ID2211
Data Mining, Basic Course

Sarunas Girdzijauskas, KTH

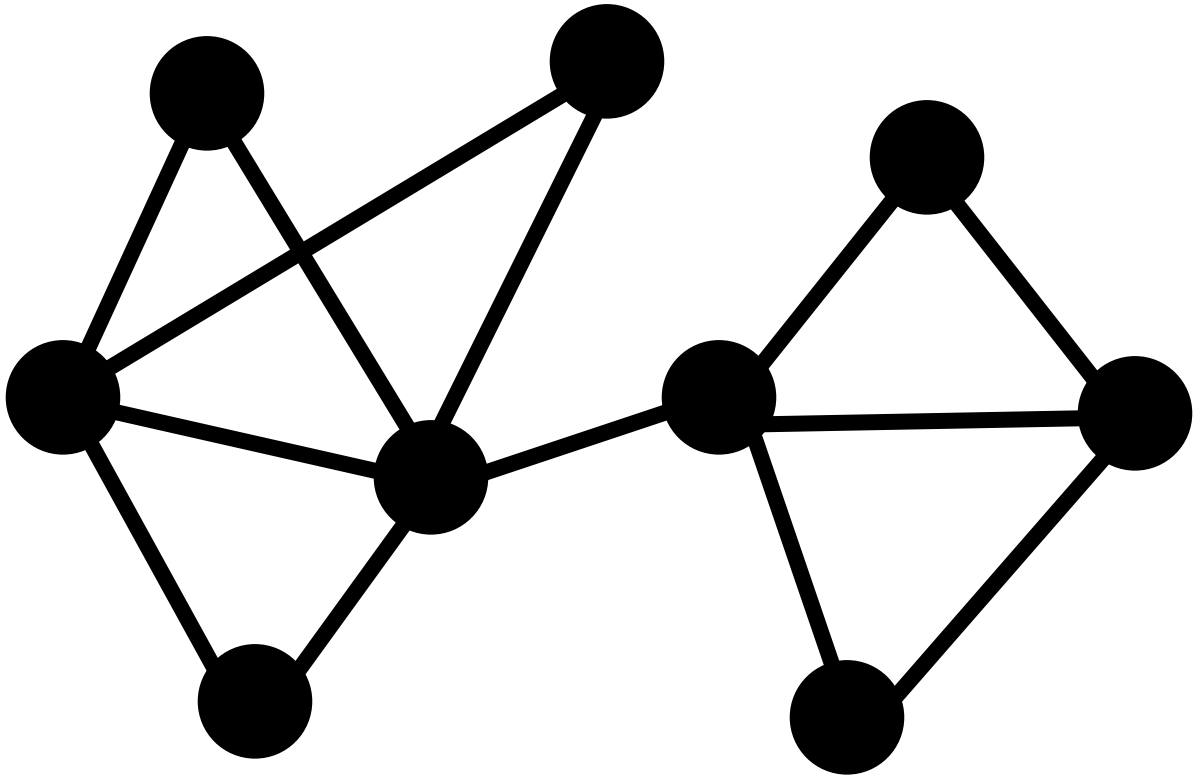
March 2022

Hybrid Lecture

What is Data?

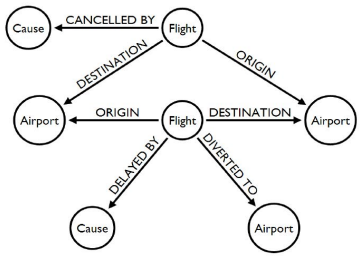
- Are these data points/objects independent?
  - Relations, Interactions, Dependencies!
- A (complex) network appears!
- Focus of this course:

Network Data Mining

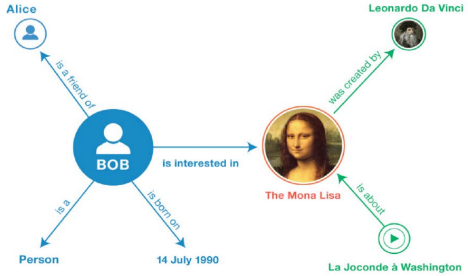


Networks are Everywhere!

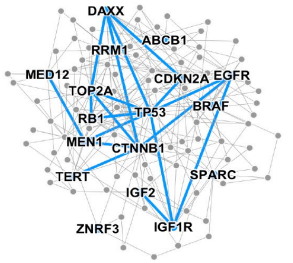
- **Examples?**
  - The Internet
  - Social networks
  - Financial Systems
  - Brain
  - Cell
  - Roads
  - Power Grids
  - Information Dissemination
  - Disease spreading
  - Etc.



