex data analysis
- Logistics of the course
  - General info
  - Who is in the class
  - What we will learn
  - Grading etc.



Logistics

Instructor: Reihaneh Rabbany

Teaching Assistant: Andy Huang

Class times: Tuesday & Thursday, 10:05-11:25

Class location: Macdonald Engi. Building 276 & Online [Zoom in Mycourses]

☛Please note that this is a seminar course and participation at the class time is required

Contact: **comp551mcgill@gmail.com**

Office hours: Thursday 12:00-1:00pm

Office: Online [Zoom in Mycourses]

Course Website: www.reirab.com/comp599.html [has all the information needed, links and access restricted items are through Mycourses]

Class admin

Instructor: Reihaneh Rabbany

Canada CIFAR AI Chair and core member at Mila

Assistant Professor in the School of Computer Science

I research on network science, data mining and machine learning, with a focus on analyzing real-world interconnected data, and social good applications

http://www.reirab.com/

Assistant: Andy Huang

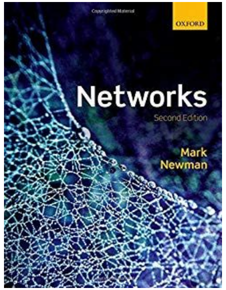
CS PhD student and a student at Mila

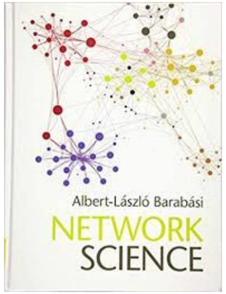
Research on network science, data mining and machine learning, with a focus on anomaly detection in temporal graph

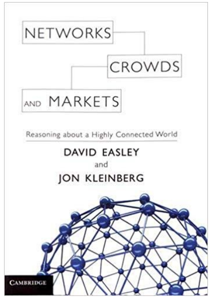
https://www.cs.mcgill.ca/~shuang43/

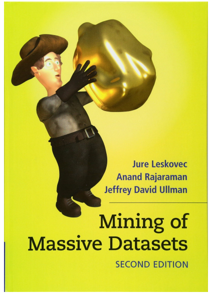
Reference Materials

- **Main textbooks**
  - **Networks: An Introduction** by M.E.J. Newman, ebook at library
  - **Network Science** by Albert-Barabasi, available online
- **Other textbooks**
  - **Networks, Crowds and Markets** by D. Easley and J. Kleinberg, available online
  - **Graph Representation Learning** by William L. Hamilton, available online
  - **Mining of Massive Datasets** by Jure Leskovec, Anand Rajaraman, Jeff Ullman, available online
- **Surveys and conference papers**
  - Web (WebConference, WSDM, ICWSM), Data (KDD, ICDM, SDM, ECML/PKDD, PAKDD), Learning (ICML, NeurIPS), Networks (ASONAM, NetSci, Complex Networks), …









What ~~you~~ we will learn

- Fundamental methods in each topic
  - Highly cited papers and basic concepts
- State of the art papers in each topic
  - Seminars on recent papers
- How to work with networked data
  - Assignments
- How to (attempt to) advance this area
  - Project

Grading details

- 50% project (10% proposal, 15% progress report, 25% final report)
- 30% assignments (3x10%)
- 10% presentations of assigned papers
- 10% reviewing assignments
  note: most of the grading is by peer-assessment
- bonus points:
  - 5 points for the best class presentation
  - 5 points for the best project proposal
  - 5 points for the best reviewer
  - 10 points for the best project
  - 1 point for each interesting point you share at the end of a class from the readings (for the current or previous lectures) which was not covered in the class

Project

- 50% project [use the format linked in the website for writeups]
  - 10% proposal
    - Writeup: 2 pages, describing what and why [8pt]
    - Presentation: 2 mins (2-3 slides) [2pt]
    - You will pitch this and get feedback
  - 15% progress report
    - Writeup: 4-5 pages, describing how and some preliminary results [12pt]
    - Presentation: 3 mins (3-4 slides) [3pt]
    - You will submit this and get feedback
  - 25% final report
    - Writeup: 8 pages, full project report [20pt]
    - Presentation: 7 mins (7-10 slides) [5pt]
    - You will submit this and get feedback and time to improve/respond before final submission
- Peer Reviewing: provide feedbacks on projects from other groups on each round
  - Proposal [2pt], progress [3pt], final [5pt]

