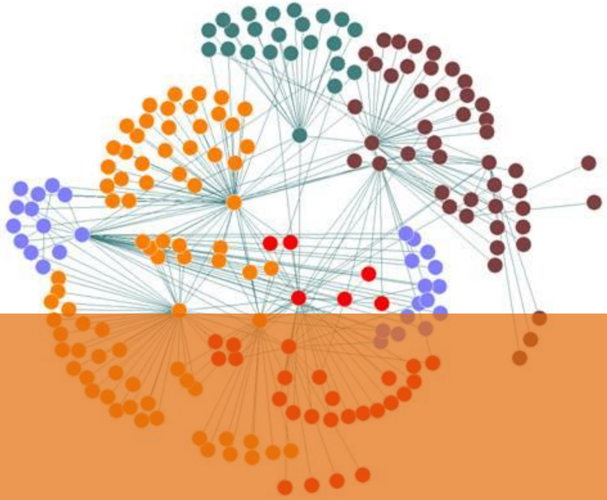


BABEŞ-BOLYAI UNIVERSITY
Faculty of Mathematics and Computer Science

Social Network Analysis

Project Requirements

Camelia Chira

camelia.chira@ubbcluj.ro

Evaluation: seminar (project)

Grade2: 50% of the final grade

Project implementation and presentation

- *Collect data, create and perform a full analysis of a complex network (details given in the seminar)*
- Team of 3-4 students
- Preliminary presentations: seminars 3-4
  - Show what is the data and idea of the project
- Final presentations: seminars 6 and 7
  - Present network analysis results (each student in the team will present his/her own contribution)
  - Submit code, data and experimental results

Team

- 3-4 students
- Possible tasks
  - Data acquisition
  - Network analysis
  - Network visualization
  - Important nodes / Communities / Spreading / etc
- Decide your team by Seminar 2/3

Project components

- **Part 1: Dataset analysis**
  - Choose network data from a repository
  - Perform a full analysis of the network

  Choose the dataset by Seminar 3.
- **Part 2: Real data analysis**
  - Collect data
  - Choose a network representation
  - Perform network analysis
- **Preliminary project presentation (Seminar 4)**
- **Final project presentation (Seminar 6-7)**

Part 1: Dataset analysis

- Choose ***one*** dataset from a network repository
  - SNAP repository: https://snap.stanford.edu/data/index.html
  - Network repository: http://networkrepository.com/
  - Pajek datasets: http://vlado.fmf.uni-lj.si/pub/networks/data/
  - Koblenz Network Collection: http://konect.cc/
  - Mark Newman's collection: http://www-personal.umich.edu/~mejn/netdata/

- The network should have at least 500 nodes and 2000 links
- Create two artificial networks starting from the real network selected:
  - The first should be a random network (Erdos-Renyi model)
  - The second should be a scale-free network (Barabasi-Albert model)
- The artificial networks should have the same number of nodes and similar number of links

Part 1: Dataset analysis

For each of the three networks (original, random and scale-free):

1. Provide a detailed **description** of the network: number of nodes, links, type of network, significance and meaning
2. Compute **network properties**: average degree, diameter, connected components and the size of the largest one, shortest paths, average clustering coefficient
3. Compute and plot: degree distribution, clustering coefficient distribution, betweenness centrality distribution, connected components size distribution
4. Identify most **important nodes** according to different measures
5. Provide a **visualization** for the network.
6. What type of network is the selected real network?

Part 2: Real data analysis

- Data acquisition
  - Download data
- Network representation
  - What are the nodes and links?
- Network analysis
  - Questions you want to answer
  - Network tools and metrics
  - Network visualization

Part 2: Real data analysis

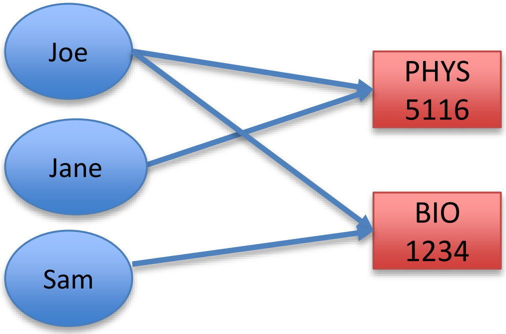
- **Data acquisition**
- Many online data sources have an **API** that allows querying and downloading the data in a targeted way
  - Example: What are all movies from 1984-1995 starring Kevin Bacon and distributed by Paramount Pictures?
  - This is done either through a web interface or through a library within a programming language
- Other sources will provide raw bulk data (e.g., Excel spreadsheets) that require processing, either manually or through a program you will write

