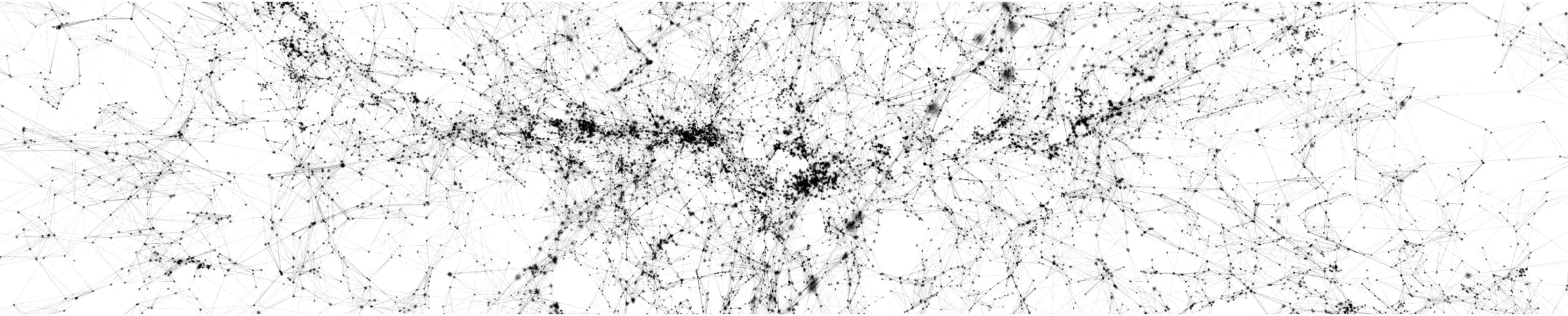
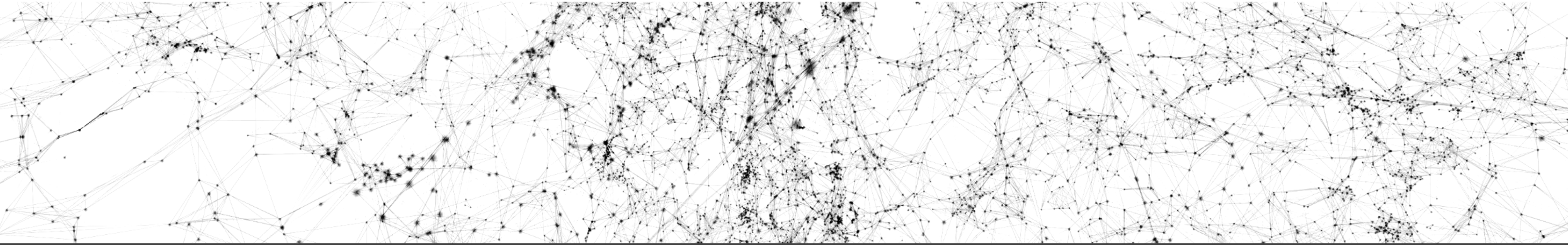


Network Science

Analysis of Complex Interconnected Data

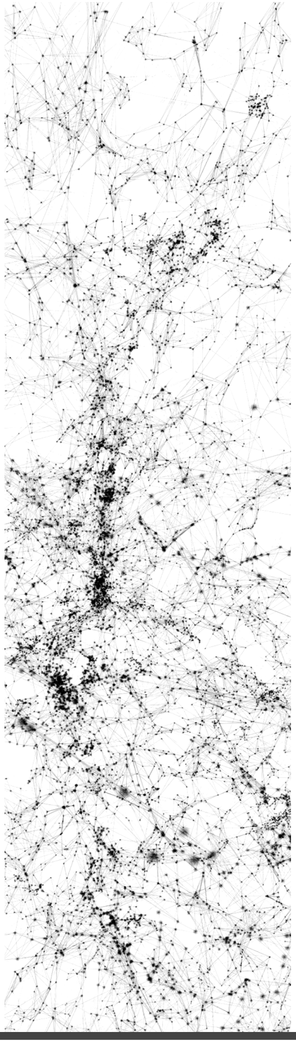
~~ Introduction ~~



McGill
School of Computer Science

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, deadlines, …



Why network science?