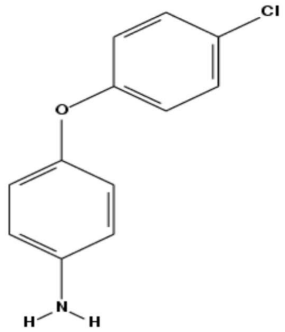
Event Graphs



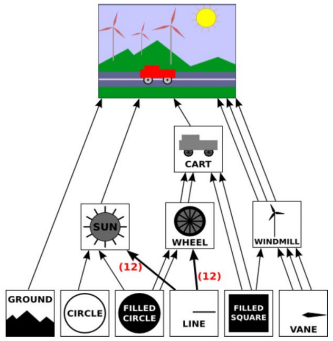
Knowledge Graphs



Disease pathways



Molecules



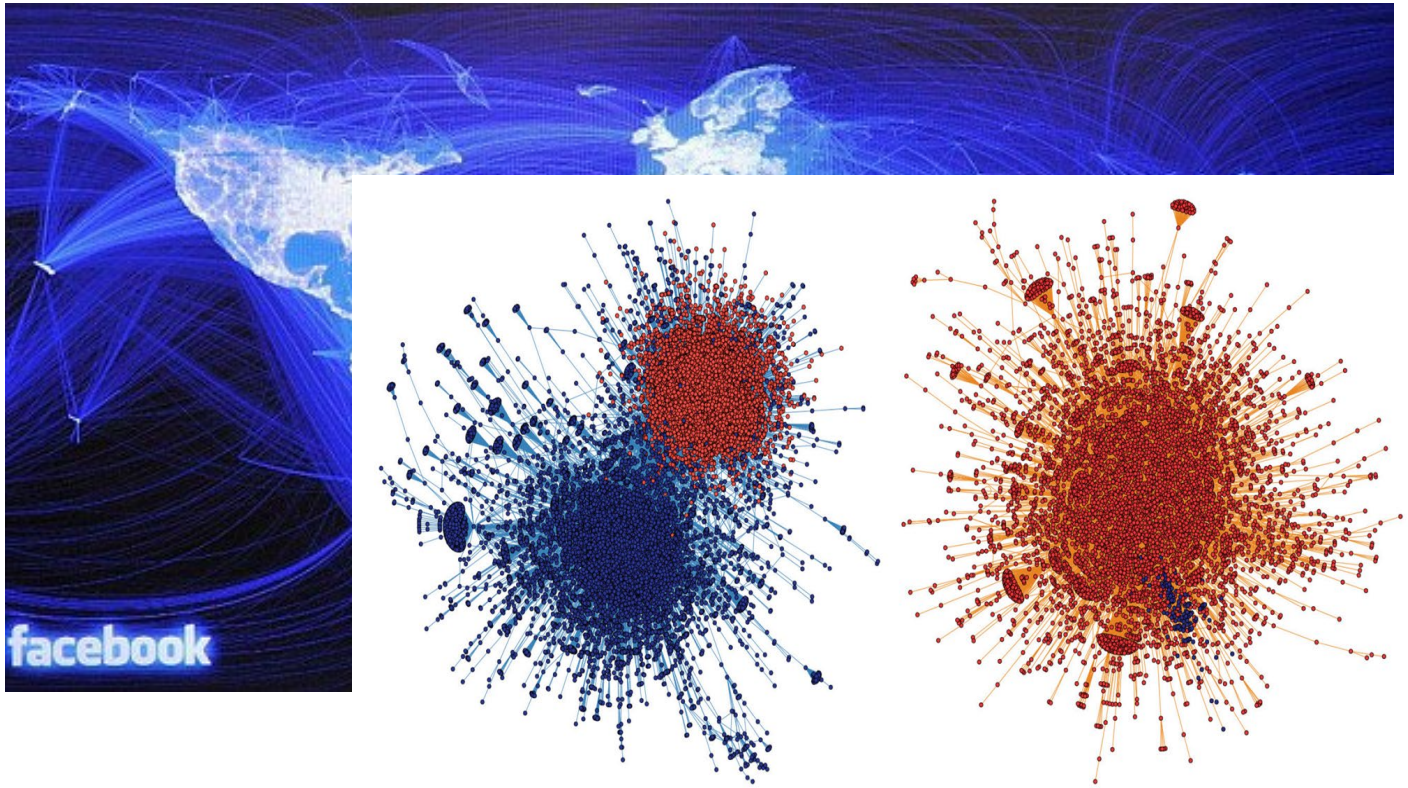
Scene Graphs

Cell-cell similarity networks

- **Networks are a general language for describing Complex Systems**
  - **In order to understand these systems and extract actionable knowledge we need to understand the networks behind them!**