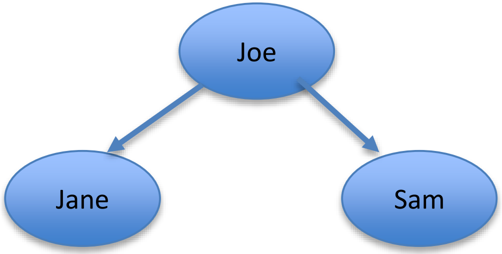
Part 2: Real data analysis

- **Network representation**

- Most datasets will admit more than one representation as a network
- Some representations will be more or less informative than others
- Figuring out the "network" that's buried in your data is part of your project!

Suppose you have a list of students and the courses they are registered for





Part 2: Real data analysis

- **Network analysis**
  - Degree distribution, clustering coefficient, shortest paths, diameter, connected components
  - Important nodes, centrality measures
  - Communities, Information spreading
  - Visualization and plots
  - Compare to random / small-world / scale-free model

Part 2: Real data analysis

- **Network analysis**
  - Measure: N, L (time dependence?), P(k) degree distribution, <l> average path length, C(k) clustering coefficient
  - Visualize communities, network robustness and spreading (whatever if appropriate for your network)

- It is not sufficient to measure things - discuss your insights:
  - What is the meaning of each quantity measured? What did you learn?
  - What was your expectation?
  - How do results compare to your expectation?

Project presentations

- Preliminary presentations: **Seminar 4**
  - Show results for dataset analysis
  - Present 3 slides to show your ideas for real data analysis: introduce the network (nodes, links), how will you collect the data (make sure N>100), what questions do you plan to ask and why
- **Final project presentations: Seminar 6-7**

Project grading

- Dataset analysis
- Network analysis
- Use of network tools and measures
- Extract information and insights from data
- Quality of project presentation

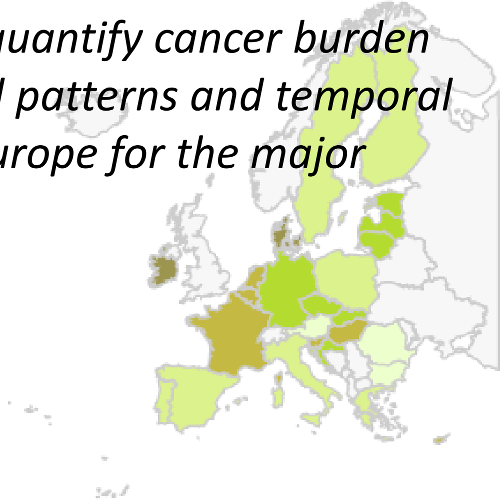
Timeline

Seminar	TO DO
Seminar 2	Decide your team
Seminar 3	Choose the dataset and perform the analysis
Seminar 4	Preliminary presentations: show dataset results and present your ideas for the real data analysis
Seminar 5	Continue real data analysis, decide tasks, get feedback
Seminar 6	Project final presentations
Seminar 7	

Ideas: Health

*“**ECIS** provides the latest information on indicators that quantify cancer burden across Europe. It permits the exploration of geographical patterns and temporal trends of incidence, mortality and survival data across Europe for the major cancer entities.”*