The world around us is interconnected, and complex systems arise 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*

Read more on wiki

Research disciplines

Analysis of complex interconnected data is multidisciplinary:

- 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, see here

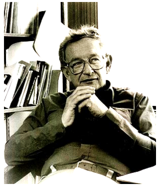
Complex Systems in Physics

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

examples: deterministic chaos, quantum entanglement, spin glasses

It is not limited to Physics phenomenons and even reaches the philosophy of science



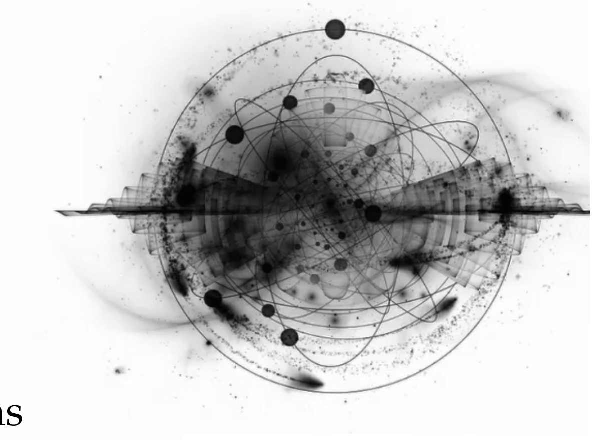
In 'More is different',
— P. Anderson, Science (1972)
Condensed matter physicist who discusses emergent phenomena; limitations of reductionism and the existence of hierarchical levels of science

X	Y
solid state or many-body physics	elementary particle physics
chemistry	many-body physics
molecular biology	chemistry
cell biology	molecular biology
.	.
.	.
.	.
psychology	physiology
social sciences	psychology

Read it [here]

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 a self-organization process
- often interactions change states of parts,
  and the states of the parts change the networks of interactions



com·plex

adjective
/ˌkämˈpleks,kəmˈpleks,ˈkämˌpleks/

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

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.

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

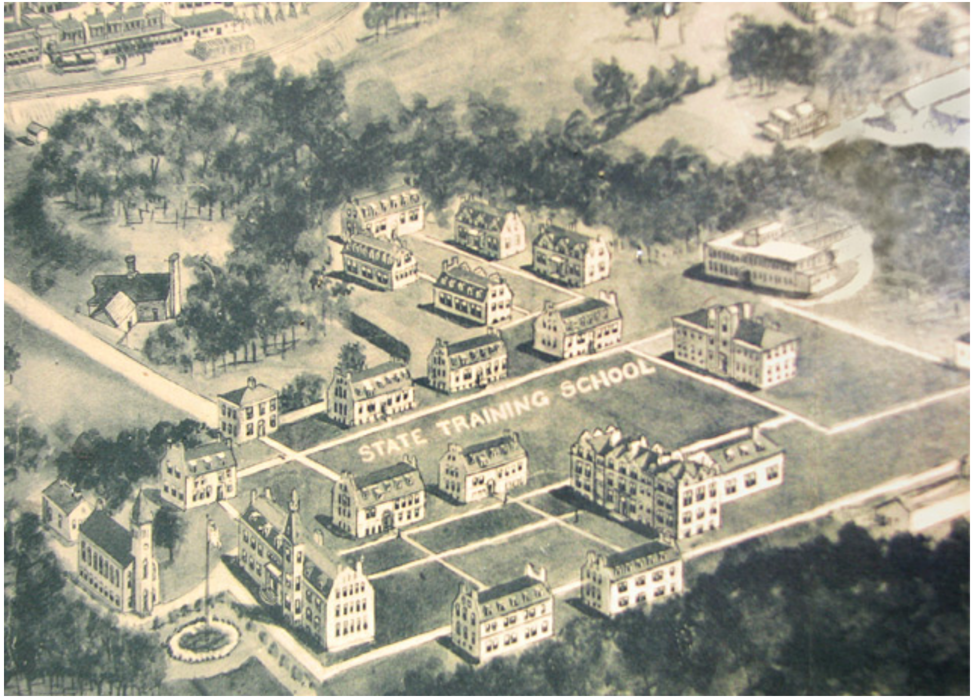
Social science should be holistic.
— Émile Durkheim (1895)
the principal architect of social science

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...)

