Data Mining II / Adv. Topics in Data Science

Web Mining: Link Analysis

Álvaro Figueira
Rita Ribeiro

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO
[dcc] DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO

1

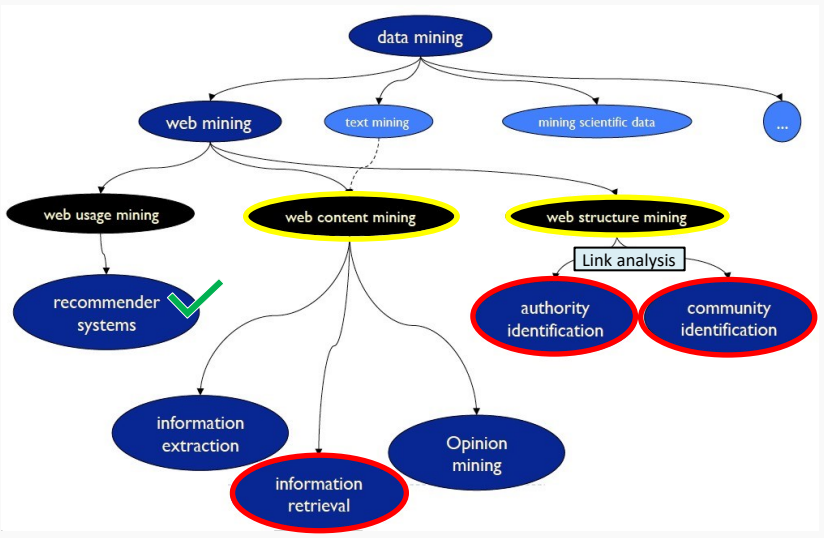
Summary

2

Introduction

3

Data Mining - a structured view



4

Web Structure Mining

- Take advantage of the information in **web hyperlinks**:
  - links are created locally;
  - web structure, as a whole, is not planned.
- Take advantage of the information in **social links**:
  - social networks.
- To understand the **structure of the web**:
  - link analysis
  - analysis of the topology of connections

5

Web Structure Mining: Graphs

- Internet can be seen as a set of different **interdependent graphs**
  - pages and hyperlinks (nodes and edges)
  - computers and communications between them (Internet)
- The Web is a particular type of graph
  - very large (5×10^{10} nodes?)
  - dynamical (frequent changes in structure and content)
  - has virtual parts (dynamic pages, harder to analyze)
  - disconnected (has islands)
  - sparse (relatively few connections)

6

Web Structure Mining: Goals

- The goals are to find:
  - **Central links** in social webs
  - Particular (or **prestigious**) web pages
  - **Communities**
    - web page clusters pointing to each other
    - groups of people who exchange some information
- The study of web structure is related to:
  - Social Network Analysis;
  - Network Science
  - Complex Networks

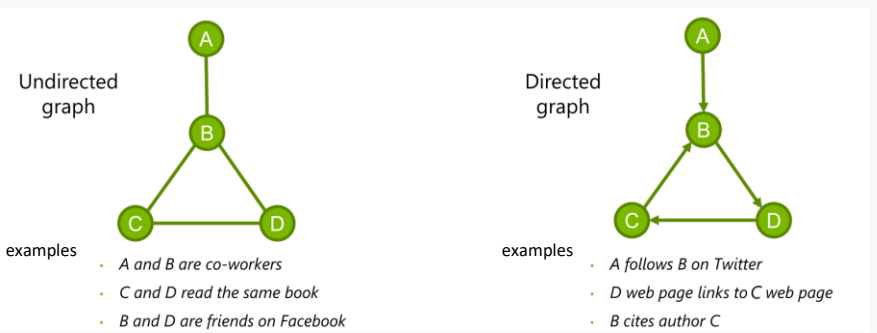
Obs:
A prestigious web page is a web page recognized as being authoritative, influential, and respected in its field

7

Graph Theory concepts

Undirected vs. Directed Graphs

• Graph edges can be undirected (symmetric relationship) or directed (possibly asymmetric relationship)



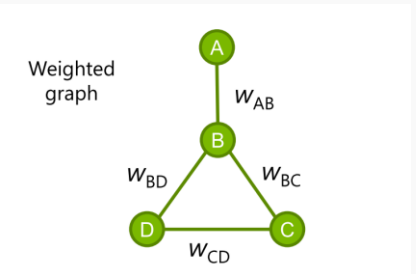
(Next 4 slides adapted from Eduarda M. Rodrigues, PBS)

8

Graph Theory concepts

Weighted Graphs

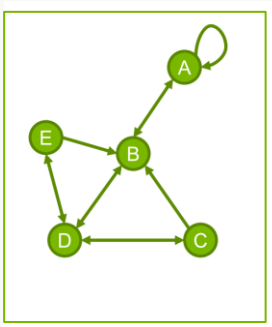
• Graph **edges can be weighted** (e.g., duration of relationship, number of books, number of citations, frequency of interaction, etc.)



