Social Networks & Recommendation Systems

IV. Network metrics.

Grzegorz Siudem

Warsaw University of Technology

MSc program in Data Science has been developed
as a part of task 10 of the project
„NERW PW. Science - Education - Development - Cooperation”
co-funded by European Union from European Social Fund.

Before classes

Remind yourself

How we measure distance in graphs?

$$d(i,j) = ?$$

Already known network metrics

- mean vertex degree

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i,$$

- mean length of paths

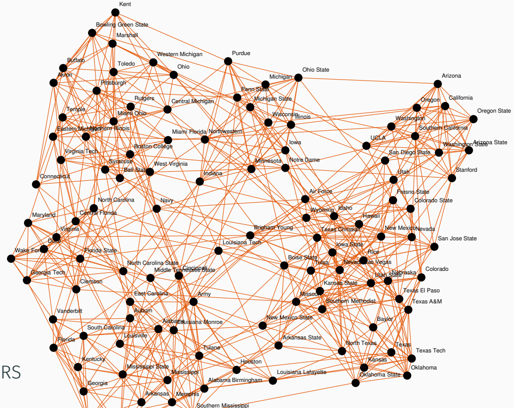
$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j).$$

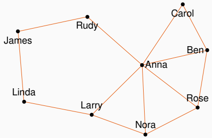
Handshaking lemma

Lecture

What network characteristics can we measure?

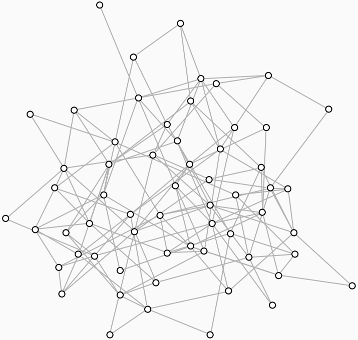
- How big is the network?

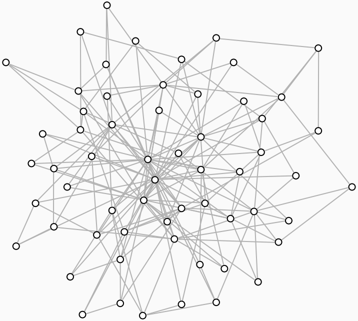




What network characteristics can we measure?

- How big is the network?
- How dense it is?





What network characteristics can we measure?

- How big is the network?
- How dense it is?
- What is the structure of the connections (topology of the network)?

?

Warning!
In complex network science topology has different meaning than in mathematics!

What are the common metrics/measures of the networks?

Natural/naive metrics:

- number of vertices N (size),

What are the common metrics/measures of the networks?

Natural/naive metrics:

- number of vertices N (size),
- number of edges E (size, density),

What are the common metrics/measures of the networks?

Natural/naive metrics:

- number of vertices N (size),
- number of edges E (size, density),
- why $\langle k \rangle = 2E/N$? (density)

What are the common metrics/measures of the networks?

Natural/naive metrics:

- number of vertices N (size),
- number of edges E (size, density),
- why $\langle k \rangle = 2E/N$? (density)
- degree of the biggest hub (celebrities?),

What are the common metrics/measures of the networks?

Natural/naive metrics:

- number of vertices N (size),
- number of edges E (size, density),
- why $\langle k \rangle = 2E/N$? (density)
- degree of the biggest hub (celebrities?),
- another?

What are the common metrics/measures of the networks?

Natural/naive metrics:

- number of vertices N (size),
- number of edges E (size, density),
- why $\langle k \rangle = 2E/N$? (density)
- degree of the biggest hub (celebrities?),
- another?

More formal approach – distributions

In the previous classes, we already talked about

- degree distribution $\mathcal{P}(k)$,

More formal approach – distributions

In the previous classes, we already talked about

- degree distribution $\mathcal{P}(k)$,
- particularly power law distributions with parameter α

$$\mathcal{P}(k) \propto k^{-\alpha},$$

- How estimate α?