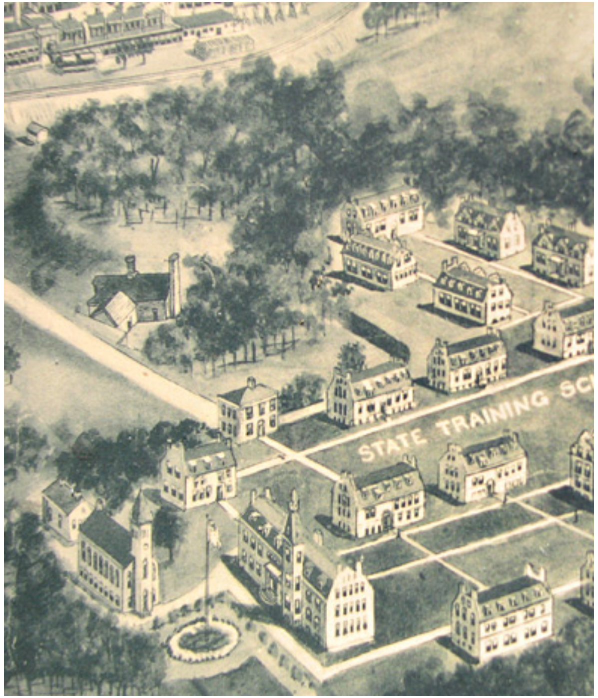
How to explain the pandemic of runaways?



New York Training School for Girls

In 1932, within two weeks
14 girls ran away
(30x more than the average)

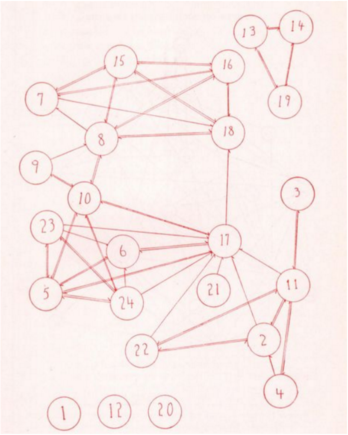
How to explain the pandemic of runaways?



Jacob L. Moreno,

Mapped out the **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



earliest graphical depictions of social networks (sociograms)

Who Shall Survive? (1934)

How to explain the pandemic of misinformation, fake news, conspiracy theories, populism, extremism, covid, …

Model the **channels for the flow of social influence and ideas**, and **infer** how individuals **behave based on how they are positioned in their social network**

W.H.O. Fights a Pandemic Besides Coronavirus: An 'Infodemic'

Working with the big tech companies, the U.N. health agency has made strides in combating rumors and falsehoods on the internet about the new infection.

Facebook, YouTube usage linked to belief in coronavirus conspiracy theories, study finds

PUBLISHED WED, JUN 17 2020·7:01 PM EDT | UPDATED THU, JUN 18 2020·1:09 AM EDT

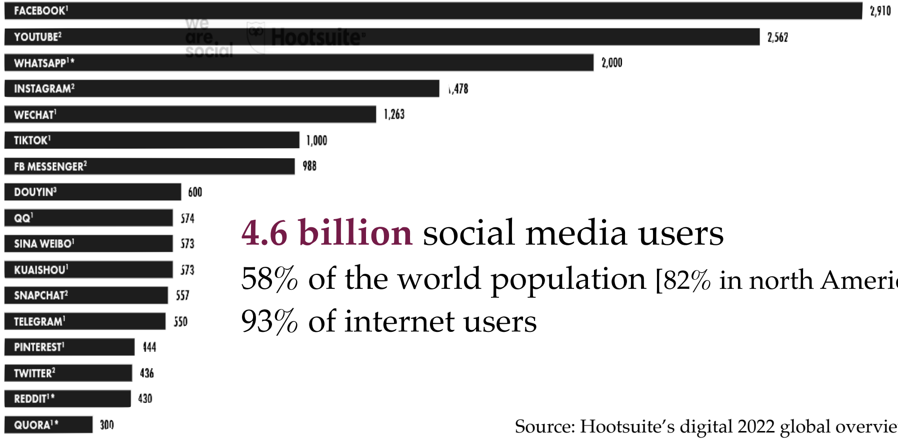

FACEBOOK UNDER FIRE

Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs

A trove of internal documents turned over to the SEC provides new details of the social media platform's role in fomenting the storming of the U.S. Capitol

By Craig Timberg, Elizabeth Dwoskin and Reed Albergotti

October 22, 2021 at 7:36 p.m. EDT



4.6 billion social media users

58% of the world population [82% in north America]

93% of internet users

Source: Hootsuite's digital 2022 global overview report

Model Complex Data as Graphs

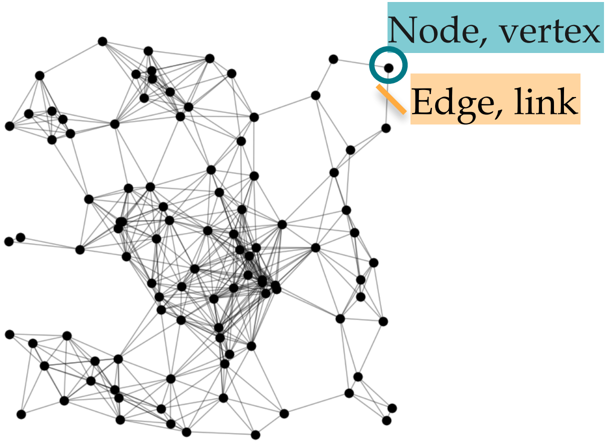
Represents interconnections between the datapoints as graphs or edge streams,