9

Graph-theoretic Data Structures

- Graph data can be **represented using different data structures**
- Edge list and adjacency list are more compact representations than adjacency matrix and are better for when the network is large and/or sparse



Edge list

A,A
A,B
B,A
B,D
C,B
C,D
D,B
D,C
D,E
E,B
E,D

Adjacency list

A → A,B
B → A,D
C → B,D
D → B,C,E
E → B,D

Adjacency matrix

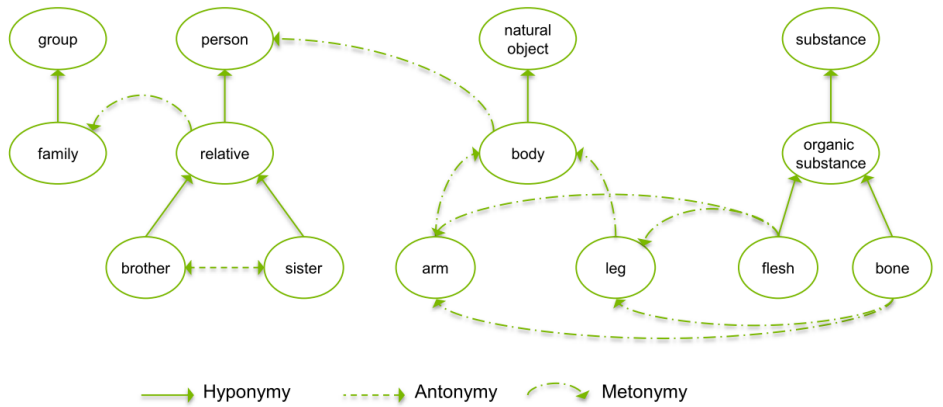
	A	B	C	D	E
A	1	1	0	0	0
B	1	0	0	1	0
C	0	1	0	1	0
D	0	1	1	0	1
E	0	1	0	1	0

Sparse representation

10

Semantic Networks (Wordnet)

Lexical relations between words:



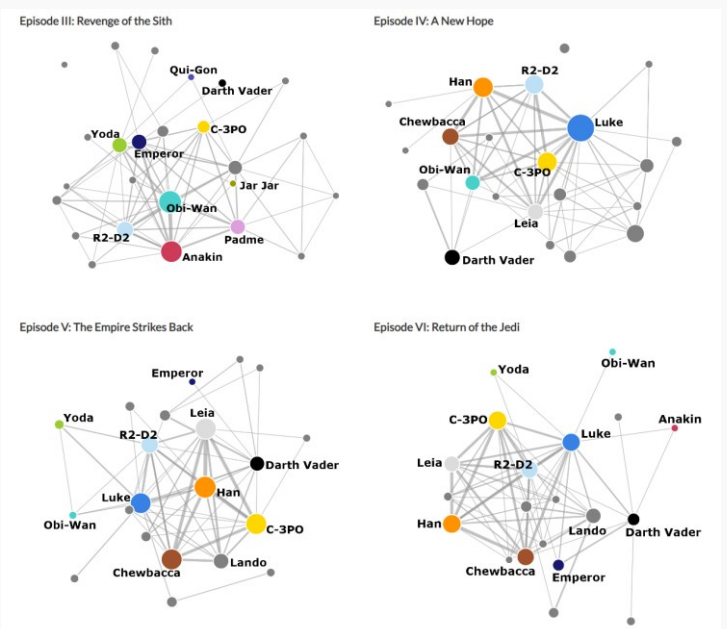
(a specific type of)

(is the opposite of)

(a word is used to represent another)