Examples: Social Networks

- **Questions to ask?**
  - Friend suggestion (link prediction)
  - Recommendation systems (News Feed, Ads)
  - Fake News/misinformation spread, Modelling Epidemics, Polarisation in opinions etc

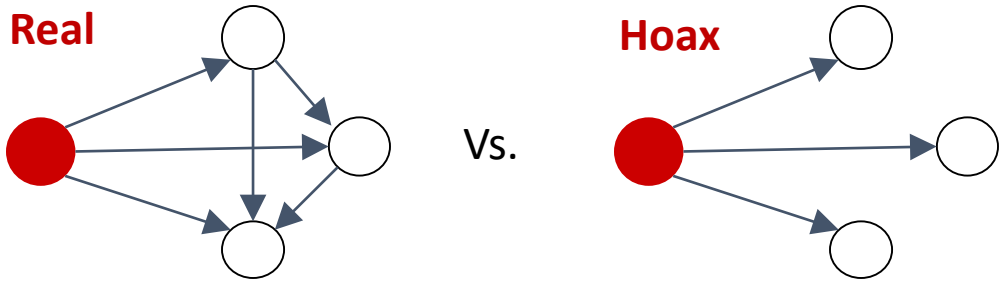


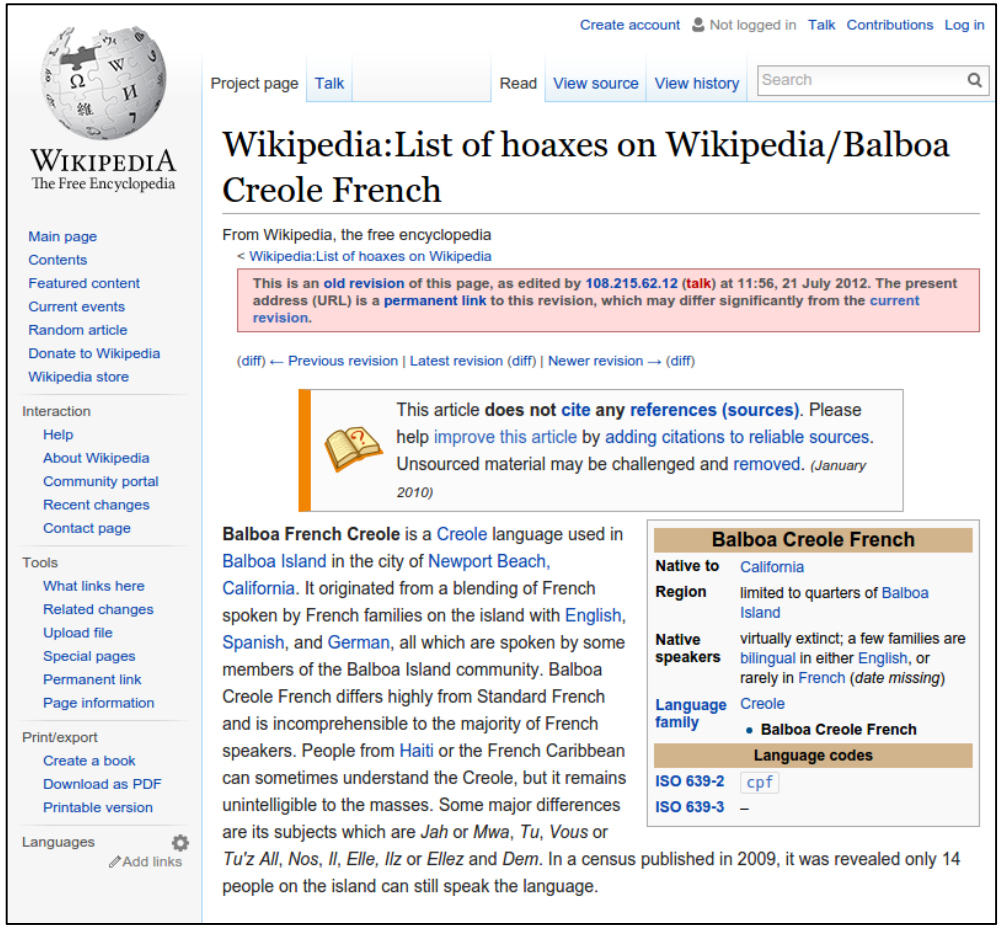
Twitter Retweet networks:
Polarized (left), Unpolarized (rigth)