A (0,1) square matrix of size N (number of nodes)

$$A \in [0,1]^{N \times N}$$

$$A_{ij} = 1 \iff (i,j) \in E$$

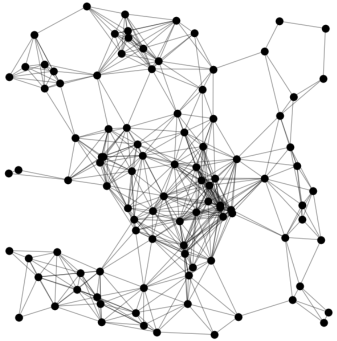
$$G(V,E), E \subseteq \{(i,j) \,|\, (i,j) \in V^2\}$$



Extension: weighted, directed, signed, bipartite, 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.

Model Complex Data as Graphs

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



connections & features of the instances are often **dynamic** and in interplay

similarity of individuals' characteristics motivates them to form relations (social selection) & characteristics of individuals is affected by the characteristics of their neighbours (social influence)

feature

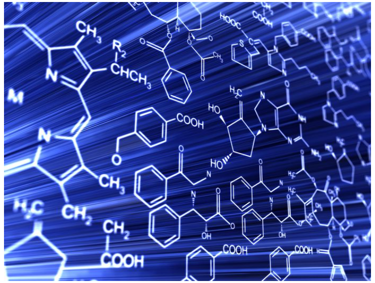
$$X = \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(N)T} \end{bmatrix} = \begin{bmatrix} x_1^{(1)}, & x_2^{(1)}, & \cdots, & x_D^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)}, & x_2^{(N)}, & \cdots, & x_D^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times D}$$

instance

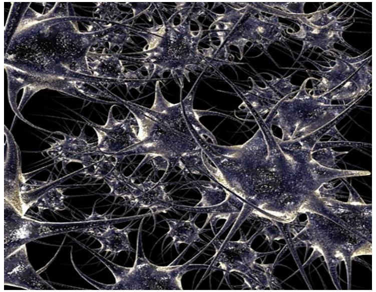
Natural sciences

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

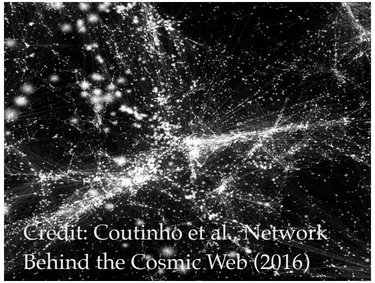
Chemistry



Biology



Physics



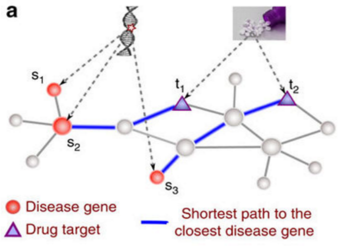
Credit: Coutinho et al., Network Behind the Cosmic Web (2016)

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 complex system, as a whole, gives us non trivial insights and is necessary to understand these systems.

Medicine

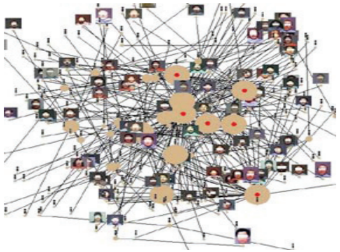


Disease Gene Network

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)