https://ecis.jrc.ec.europa.eu/

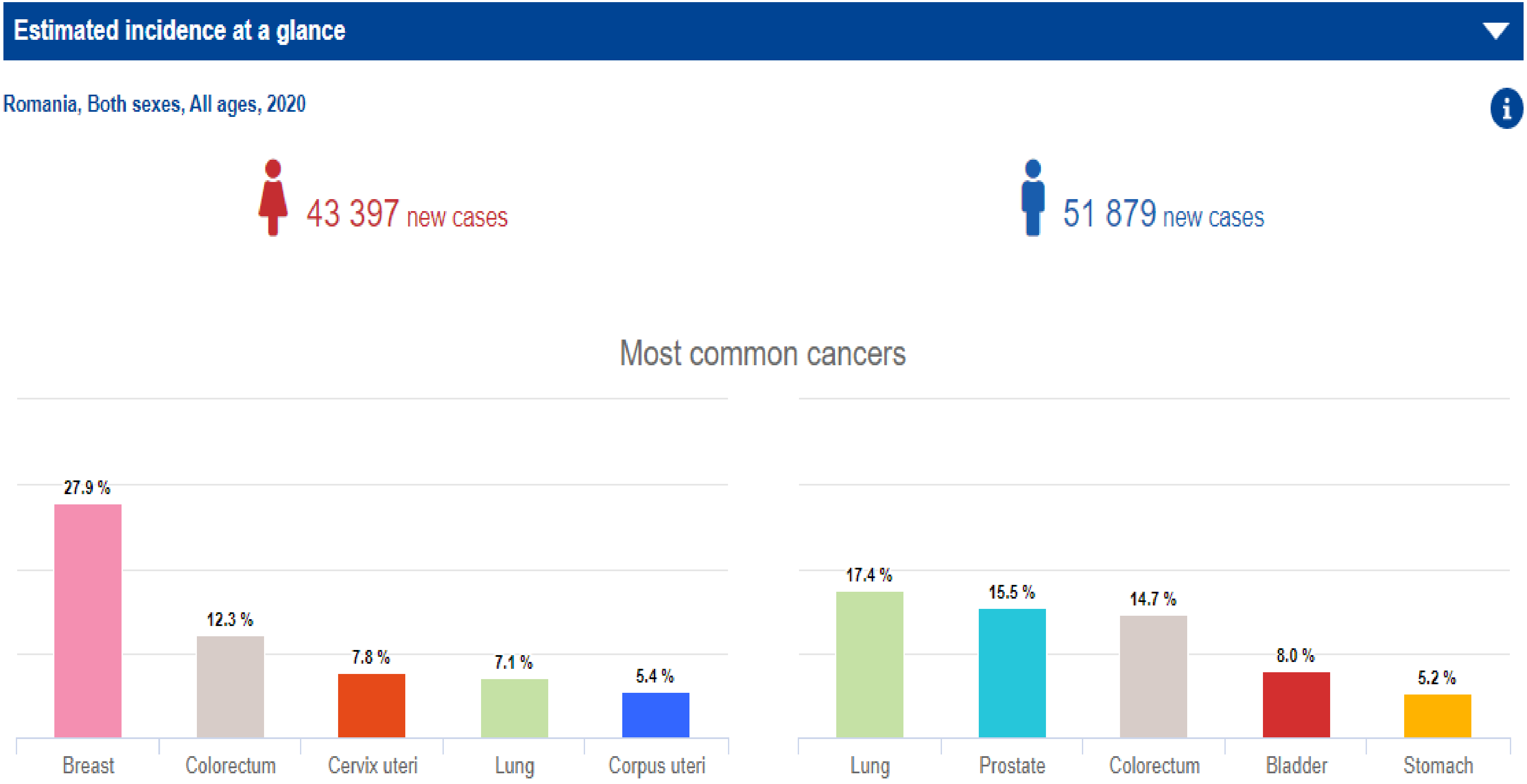


- Possible investigations
  - Correlations between location, prevention methods and measured indicators
  - Breast cancer: correlate screening data and indicators across regions
  - What can we do to improve screening and prevent breast cancer?

Incidence and mortality estimates 2020

National estimates of cancer incidence and mortality in 2020, for the major cancer sites in European countries.