Grading & policies

- 30% assignments (3x10%)
  - Basic programing with networked data
    - Assignment one: patterns in real world networks [explore]
    - Assignment two: random network and community detection [unsupervised]
    - Assignment three: node and link prediction [supervised]

Grading & policies

- 10% presentations of assigned readings (one/two presentations)
    - You need to be able to answer questions and fully understand the paper, read the background papers necessary, if code, data is released, check those out, etc.
    - Each presentation is 20 minutes, you need to practice to be ontime. We will run this similar to conference presentations
    - Cover each with equal emphasis/time allocation: problem def, motivation, main intuition, methodology, experiment setup (data, tasks, evaluation), main finding, and results
        - Don’t spend all of the time taking about only one component, e.g. details of the method
- How you get marked?
    - Average score given by the listeners, peers and instructor

Collaboration

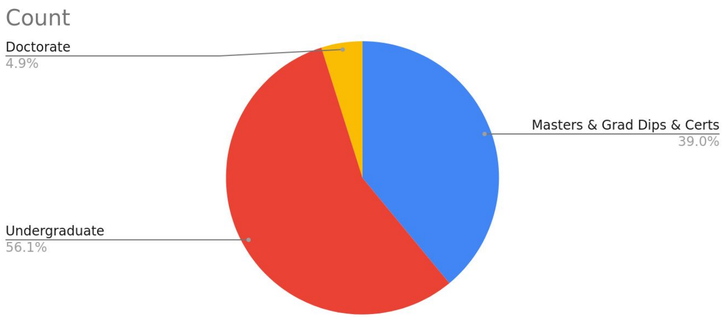
Welcome, but you need to acknowledge, cite any used resources

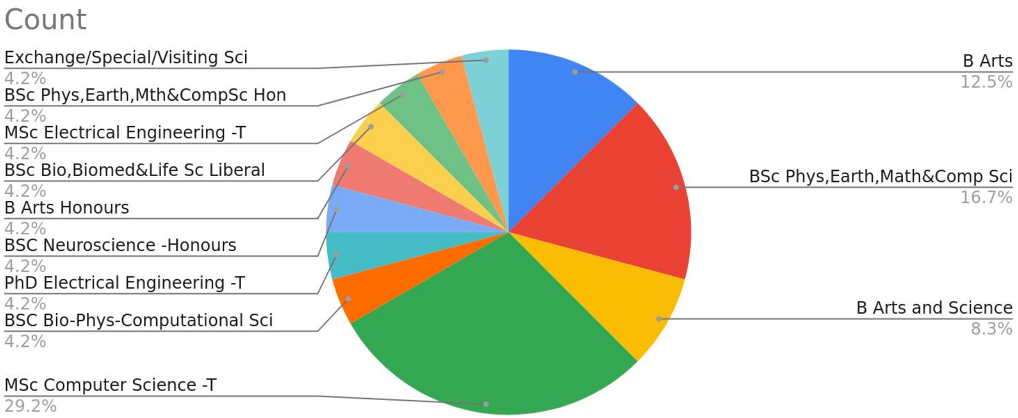
You should not copy and paste anything more than 3 consecutive words, in coding or write ups. This and other forms of plagiarism will be reported

Class composition

Quick round of introductions

- Name
- Your background
- Any particular reason for taking this class





Further optional readings

- The first ideas: Six degrees of separation & small world experiment
  - First mentioned in a novel in 1929, then validated in real world through experiments in 1967
- Funding papers:
  - Emergence of scaling in random networks, 1999
  - On power-law relationships of the Internet topology, 1999
- Interesting read: More is different (loosely relevant)
- Watch:
  - Connected Movie
  - Mark Newman 1 - The Connected World
  - Networks are everywhere with Albert-László Barabási
  - Mark Newman - The Physics of Complex Systems

Childhood's end by Arthur C. Clarke