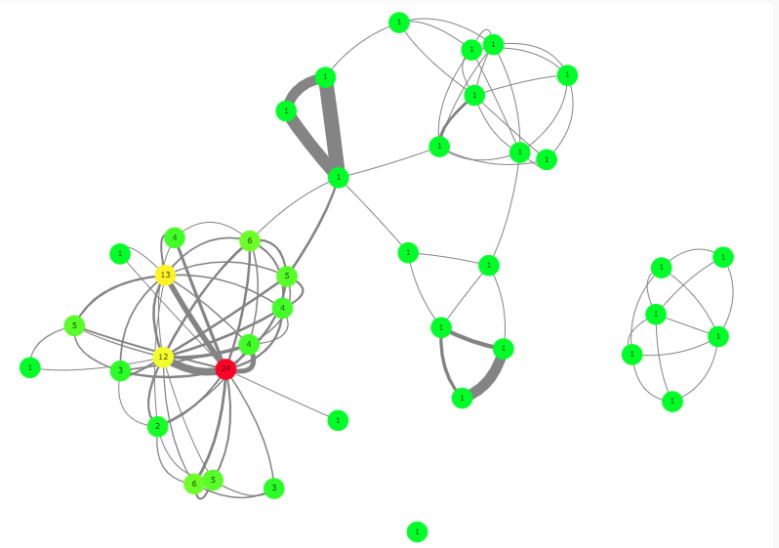
11

The Star Wars saga



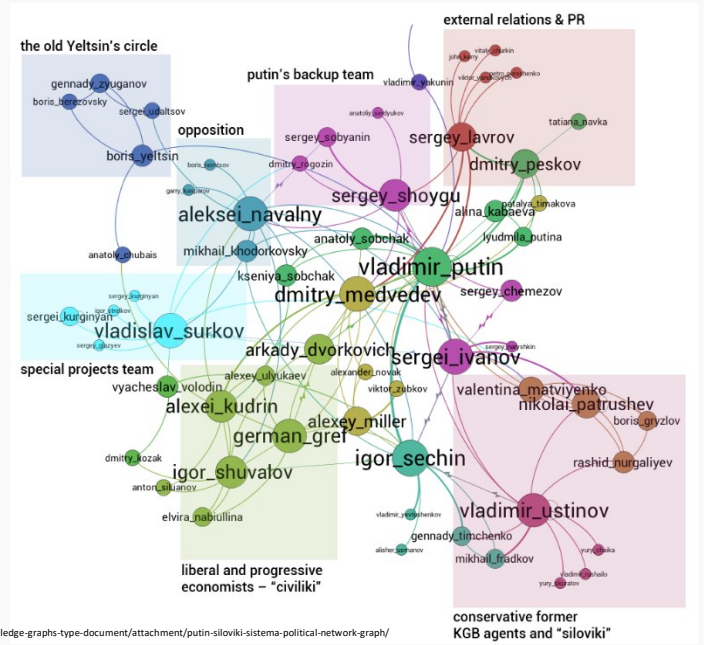
12

Co-author's Network



13

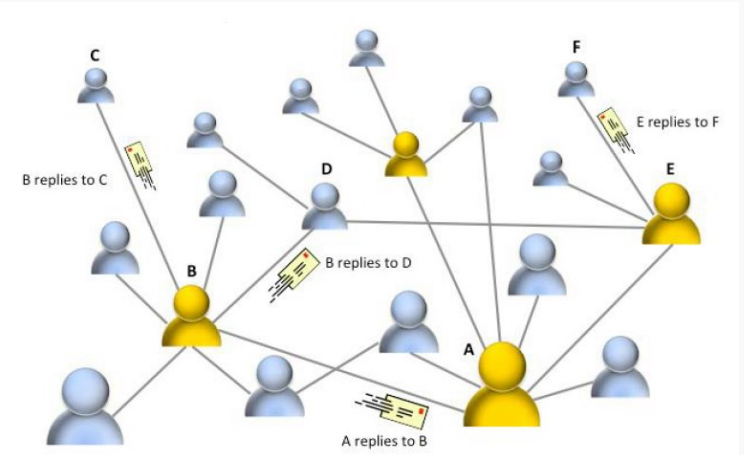
Political / Social Networks



Source:
https://noduslabs.com/research/knowledge-graphs-type-document/attachment/putin-siloviki-sistema-political-network-graph/

14

Social Media Interaction – an online discussion group

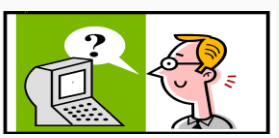


15

Types of participants in online communities (SNA)

- **Leaders**
  - Provide answers and social support, are the institutional memory, and police the space
  - Want visibility and recognition for their efforts

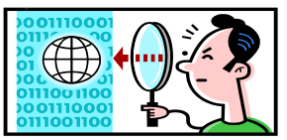
- **Questioners / Curious**
  - Go to the community for help, expecting quick and accurate answers and have various levels of expertise
  - Want to be guided in their community interaction



16

Types of participants in online communities (SNA cont.)

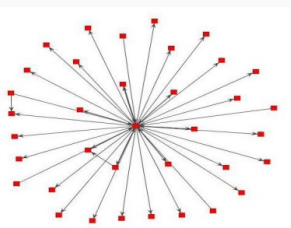
- **Silent searchers / Lurkers**
  - Do not actively participate in the community
  - Just consume the product of the interaction of the leaders and questioners, searching for previous answers to similar problems



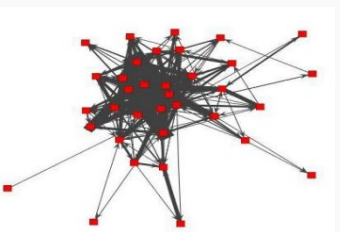
- **Content contributors**
  - Spend a considerable amount of their personal and work time contributing content to the community
  - Write FAQ, how-to articles, share code, etc.