https://ecis.jrc.ec.europa.eu/



Estimated incidence at a glance

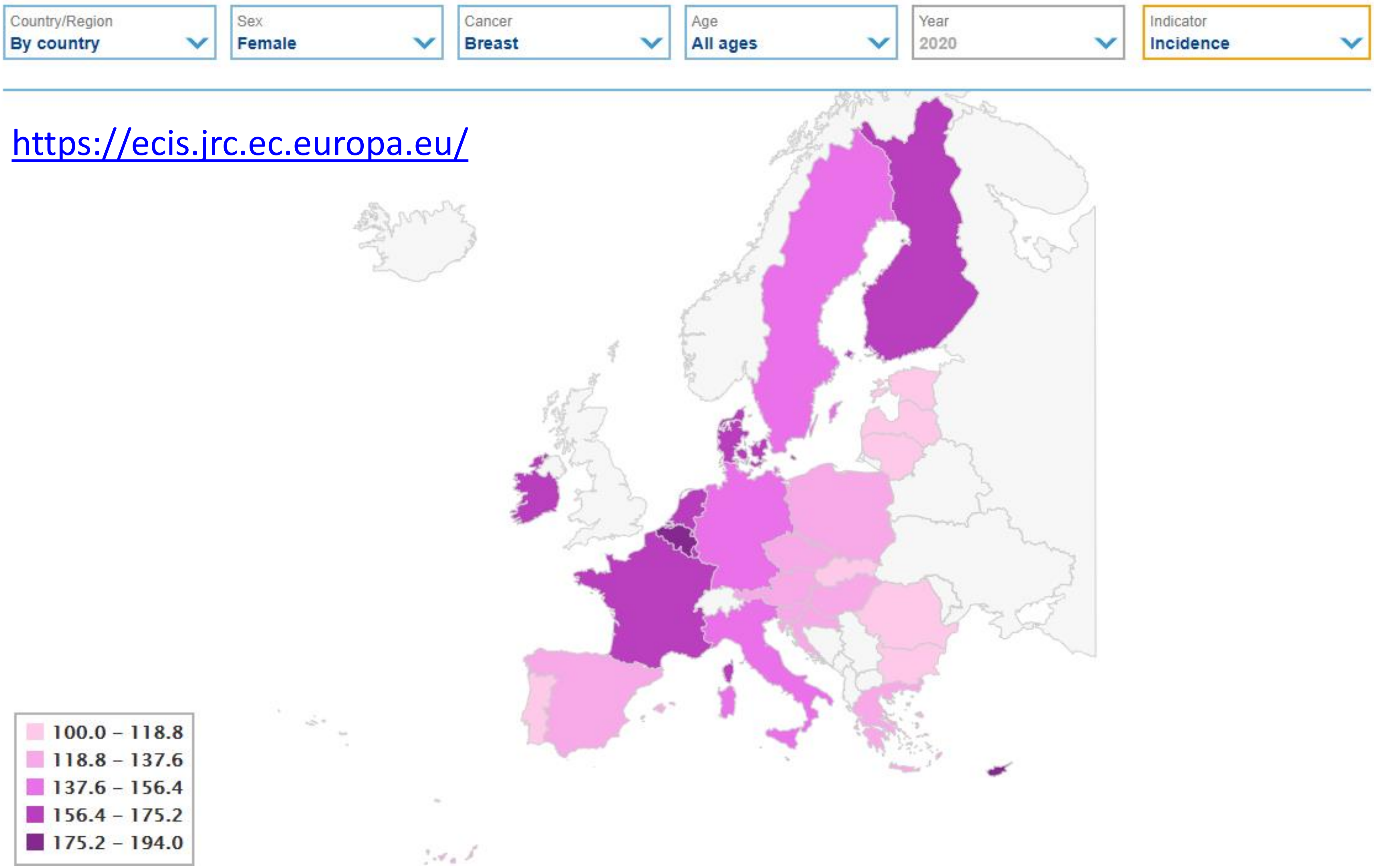
Romania, Both sexes, All ages, 2020

43 397 new cases

51 879 new cases

Most common cancers

https://ecis.jrc.ec.europa.eu/



Country/Region: By country

Sex: Female

Cancer: Breast

Age: All ages

Year: 2020

Indicator: Incidence

▨	100.0 – 118.8
▨	118.8 – 137.6
▨	137.6 – 156.4
▨	156.4 – 175.2
▨	175.2 – 194.0

Ideas: Economics

Data about financial transactions, companies, online sales, etc

Use UCI repository data and analyse it with network tools
Ex. Online Retail Dataset
https://archive.ics.uci.edu/ml/datasets/online+retail#

- Possible investigations
  - Similarity between customers
  - Identification of important nodes and chains
  - Community detection
  - Correlations between location, customers, products

Ideas: Books

- Contains books, ratings, reviews, recommendations, etc.