More formal approach – distributions

In the previous classes, we already talked about

- degree distribution $\mathcal{P}(k)$,
- particularly power law distributions with parameter α

$$\mathcal{P}(k) \propto k^{-\alpha},$$

- How estimate α?

Who did the homework?

- M.E.J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics, **46**, 323-351 (2005).

More formal approach – distributions

In the previous classes, we already talked about

- degree distribution $\mathcal{P}(k)$,
- particularly power law distributions with parameter α

$$\mathcal{P}(k) \propto k^{-\alpha},$$

- How estimate α?

Who did the homework?

- M.E.J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics, **46**, 323-351 (2005).

More on this topic in the project part.

Measures of node correlation

We already know:

Two-nodes correlations $\mathcal{R}(k_i, k_j)$ i.e. probability that randomly chosen edge connects vertices with degrees k_i and k_j

$$\mathcal{R}(k_i, k_j) = \frac{P(k_i, k_j)}{P_u(k_i, k_j)},$$

and P_u correspondes to uncorrelated network with the same distribution.

Unfortunately it is not very practical tool.

Measures of node correlation

We already know:

Two-nodes correlations $\mathcal{R}(k_i, k_j)$ i.e. probability that randomly chosen edge connects vertices with degrees k_i and k_j

$$\mathcal{R}(k_i, k_j) = \frac{P(k_i, k_j)}{P_u(k_i, k_j)},$$

and P_u correspondes to uncorrelated network with the same distribution.

Unfortunately it is not very practical tool.

We also talked about:

Conditional probability

$$\mathcal{P}(k_i|k_j) = \frac{\mathcal{P}(k_i, k_j)}{k_j \mathcal{P}(k_j)/\langle k \rangle}$$

Is this can be *well* estimated? Unfortunately not...

How to lower the correlation?

Random switch:

It preserves vertices degrees.

Why are we doing this?

- to get rid of unwanted correlations,
- to determine their significance for a given network,
- to obtain a reference model with the same distribution.

Measures of node correlation – continuation

Let's introduce:

Average degree of the nearest node (for node of degree k_i)

$$\langle k\rangle_{nn}(k_i) = \frac{1}{k_i}\sum_{j=1}^{N} a_{ij}k_j = \sum_{\ell} \ell\mathcal{P}(\ell|k_i).$$

We ask about the relationship between $\langle k\rangle_{nn}$ and the degree k_i.

Measures of node correlation – continuation

Let's introduce:

Average degree of the nearest node (for node of degree k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

We ask about the relationship between $\langle k \rangle_{nn}$ and the degree k_i.

What does this measure measure?

- if $\langle k \rangle_{nn}(k_i)$ is an increasing function then network is assortative.

Measures of node correlation – continuation

Let's introduce:

Average degree of the nearest node (for node of degree k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

We ask about the relationship between $\langle k \rangle_{nn}$ and the degree k_i.

What does this measure measure?

- if $\langle k \rangle_{nn}(k_i)$ is an increasing function then network is assortative.
- if $\langle k \rangle_{nn}(k_i)$ is a decreasing function then network is dissassortative.

Measures of node correlation – continuation

Let's introduce:

Average degree of the nearest node (for node of degree k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

We ask about the relationship between $\langle k \rangle_{nn}$ and the degree k_i.

What does this measure measure?

- if $\langle k \rangle_{nn}(k_i)$ is an increasing function then network is assortative.
- if $\langle k \rangle_{nn}(k_i)$ is a decreasing function then network is dissassortative.
- if function is constant the network is uncorrelated.

Measures of node correlation – continuation

Let's introduce:

Average degree of the nearest node (for node of degree k_i)

$$\langle k \rangle_{nn}(k_i) = \frac{1}{k_i}\sum_{j=1}^{N} a_{ij}k_j = \sum_{\ell} \ell \mathcal{P}(\ell | k_i).$$

We ask about the relationship between $\langle k \rangle_{nn}$ and the degree k_i.

What does this measure measure?

- if $\langle k \rangle_{nn}(k_i)$ is an increasing function then network is assortative.
- if $\langle k \rangle_{nn}(k_i)$ is a decreasing function then network is dissassortative.
- if function is constant the network is uncorrelated.
- what in the case of non-monotonic behavior?