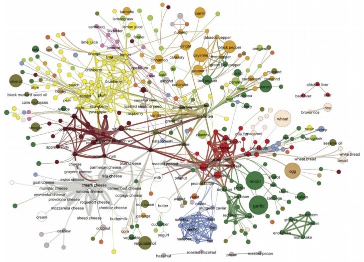
Economics

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

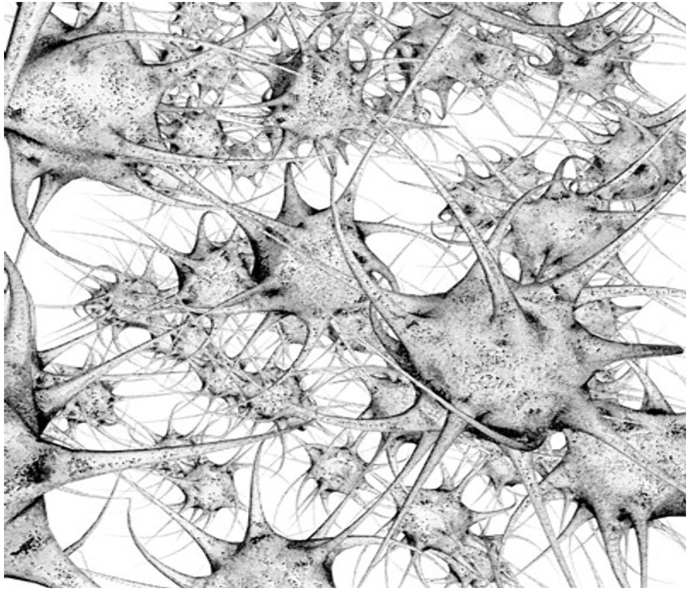
Credit: Ahn et al. (2011)

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

Different scales

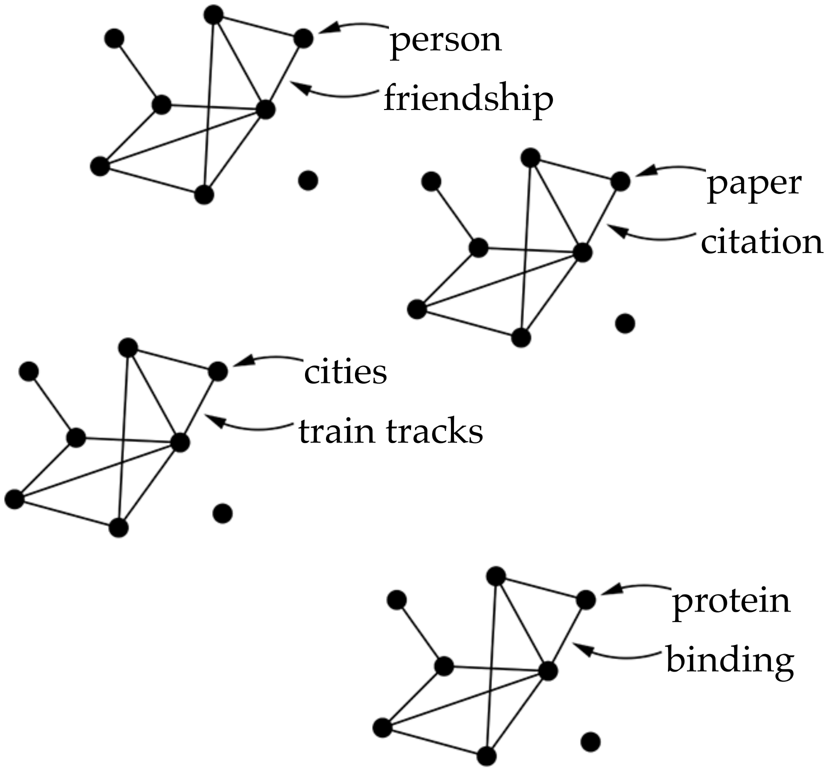
Interconnected systems exist at different scales, for instance in biology we have networks

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



Benchmark 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



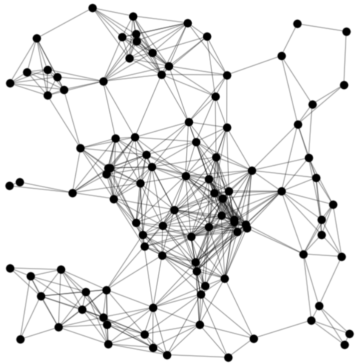
From: The structure and function of complex networks by Newman. SIAM review. 2003;45(2):167-256.

If interested, take a look at part one of Newman's book on different types of network: [Chapters 2-5 here]

Graph Mining in CS

Analyzing, modelling complex data (not iid, structured)

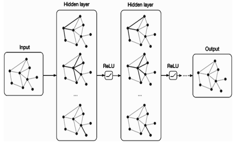
Comes as flavours of (statistical) relational learning, learning in structured settings, graph neural nets, graph representation learning, etc.



Recent Trend: Deep Learning for Graphs

2015~2025

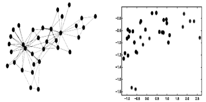
- o **Graph Neural Networks**
- o Deep Learning for Networks
- o High-Order Networks [Benson et al.]