- API available at https://www.goodreads.com/api

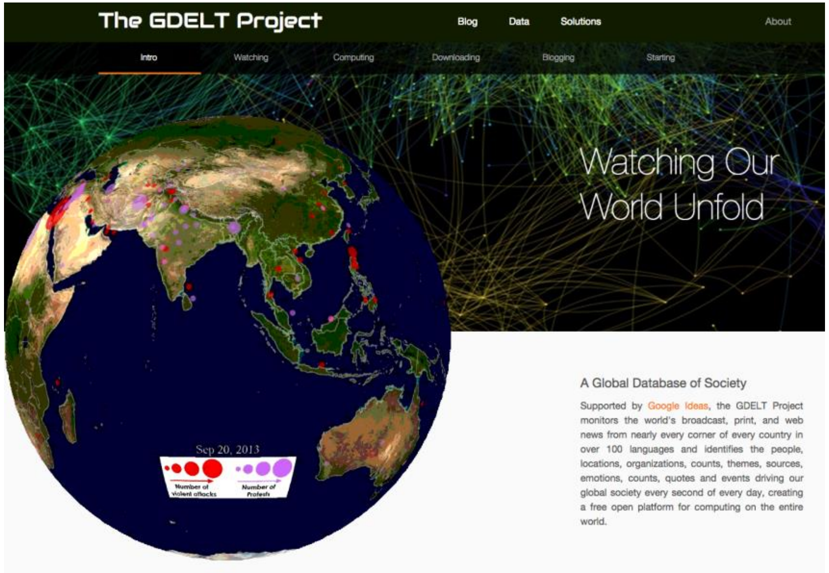
- Potential areas of investigation:
  - Similarity network of books
  - Community detection (discovering genres)

Ideas: Mendeley

http://www.mendeley.com/

- Large scientific publication database/social network for researchers
- API available (dev.mendeley.com)
- Potential investigation: use readership to assign authorship credit
  - Data consist of user profiles + papers the user has read
  - Publications (nodes) are linked if they are both present in one or more users’ lists

Ideas: GDelt



Watching Our World Unfold

A Global Database of Society

Supported by Google Ideas, the GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

https://www.gdeltproject.org/

- Data about monitoring news (broadcast, print and web) from 1979 to today in the entire world. It identifies names, places, organizations, emotions, counts.
- Offer raw data files and/or possibility of querying a database

- Potential investigation: (i) study the individual – individual network (two individuals are connected if they appear in the same news) over time, see how leaders emerge; (ii) study the network of locations, with two locations connected if the same news is reported. How do news travel over space?

- The dataset can be used for many more projects!

Ideas: Comics

Global comics database

http://www.comics.org/

- Wiki and advanced search interface available

- Varied data about each comic: publisher, who wrote script/penciled/inked, publication date

- Potential areas of investigation:
  - Comics linked by common characters
  - Collaboration network between artists

Ideas: Sport

Databases of statistics at player level (individual stats) and team level (team compositions, hall of fame, managers, etc)

Ex. Baseball data   http://seanlahman.com/baseball-archive/statistics/

- Possible investigations
  - Are there characteristics of the network that distinguish hall-of-famers?
  - Mobility of players/managers across teams
  - etc

Get inspired: Papers

- Wang T, Brede M, Ianni A, Mentzakis E (2018) Social interactions in online eating disorder communities: A network perspective. PLOS ONE 13(7): e0200800. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0200800
- Aiello, L.M., Schifanella, R., Quercia, D. *et al.* Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Sci.* **8,** 14 (2019). https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-019-0191-y
- Obrist Marianna, Tu Yunwen, Yao Lining, Velasco Carlos, Space Food Experiences: Designing Passenger's Eating Experiences for Future Space Travel Scenarios, 1(3), Frontiers in Computer Science 2019, https://www.frontiersin.org/article/10.3389/fcomp.2019.00003
- Asano, Y.M., Biermann, G. Rising adoption and retention of meat-free diets in online recipe data. *Nat Sustain* **2,** 621–627 (2019). https://www.nature.com/articles/s41893-019-0316-0
- Araceli Arellano-Covarrubias, Carlos Gómez-Corona, Paula Varela, Héctor B. Escalona-Buendía, Connecting flavors in social media: A cross cultural study with beer pairing, Food Research International, Volume 115, 2019, Pages 303-310. http://www.sciencedirect.com/science/article/pii/S0963996918309517
- Simas Tiago, Ficek Michal, Diaz-Guilera Albert, Obrador Pere, Rodriguez Pablo R., Food-Bridging: A New Network Construction to Unveil the Principles of Cooking , Frontiers in ICT, vol.4, 2017. https://www.frontiersin.org/article/10.3389/fict.2017.00014

Camelia Chira
http://cs.ubbcluj.ro/~cchira

Content based on the textbook:
A.-L. Barabási, Network Science, Cambridge University Press, 2016.
http://networksciencebook.com/