Quantifying Network Structure: Basic Properties, Centralities, Communities

IV124

Josef Spurný & Eva Výtvarová

Faculty of Informatics, Masaryk University

March 17, 2023

Networkx

- Python library
- basic functions for manipulation and analysis
- basic visualization using `matplotlib`
- http://networkx.github.io/

Useful for:

- interactive work using `ipython`
- scripting
  - batch processing
  - reproducibility!

igraph http://igraph.org/ as an alternative

Gephi

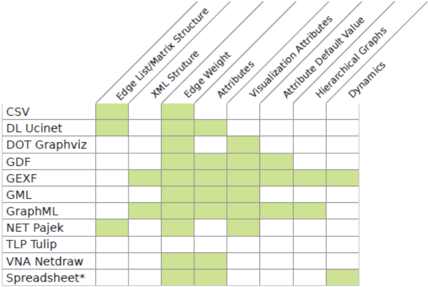
- multiplatform graphic app (Java)
- robust visualization
- many measures and plugins
- time-varying and dynamic graphs

Useful for:

- interactive exploratory analysis
- visualization

Cytoscape `http://www.cytoscape.org/`, UCINET `http://bit.ly/1zkNUk6`, yEd `http://bit.ly/1piw0j0` as alternatives

Gephi – file formats



	Edge List/Matrix Structure	XML Struture	Edge Weight	Attributes	Visualization Attributes	Attribute Default Value	Hierarchical Graphs	Dynamics
CSV	🟩		🟩					
DL Ucinet	🟩		🟩	🟩				
DOT Graphviz			🟩		🟩			
GDF			🟩	🟩	🟩	🟩		
GEXF		🟩	🟩	🟩	🟩	🟩	🟩	🟩
GML			🟩	🟩	🟩			
GraphML		🟩	🟩	🟩	🟩	🟩	🟩	
NET Pajek	🟩		🟩		🟩			
TLP Tulip								
VNA Netdraw			🟩	🟩				
Spreadsheet*			🟩	🟩				🟩

GEXF file format

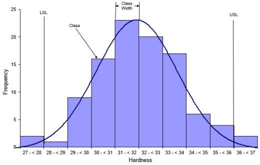
```
    <?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://gexf.net/1.2" version="1.2">
    <meta lastmodifieddate="2009-03-20">
        <creator>Gexf.net</creator>
        <description>A hello world! file</description>
    </meta>
    <graph mode="static" defaultedgetype="directed">
        <nodes>
            <node id="0" label="Hello" />
            <node id="1" label="Word" />
        </nodes>
        <edges>
            <edge id="0" source="0" target="1" />
        </edges>
    </graph>
</gexf>
```

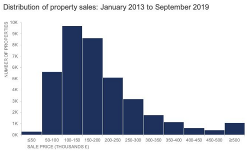
Connected Components, Histogram

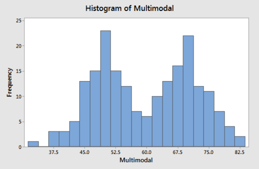
Is a network fully connected?

- strongly connected components
- weekly connected components

Histogram – an approximate representation of the distribution of numerical data




Distribution of property sales: January 2013 to September 2019


Histogram of Multimodal

Node Degree

- # edges connecting a node with others
- # non-zero elements in a row (a column) in an adjacency matrix
- in-degree/out-degree in oriented (directed) networks

Interpretation?

Average node degree

For an undirected network

- $\overline{k} = \frac{2|E|}{|V|}$
- every edge contributes to two nodes

Degree distribution $P(k)$

- probability that a random node has a degree of k
- an average is not enough description parameter in real networks

Paths and distance

A path in a network

- a sequence of edges connecting two nodes
- path length in unweighted networks: # edges
- path length in weighted networks: depends on the semantics

A distance of two nodes d

- a length of the shortest path
- there can be more than one shortest path
- $d = \infty$ if two nodes are unconnected

A network diameter D

- the longest distance between any two nodes in a network

Paths and distance II.

Computing distance

- Unweighted network
    - breadth-first search
- Weighted network
    - Dijkstra algorithm

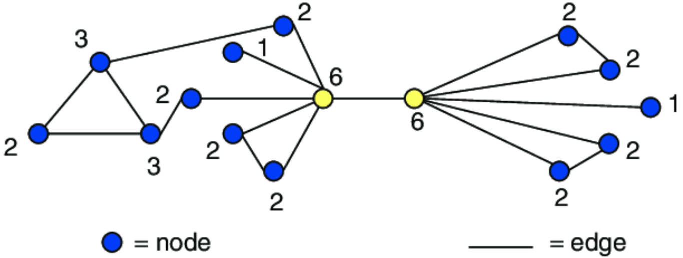
Average path length $\overline{d}$

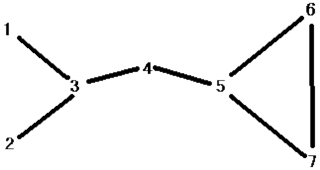
- all-to-all
- Floyd-Warshall algorithm

Interpretation

- efficiency of spreading e.g. information
- network integration

Path length – examples





Node Importance

Types of questions we are interested in:

- which individuals are key for disease spreading?
- how to target attacks on a network?
- how to improve spreading information in an organization?
- which web pages are more valuable than others?
- which people have the highest influence on forming group opinion?
- ...