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. "Political Polarization on Twitter." (2011)

Application: Misinformation

- **Q: Is a given Wikipedia article a hoax?**
  - Real articles link more coherently:





Create account Not logged in Talk Contributions Log in

Project page | Talk | Read | View source | View history | Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information

Print/export
Create a book
Download as PDF
Printable version

Languages
Add links

Wikipedia:List of hoaxes on Wikipedia/Balboa Creole French

From Wikipedia, the free encyclopedia
< Wikipedia:List of hoaxes on Wikipedia

This is an old revision of this page, as edited by 108.215.62.12 (talk) at 11:56, 21 July 2012. The present address (URL) is a permanent link to this revision, which may differ significantly from the current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

This article **does not cite any references (sources)**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(January 2010)*

Balboa French Creole is a Creole language used in Balboa Island in the city of Newport Beach, California. It originated from a blending of French spoken by French families on the island with English, Spanish, and German, all which are spoken by some members of the Balboa Island community. Balboa Creole French differs highly from Standard French and is incomprehensible to the majority of French speakers. People from Haiti or the French Caribbean can sometimes understand the Creole, but it remains unintelligible to the masses. Some major differences are its subjects which are *Jah* or *Mwa*, *Tu*, *Vous* or *Tu'z All*, *Nos*, *Il*, *Elle*, *Ilz* or *Ellez* and *Dem*. In a census published in 2009, it was revealed only 14 people on the island can still speak the language.

Balboa Creole French	
Native to	California
Region	limited to quarters of Balboa Island
Native speakers	virtually extinct; a few families are bilingual in either English, or rarely in French (*date missing*)
Language family	Creole
	• Balboa Creole French
Language codes	
ISO 639-2	cpf
ISO 639-3	–

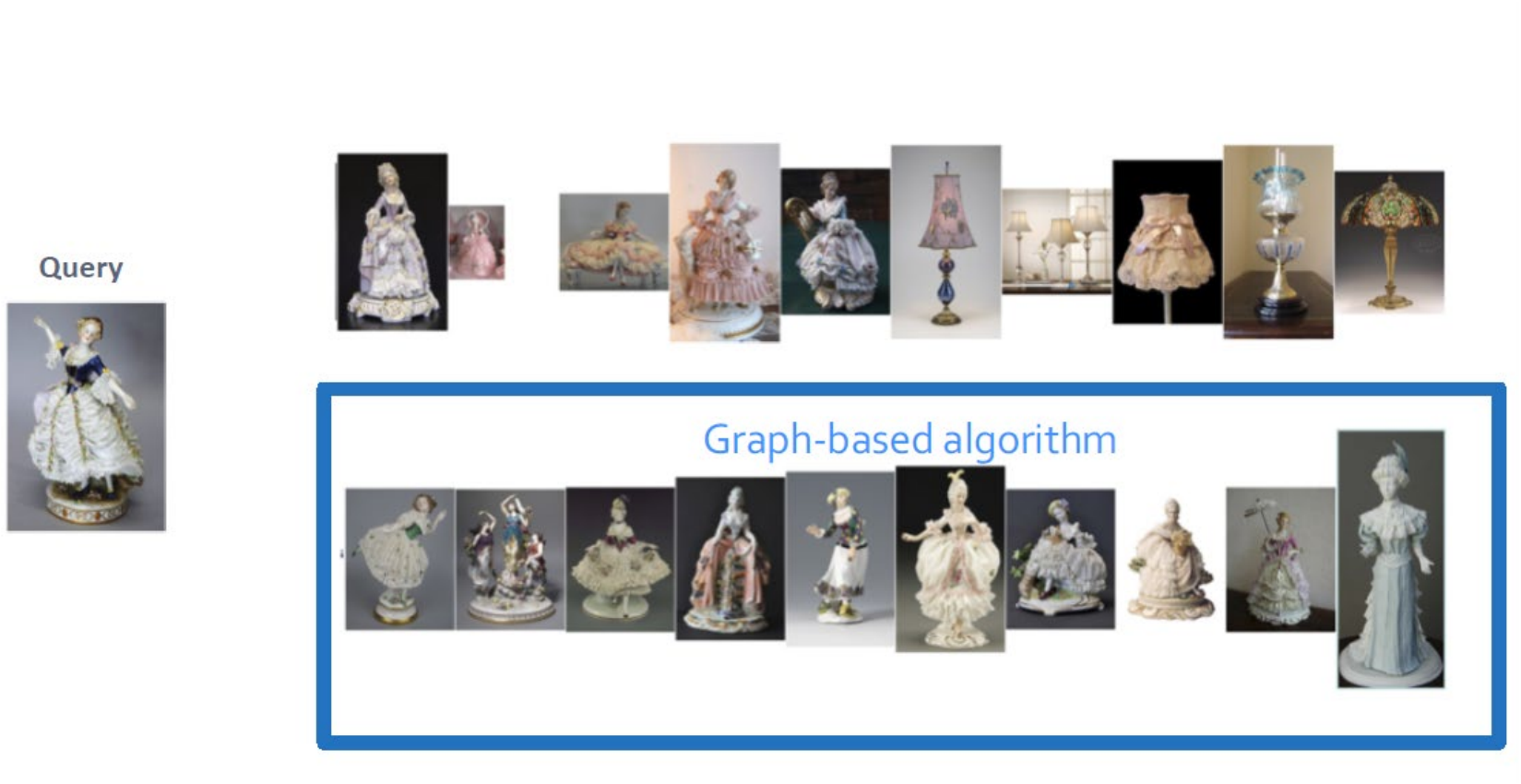
Hoax article detection performance:

50%	66%	86%
Random	Human	WWW ‘16

Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. Kumar et al. WWW ‘16.

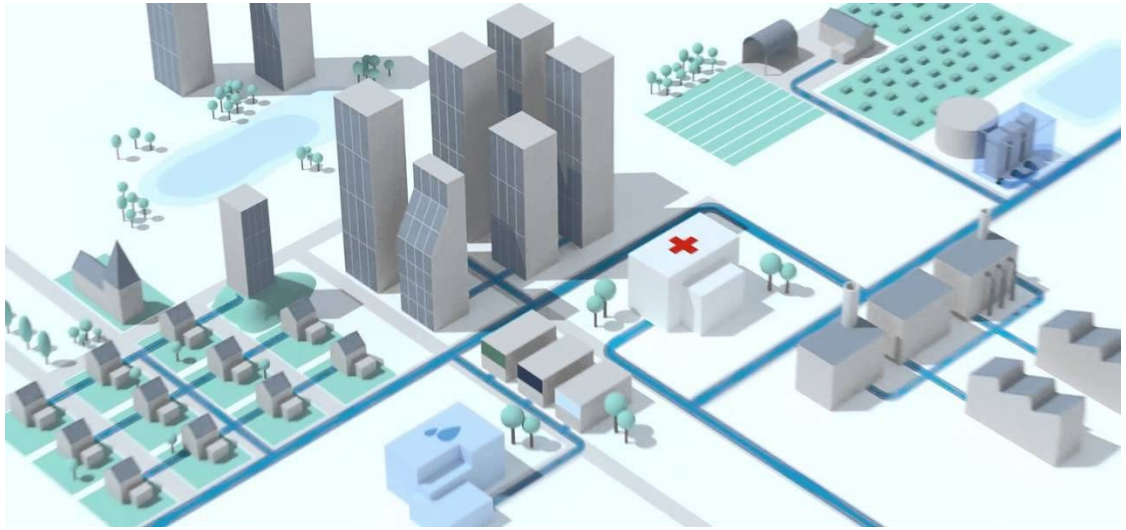
Application: Recommendation

- Pinterest recommendations



Examples: Infrastructure

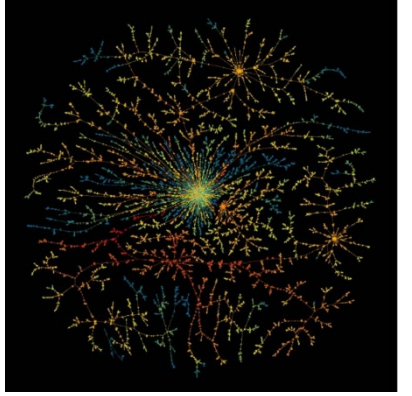
- How Robust are our networks to failures and malicious attacks? How Efficient?
  - Power grid
  - Water supply
  - Airline Networks