21st Century: More CS

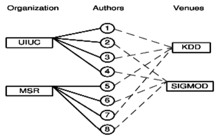
2010~2014

- o Info. vs. Social Networks (Twitter) [Kwak et al.]
- o **Signed** Networks [Leskovec et al.]
- o Semantic Social Networks [Tang et al.]
- o Four Deg. Of Separation [Backstrom et al.]
- o Structural Diversity [Ugander et al.]
- o Computational Social Science [Watts]
- o **Network Embedding** [Perozzi et al.]



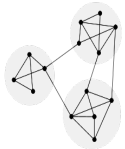
2005~2009

- o Graph Evolution [Leskovec et al.]
- o 3 Deg. Of Influence [Christakis & Fowler]
- o Social **Influence** Analysis [Tang et al.]
- o Six Deg. Of Separation [Leskovec & Horvitz]
- o Network **Heterogeneity** [Sun & Han]
- o Network **Embedding** [Tang & Liu]
- o Computer Social Science [Lazer et al.]



2000~2004

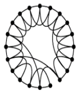
- o Influence Max'n [Domingos & Kempe et al.]
- o **Community Detection** [Girvan & Newman]
- o Network Motifs [Milo et al.]
- o Link Prediction [Liben-Nowell & Kleinberg]

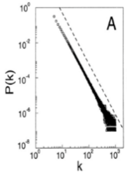


Late 20th Century: CS & Physics

1998~1999

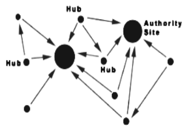
- o **Small Worlds** [Watts & Strogatz]
- o **Scale Free** [Barabasi & Albert]
- o **Power Law** [Faloutsos x3]





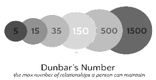
1997

- o **HITS** [Kleinberg]
- o **PageRank** [Page & Brin]
- o Hyperlink Vector Voting [Li]



1992

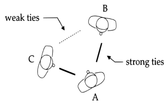
- o Structural Hole [Burt]
- o **Dunbar's Number** [Dunbar]



20th Century: Sociology

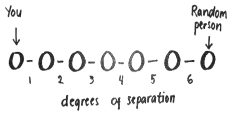
1970s

- o The Strength Of **Weak Tie** [Granovetter]



1960s

- o **Small Worlds** [Migram]

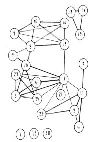


1950s

- o **Random Graph** [Erdos, Renyi, Gilbert]
- o Degree Sequence [Tuttle, Havel, Hakami]
- o **Homophily** [Lazarsfeld & Merton]
- o Balance Theory [Heider et al.]

1930s

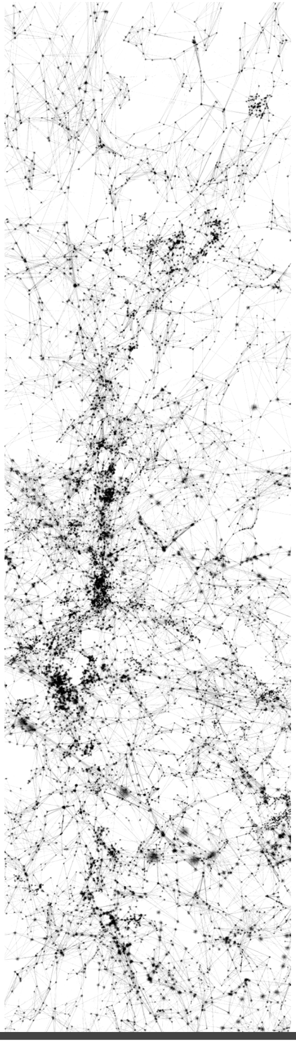
- o **Sociogram** [Moreno]



Based on Slides from Jie Tang

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

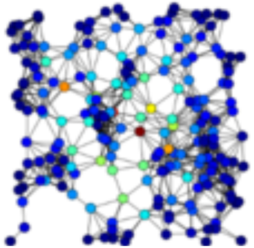
$$R : v \mapsto \mathbb{R}$$

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

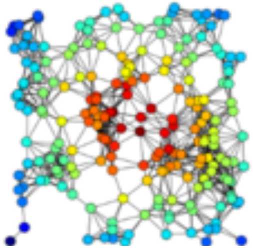
$$S : (u, v) \mapsto \mathbb{R}$$

Ranking nodes

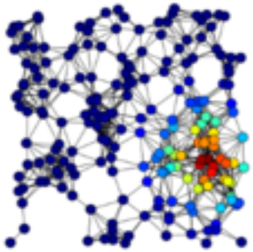
- Degree Centrality
  - marginals of the adjacency matrix
- Closeness Centrality
  - average length of the shortest paths
- Betweenness Centrality
  - number of shortest paths
- Eigenvector Centrality
  - connections to high-scoring nodes
  - e.g. Katz & PageRank