Centrality as Node Importance

The importance of a node depends on:

- its attributes
- *location in the network*

A choice of a suitable measure depends on:

- research question
- semantics of particular network of interest

Node Degree as Centrality

Node with high degree:

- is highly connected to the rest of the network
- has a direct impact on a large number of other nodes (neighbours)

In directed network:

- in-degree and out-degree
- substantial difference in interpretation!

Node degree does not indicate the importance of its neighbours

Node Degree: Example I.

World Trade Network

- directed network
- degree refers to the number of business partners
  - in-degree: import
  - out-degree: export

An evolution of nodes with the highest degree centrality reflects structural changes in world trade

- higher overall connectedness (lesser differences)
- changes in the composition of the most central group

World Trade Network[1]

in-degree			out-degree		
1960					
1.	0.6438	UK	1.	0.5987	USA
2.	0.5954	Netherlands	2.	0.5861	UK
3.	0.5866	France	3.	0.5740	France
2000					
1.	0.8920	USA	1.	0.8636	USA
1.	0.8920	Germany	1.	0.8636	UK
3.	0.8808	UK	1.	0.8636	France

[1] De Benedictis, L., & Tajoli, L. (2011)

Node Degree: Example II.[2]

Protein network of *Helicobacter pylori* bacteria

- undirected network, edges represent known physical interaction (catalysis, signalization...)
- elimination of a specific protein has known effects

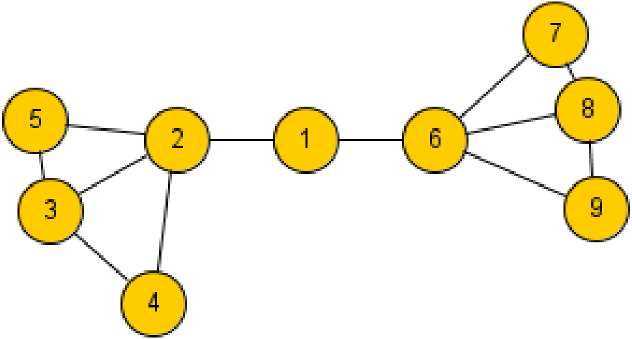
Network Robustness

- relatively high tolerance to random mutations
- removal of high node degree proteins is fatal
- correlation between node degree and severity of consequences $r = 0.75$

[2] Jeong, Hawoong, et al. (2001)

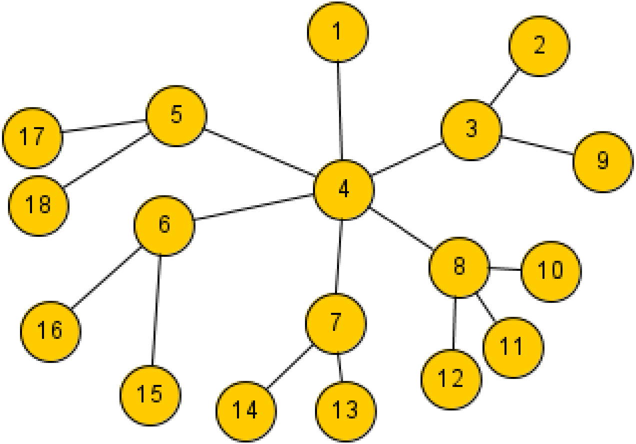
Shortest Path and Centrality

Even nodes with a low degree may be important



Shortest Path and Centrality

Even nodes with a low degree may be important



Closeness Centrality

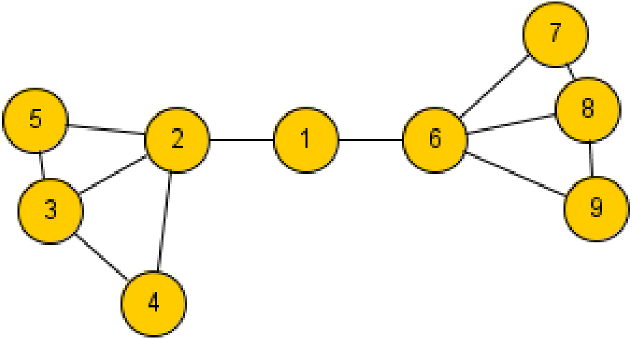
"To be in the center of action"