Water supply distribution networks

Airline networks

Analysis of Networks



- **How do we start Analysis of Networks?**
  - To understand how the **structure and evolution of the network** (topology) impacts the system
  - To understand network **processes and dynamics** that unfolds on the "structural skeleton"
- Questions that we can ask:
  - What are the **Patterns and statistical properties** of network data?
  - Can we **model** the networks?
  - Why networks are the way they are? Can we find the **underlying rules** that build these networks?
  - How does the structure **evolve** over time?
  - Can we **predict behavior**?

Why Now?

- **Universal language for describing complex data**
  - Networks from science, nature, and technology are more similar than one would expect
- **Shared vocabulary between fields**
  - Computer Science, Social science, Physics, Economics, Statistics, Biology
- **Data availability & computational power**
  - Web/mobile, bio, health, and medical
- **Impact!**
  - Social networking, Social media, Drug design

What will we talk about in this class?

- We will observe **Network Properties** (Diameter, Scale-free/power law network, Small-World behavior)
- We will create **Network Models** that fit our observations (Erdos Renyi random graphs, Kleinberg's model models etc).
- And we will create **Algorithms** that could unfold on our networks
  - Predict the type/color of a given node (Node classification)
  - Predict whether two nodes are linked (Link Prediction)
  - Identify densely linked clusters of nodes (Community Detection)
  - Measure similarity of two nodes/networks (Network Similarity)

Networks vs. Graphs

- **Network** refers to a **real system**, while **Graph** to a **mathematical representation** of a network
  - **Node** vs. **vertex**, **link** vs. **edge**
  - We will use the terms interchangeably

What graphs represent?

- It is important to chose **proper graph representation** of a real system
  - Connecting all the scientific papers that cite each other will result in a **citation graph**
  - Connecting all airports that have direct flights will allow us to explore **flight route graph**
  - Connecting all the actors that co-appeared in the same move together will result in a **co-appearance graph** (e.g, Kevin Bacon graph)
  - Connecting all the actors that share the same consonants in their family names will result **in what?..**
- **Choice of the proper network** representation of a given domain/problem determines our ability to use networks successfully
  - The way **nodes and links are assigned** determines the type of questions we can study
  - Sometimes there are multiple ways to represent the system
    - E.g., Social Networks:
      - 1) Nodes=people, Edge=if they are friends on Facebook
      - 2) Nodes=people, Edge=if they follow each other on Twitter

Complexity of the Networks (cont.)

- Take an social network of 6 people
  - Claim: “I would able to find either 3 mutual friends or 3 mutual strangers (or both) on any instance of a social network of size 6”.
  - Can you prove it?
- What if we have 5 people? Am I guaranteed to find 3 mutual friends or strangers?
  - Can you show a counter example?
- What if we have 7 people?
- **Ramsey’s** number R(3,3)=6
- Can be extended to R(r,s)
- R(4,4)= ?
  - R(4,4)= 18
- R(5,5)= ?
  - R(5,5)= we don’t know (is between 43-48)
- **Why don’t we solve it with brute force? ;-)**
  - How many networks a computer would need to check for 43?

Complexity of the Networks

- How many different networks can we produce if we have N nodes?
  - How many possible links in your class?
  - M=0,5*N*(N-1), i.e., N choose 2
    - If N=10, we have M=45
    - If N= 30, we have M=435
    - If N=100 we have M=4950
  - The number of possible networks 2^M
    - Wikipedia says that number of atoms in the observable universe is 2^{80}

Complexity of Networks (cont.)

- How many **different networks** can we produce if we have N nodes?
  - How many possible links in your class?
  - M=0,5*N*(N-1), i.e., N choose 2
    - If N=10, we have M=45
    - If N= 30, we have M=435
    - If N=100 we have M=4950
  - The number of possible networks 2^M
    - Wikipedia says that number of atoms in the observable universe is 2^{80}
  - Coming back to Ramsey's number: for 43 nodes the number of possible networks to naïvely evaluate is 2^{903}

"Erdős asks us to imagine an alien force, vastly more powerful than us, landing on Earth and demanding the value of R(5,5) or they will destroy our planet. In that case, he claims, we should marshal all our computers and all our mathematicians and attempt to find the value. But suppose, instead, that they ask for R(6,6). In that case, he believes, we should attempt to destroy the aliens."