17

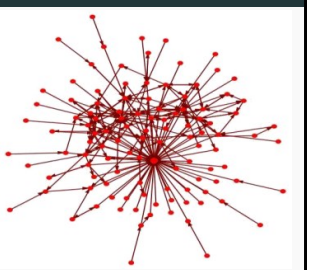
Social Media Interaction - Social Roles



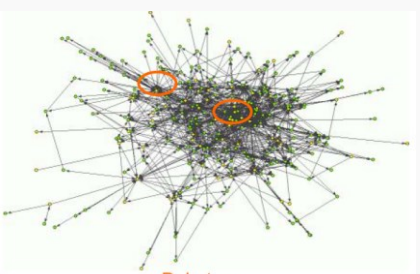
Answer person



Discussion group



Reply/Magnet



Debaters
Political Discussion (guns)

Topic Experts
Technical Discussion

18

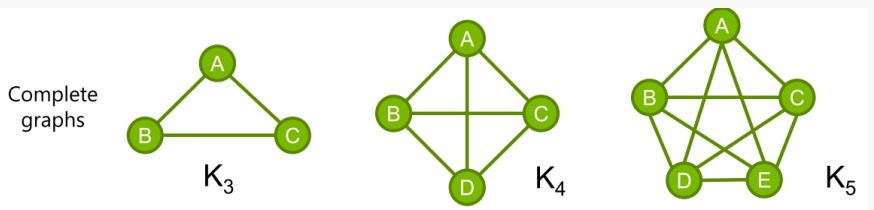
How good is a network visualization?

- Avoid edge crossings and node occlusions
  - standard layout algorithms don't help much when the size of the network is above a few hundred nodes and the network is relatively dense in the number of links
- A coherent network layout algorithm
  - an inadequate choice of network layout algorithm may hinder the discovery of patterns in the network structure
- Lack of contextual information
  - interpretation of the network structure often requires visualizing additional information about the nodes and links
- Not many software toolkits out there – Wikipedia page lists around 15 entries
  - Gephi, GraphViz, NodeXL, igraph, Pajek, …

19

Complete Graphs (Cliques)

- K_n is the complete graph (clique) with n vertices
  - each vertex is connected to every other vertex
  - there are n*(n-1)/2 undirected edges



20

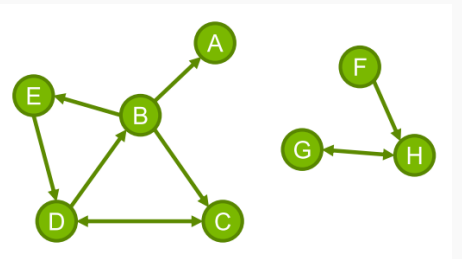
Paths in graphs

- A **graph path** is a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge
- Examples:
  - a passenger taking a **sequence of airline flights**
  - a piece of **information being passed** from person to person in a social network
  - a computer user or piece of software visiting **a sequence of Web pages** by following links
- The **length of a path** is the number of edges in the sequence

21

Shortest Path and Graph Diameter

- The **shortest path**, or geodesic distance, between two nodes is the shortest sequence of links connecting them (may not be unique)
- **Eccentricity of a node**: largest shortest path distance between the node and any other node in the graph
- The **graph diameter** is the largest shortest path in the graph



Node	Eccentricity
A	-
B	2
C	3
D	2
E	3
F	2
G	1
H	1

Exercise: Find the eccentricity of each node. What is the shortest path between D and A? Diameter?

22

Special Types of Graphs

Tree:
- is a connected acyclic graph (i.e., no cycles exist)
- Two nodes have exactly one path between them

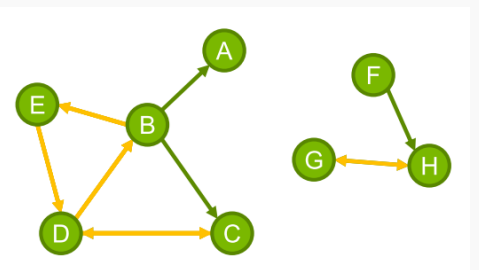
Clique:
- A fully connected graph
- No node is more important than any other

Path:
- Two nodes have exactly one shortest path between them

23

Strong and Weak Components

- Each node within a **strongly connected component** (**SCC**) can be reached from every other node in the component by following directed links
- Each node within a **weakly connected component** (**WCC**) can be reached from every other node in the component by following links in either direction



WCC:
- A, B, C, D, E
- F, G, H

SCC:
- A
- B, C, D, E
- F
- G, H

Exercise: List all the WCCs and SCCs of this network.

24