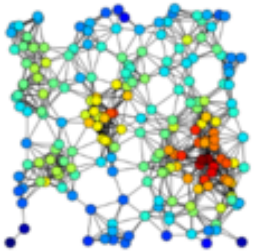
Betweenness



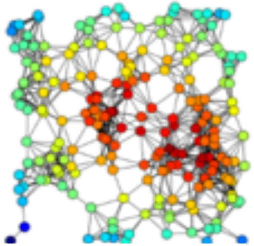
Closeness



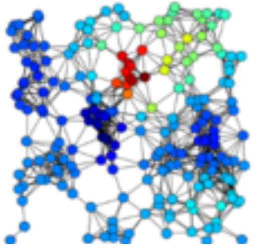
Eigenvector



Degree

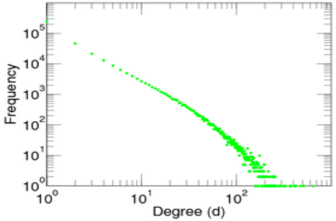


Harmonic



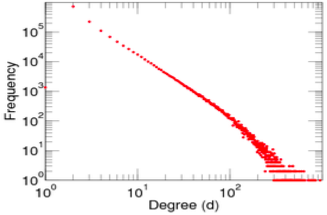
Katz

Degree distribution is heavy tailed [Example Pattern]



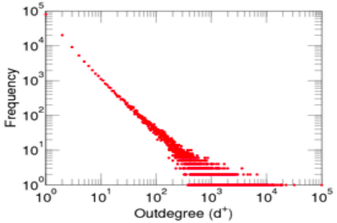
Actor degree distribution

[Actor-Movies](#)



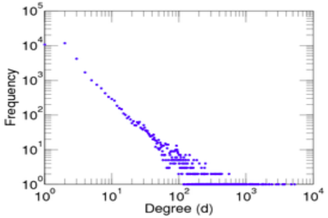
Author degree distribution

[Researcher-Publications](#)



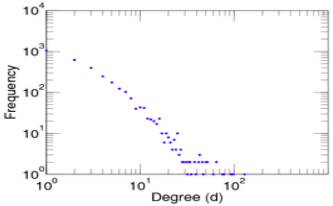
Outdegree distribution

[Wiki communications](#)



Degree distribution

[Internet Topology](#)



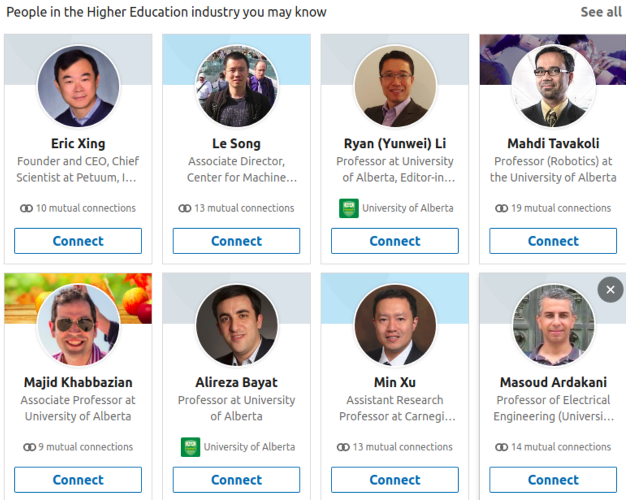
Degree distribution

[Protein Interactions](#)

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

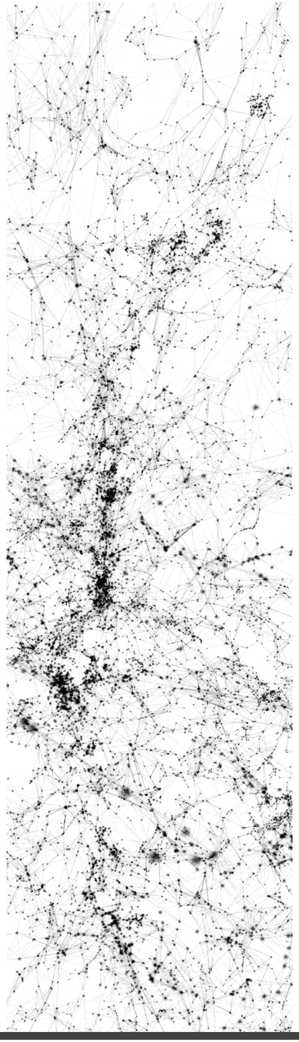
Link Prediction [Example Task]