Course Structure

- 13 Lectures
- 6 scientific paper reviews: 2 for Reaction Paper and 4 for peer-review
- Project proposal
- Final Project
- Final Grade:
  - Reaction Paper 10%
  - Project proposal - Pass/Fail
  - Project Report 40%
  - Project Presentation 10%
  - Exam 40%

Literature and Content

- Content is based on:
  - Jure Leskovec, A. Rajaraman and J. D. Ullman, ”**Mining of massive datasets**” Cambridge University Press, 2012
    - Some slides are from Online Coursera Course based on that book, as well as on Stanford’s course ”Analysis of Networks” by J.Leskovec
  - David Easley and Jon Kleinberg “**Networks, Crowds, and Markets: Reasoning About a Highly Connected World**” (2010).
  - John Hopcroft and Ravindran Kannan ”**Foundations of Data Science**” (2013).
  - And many research papers...

Similar Courses

- **Analysis of Networks**, J.Leskovec
- **The Structure of Information Networks**, Jon Kleinberg
- **Networks: Theory and Application**, Lada Adamic
- **Structure and Dynamics of Networked Information**, David Kempe
- **Information Networks**, Panayiotis Tsaparas

Prerequisites

- Probability and statistics
- Linear algebra
- Programming (be able to write non-trivial programs)

Course Team

- **Teacher and Examiner**
  - Sarunas Girdzijauskas
  - sarunasg@kth.se

- **TA and Lecturer**
  - Zekarias Tilahun

  zekarias@kth.se

Course Project

- **Teams of ~4 students** each (please start forming groups now, but the latest end of this week).
- Project within (Network) Data Mining Domain
- **Task 1.1:** Select 2 papers (per person) for the review.
  - start now, but at the latest Monday March 28
- **Task 1.2**: Reaction Paper - write a review/critique for the selected papers (individual).
  - Deadline to submit Friday April 8
- **Task 2**: Peer reviews – two reaction papers to review (per person)
  - Deadline Friday April 15
- **Task 3**: Project proposal, 2-page proposal. You should have in mind the dataset you are going to work on.
  - Deadline to submit Friday April 22
- Presentation/Feedback session on Friday May 6
- **Task 4**: Project Milestone (Draft reports uploaded and short oral presentations to the TA)
  - On Tuesday May 10

Task 5.1: Project Presentations
  - On Tuesday May 24.
- **Task 5.2**: Final Project Report
  - Deadline June 11

Project Ideas

- **Prediction of Fake news**
  - Online news articles and references
  - Clustering, community detection
- **User Classification in Social Networks**
  - Social network datasets, Github network
  - Clustering, community detection, label propagation
- **Link Prediction in Social Networks**
  - Social network datasets
  - Link Prediction, Centralities
- **Recommendations for X system users**
  - E.g., Amazon, Netflix datasets
  - Recommender systems, Link prediction
- **Predicting advancements in career based on early patronage**
  - Chinese Political Elite Dataset
- **Analysis of Blockchain Crypto-currency transaction/currency flow networks**
  - Bitcoin forensics
  - Node Centralities, community detection
- **Analysis of evolution of Communities in Public Transport Networks**
  - Urban commuting dataset
  - Community Detection, Link Prediction
- **Robustness of Public Infrastructure Networks (Power Grid/Airport/Railway networks etc).**
  - Navigability, Shortest path, Connected components.
- **Detecting Heavily affected Cities after Natural disasters**
  - Facebook Movement Data
  - Community Detection

Link Prediction, Recommender systems, Clustering, Community detection/Anomaly detection, Identifying the most influential nodes or ties (edges), Network robustness, Navigation in the networks etc.

Project Ideas (Cont.)

- Search for the **available datasets**.
  - Identify a domain that intersts you!
- Think what **interesting (network) mining problem** you could come up on that dataset from that domain
- Look through **the papers that are relevant** to your problem and select them for the Reaction Paper
- This will lead to **a good Project Proposal**

Reaction Paper

- **Task 1: Reaction paper (2-3 pages)**
  - The goal of the reaction paper is to
    - familiarize more in depth with the topics covered in class
    - go beyond what will be covered in class
  - Serve as a basis for project proposal
  - *Task 1.1* Each person picks two papers that are clearly related to course topics
    - Links are given to paper suggestions
    - Feel free to suggest your own papers.
      - Papers should be approved by me or the TA
    - The template is provided in CANVAS.