- inversely proportional to avg shortest path to other nodes
- advantageous position for spreading information in terms of influencing other nodes

Definition

- $C_c(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$
- normalized $C'_c(i) = \frac{C_c(i)}{N-1}$

Betweenness Centrality



Betweenness Centrality

Represents brokerage

- nodes connecting clusters
- advantageous position for spreading information

Definition

- $C_b(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}$
- g_{jk} is the number of shortest paths between j and k
- $g_{jk}(i)$ is the number of shortest paths between j and k which go through i

Betweenness Centrality: Example [3]

Co-authorship network (library and information science)

- nodes: authors, links: an article written together
- analysis of the author’s impact (number of citations of all articles)

- betweenness centrality correlates with impact
- node degree indicates the number of co-authors
- betweenness refers to interdisciplinary projects

[3] Yan, E., & Ding, Y. (2009)

Betweenness Centrality: Example[4]

Network of transferring patients between hospitals

- nodes: hospitals in USA, links: transfers between ICU
- a case of spreading treatment-resistant infection

The problem of allocating limited resources for quarantine

- random, by degree, by betweenness, iteratively by disease exposure
- betweenness proved to be the best from static (preventive) allocations

[4] Karkada, Umanka H., et al. (2011)

Centralities: Differences

low high	degree	closeness	betweenness
degree		in the middle of a cluster, distant from the rest of the network	links of a node are redundant for the network
closeness	a node immediately close to an important node		alternative shortest paths, many nodes are close to each other
betweenness	a bridge between clusters, maintains important links	connects a distant cluster with the rest of the network	

Eigenvector Centrality

The importance of a node depends on the importance of its neighbours

- considers global network topology
- recurrent definition
- multiple variants, e.g., PageRank

What is an Eigenvector

- $\mathbf{Au} = \lambda\mathbf{u}$
- A is matrix, u is vector, λ is number
- how does it indicate centrality?

Eigenvector Centrality: Induction

Let's begin with

- $C_{eig}(i) \propto \sum_{i \neq j} A_{ij} C_{eig}(j)$
- as initial value of $C_{eig}(0)$, we will use e.g. the degree

Iteration for $x_i = C_{eig}(i)$

- $x_i(t+1) = \sum_{j \neq i} A_{ij} x_j(t)$
- which, in essence, is multiplication of matrix by a vector
- $\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t)$, therefore $\mathbf{x}(t) = \mathbf{A}^t \mathbf{x}(0)$
- through exponentiation, we obtain a method with a dominant eigenvector as its solution

Eigenvector Example[5]

retweet network during presidential election debates

- nodes: accounts, links: @ user references and # topics
- how to identify important nodes and what is the structure of communication?

Important nodes

- degree is not sufficient: it creates an advantage for news agency entities
- C_{eig} is able to correctly identify debate participants

C_{eig} can be used for the analysis of a network where we do not know the important nodes in advance.

[5] Shamma et al. (2009)

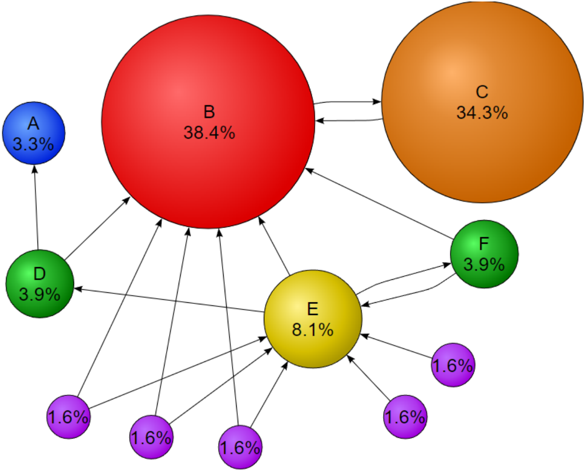
Eigenvector Centrality: PageRank

PageRank as a specific variant of C_{eig}

- an algorithm used by Google Search to rank web pages in their search engine results
- named after both the term "web page" and co-founder Larry Page
- a way of measuring the importance of website pages

- based on random walks through the network
- suitable for directed graphs (teleportation)
- (C_{eig} fails on nodes outside strongly connected components)
- A_{ij} modified: represents a probability of transition between nodes (sum over columns equals 1)

PageRank



PageRank: Example[6]

Citation network

- Physical Review journals
- nodes: articles, links: citations

Importance of an article

- usually determined by the number of its citations (degree)
- degree undervalues key works which allowed ground-breaking articles – PageRank does not do this
- yet PageRank and node degree positively correlate
- outliers: hidden treasures

[6] Chen, Peng, et al. (2007)

MUNI

FACULTY

OF INFORMATICS