- Modelling of the network evolution
- Predict likely interactions, not explicitly observed
- Link recommendation: “friend” suggestion in social networks



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, deadlines, …



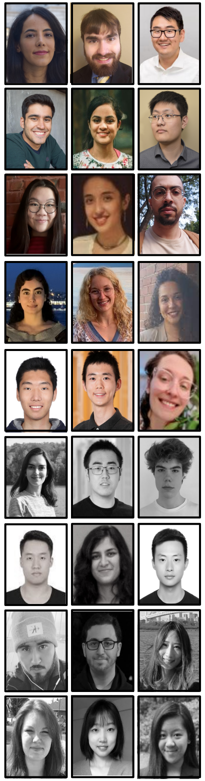
Logistics

Instructor: Reihaneh Rabbany [Office hours: Tuesdays 11:30-12:30pm, Zoom]

Teaching Assistants: TBD

Contact: **netscimcgill@gmail.com**

Course Website: www.reirab.com/comp511.html [has all the information needed, links and access restricted items are through Mycourses]



CDLab
McGill

Mila

McGill

Algorithms
analyzing
temporal graphs

- Change point & anomaly detection
- Diffusion on [dynamic] graphs
- Active learning in graphs
- Weak and noisy label / data
- Graph representation learning
- Node classification and link prediction

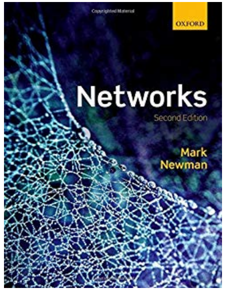
Applications
health & safety
of online
societies

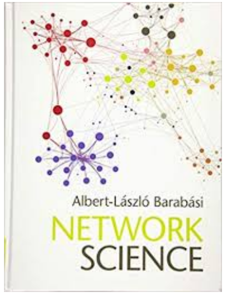
- Crime & online markets — Infrared (AI for Humanity, Mila)
- **Politics & online media** — CÉCD CSDC, Fodem (groupe de recherche sur le fonctionnement de la démocratie)
- Toxicity & online games — UBISOFT LA FORGE

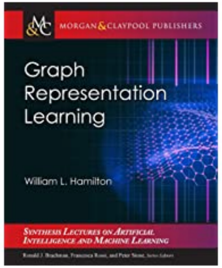
Complex Data
interconnected (graphs), multimodal (text, images),
noisy, adversarial, evolving, hard-to-label

Reference Materials

- **Main textbooks**
  - **Networks: An Introduction** by M.E.J. Newman, ebook at library
  - **Network Science** by Albert-Barabasi, available online
  - **Graph Representation Learning Book** by William L. Hamilton, 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 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 publications
- 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:
  - 2 points for the best class presentation
  - 2 points for the best project proposal
  - 2 points for the best reviewer
  - 5 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 [specific writing format linked in the website]
  - 10% proposal
    - Writeup: 2 pages, describing what and why [8pt]
    - Presentation: 2 mins (2-3 slides) [2pt] *
    - Pitch this and get feedback (review peer submissions)
  - 15% progress report
    - Writeup: 4-5 pages, describing how and some preliminary results [12pt]
    - Presentation: 3 mins (3-4 slides) [3pt] *
    - Submit this and get feedback (review peer submissions)
  - 25% final report
    - Writeup: 8 pages, full project report [20pt]
    - Presentation: 7 mins (7-10 slides) [5pt] *
    - Submit this and get feedback and time to improve/respond before final submission
- Peer Reviewing [10%]: provide feedbacks on projects from other groups on each round
  - Proposal [2pt], progress [3pt], final [5pt].

* tentative, depends on number of projects

Grading & policies

- 30% assignments (3x10%): basic programming 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 presentation)
  - showing full understanding of the paper and related background
  - being able to answer questions
  - proper timing: each presentation is 10 minutes
  - proper depth/breath: covering with equal emphasis/time allocation: problem def, motivation & intuition, methodology, experiment setup (data, tasks, evaluation), findings & results
    - e.g. don't get tangled in explaining the theory of the method, loosing the big picture
- How you get marked?
  - Average score given by the listeners, peers and instructor

Collaboration

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

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

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

Class composition

A Quick round of Discussions
if we have time also intro

- Name
- Your background & interests
- Any particular reason for taking this class
- Python, linear algebra & ML background?