Task 1.2: Reaction Paper - Contents

- Think beyond what you read!
- The reaction part of the paper should address the following questions:
- **Summary**
  - What is main technical content of the papers?
  - What is the connection between the papers you are discussing?
- **Critique**
  - What are strengths and weaknesses of the papers and how they be addressed?
  - What were the authors missing?
  - Was anything particularly unrealistic?
- **Brainstorming (that then leads to the project proposal)**
  - What are promising further research questions in the direction of the papers?
  - How could they be pursued?
  - An idea of a better model for something? A better algorithm? A test of a model or algorithm on a dataset or simulated data? Comparing to other algorithms? Or analyzing datasets from other domains?
- **Conclusion**
  - The reaction paper should be concluded with a section with a description of some promising further research directions and questions, and how could they be pursued.

Reaction Paper (cont.)

- Reaction paper should not just be summaries of the papers you read!
  - The purpose of the reaction paper is to survey the related work and identify what are strengths and weaknesses of the papers and how they may be addressed.
- Special focus on the “Brainstorming” part
  - Answering the questions within that part can be a very good way to explore a potential project topic.

Peer Reviews

- Each of you will be assigned to review two reaction papers submitted by two of your colleagues.
- **Before evaluating the reaction papers, please carefully read the two papers that each reaction paper discusses first**.
  - This will bring the total number of papers you need to read to six.
- Peer Review instructions are given in Canvas

Project Proposal

Task 2: Project proposal (1-2 pages).

- Preferably builds on the reaction papers of the group members
- Focus on what are some promising further research directions and questions:
  - How precisely do you plan to pursue them?
  - What methods/data do you plan to use?
  - You should try to provide a concrete proposal for a model(s) or algorithm(s) that potentially extend(s) or improve(s) the topics discussed in the papers you've read.

Project Proposal - Contents

The following questions should be answered

- What is the problem you are solving?
- What data will you use (how will you get it)? – we will provide you with initial links
- What work do you plan to do the project?
- Which algorithms/techniques/models you plan to use/develop? Be as specific as you can!
- How will you evaluate your method? How will you test it? How will you measure success?
- What do you expect to submit/accomplish by the end?

Project Proposal (cont.)

- The project should contain some experimentation on real and/or synthetic data
  - Big plus if some amount of mathematical analysis is present too.
- The result of the project will typically be an approx. 8 page paper, describing the approach, the results, and the related work.

- No Cheating!

Project Milestone

- Project Milestone (3-5 pages) is a draft of your final report but without your major results
- Provide a complete picture of your project even if certain key parts have not yet been implemented/solved.
- Include the parts of your project which have been completed so far, such as:
  - Thorough introduction of your problem
  - Review of the relevant prior work
  - Description of the data collection/processing
  - Description of any initial findings or summary statistics from your dataset
  - Description of any mathematical background necessary for your problem
  - Formal description of any important algorithms used
  - Description of general difficulties with your problem which bear elaboration
- Make sure to at least outline the parts which have not yet been completed so that it is clear specifically what you plan to do for the final version.
- **Give an oral update (15min per group) on the project to the TA and me (probably on May 6th, will doodle the exact time).**

Project Report

- The final project report is max 8 page paper, describing the introduction, related work, approach, results and conclusion.
  - You will be graded individually
  - Write a brief summary of the **contributions of individual team members** to the project
- Final presentation session on May 24th.
- Grand prize to the winning team!

Project Report – Contents and Evaluation

- **Introduction/Motivation/Problem Definition (15%)**
  - What is it that you are trying to solve/achieve and why does it matter.
- **Related Work (10%)**
  - How does your project relate to previous work. Please give a short summary on each paper you cite and include how it is relevant.
- **Model/Algorithm/Method (30%)**
  - This is where you give a detailed description of your primary contribution. It is especially important that this part be clear and well written so that we can fully understand what you did.
- **Results and findings (35%)**
  - Evaluation of the solution to whatever empirical, algorithmic or theoretical question you have addressed and what do these evaluation methods tell you about your solution.
  - It is not so important how well your method performs but rather how interesting and clever your experiments and analysis are. **Negative results are fine!!**
  - We are interested in seeing a clear and conclusive set of experiments which successfully evaluate the problem you set out to solve. Make sure to interpret the results and talk about what can we conclude and learn from your evaluations.
- **Style and writing (10%)**
  - Overall writing, grammar, organization and neatness.