Measures of node correlation – continuation

In practice, all the measures learned are too complex...

So it remains for us to calculate the correlation coefficient

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$$

with the following notation

- e_{jk} – joined probability distribution of the others vertices.
- degree distribution for other vertices $q_k = \sum_j e_{jk}$, but from the other hand $q_k = \frac{(k+1)p_{k+1}}{\sum_{j\geq 1} jp_j}$

the above leads to

$$r = \frac{\frac{1}{M}\sum_i k_i j_i - \left[\frac{1}{2M}\sum_i (j_i + k_i)\right]^2}{\frac{1}{2M}\sum_i (j_i^2 + k_i^2) - \left[\frac{1}{2M}\sum_i (j_i + k_i)\right]^2},$$

where $i = 1, 2, \ldots, M$ are indices of edges, and j_i, k_i are degrees of the vertices attached to i.

Clustering coefficient

Homophily phenomenon

Source: gazeta.pl

Heider equilibrium

P-O-X Heider model in a nutshell:

Heider equilibrium

P-O-X Heider model in a nutshell:

- Friend of my friend is my friend.

Heider equilibrium

P-O-X Heider model in a nutshell:

- Friend of my friend is my friend.
- Friend of my enemy is my enemy.

Heider equilibrium

P-O-X Heider model in a nutshell:

- Friend of my friend is my friend.
- Friend of my enemy is my enemy.
- Enemy of my friend is my enemy.

Heider equilibrium

P-O-X Heider model in a nutshell:

- Friend of my friend is my friend.
- Friend of my enemy is my enemy.
- Enemy of my friend is my enemy.
- Enemy of my enemy is my friend.

Heider equilibrium

P-O-X Heider model in a nutshell:

- Friend of my friend is my friend.
- Friend of my enemy is my enemy.
- Enemy of my friend is my enemy.
- Enemy of my enemy is my friend.

However, this applies to directed social networks...

Let us simplify our consideration and limit them to undirected networks.

Clustering coefficient

Definition

The (vertex) clustering coefficient is the ratio of the number of E_i existing edges between the neighbors of the vertex to all possible edges between these neighbors

$$C_i = \frac{2E_i}{k_i(k_i - 1)}.$$

Coefficient of the whole network is an average of the coefficient for every vertex

$$C = \langle C_i \rangle.$$

Clustering coefficient

Definition

The (vertex) clustering coefficient is the ratio of the number of E_i existing edges between the neighbors of the vertex to all possible edges between these neighbors

$$C_i = \frac{2E_i}{k_i(k_i - 1)}.$$

Coefficient of the whole network is an average of the coefficient for every vertex

$$C = \langle C_i \rangle.$$

Alternative definition of the clustering coefficient:

$$C_\triangle = \frac{3 \times \text{number of triangles}}{\text{number of paths of length 2}}.$$

We are counting *motifs* in networks

usually comparing *Z-score* with the ansamble of uncorrelated networks

$$Z = \frac{p - \langle p \rangle}{\sigma}.$$

How to measure how small the world in the network is?

Mean distance

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j)$$

Efficiency

$$\mathcal{E} = \frac{1}{N(N-1)} \sum_{i \neq j} [d(i,j)]^{-1}.$$

How to measure how small the world in the network is?

Mean distance

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j)$$

Efficiency

$$\mathcal{E} = \frac{1}{N(N-1)} \sum_{i \neq j} [d(i,j)]^{-1}.$$

Question:

What are the differences between these two metrics? Which one is *better*?

(Vertices) betweenness centrality

Which vertex is the most important in the network?

We are looking for the most important *transfer stations*.

Notation:

- δ_{jk} is the number of shortest paths connecting the nodes j and k,
- $\delta_{jk}^{(i)}$ is the number of shortest paths connecting the nodes j and k through the node i.

Definition

$$B_i = \frac{2}{(N-1)(N-2)} \sum_k \sum_{j>k} \frac{\delta_{jk}^{(i)}}{\delta_{jk}}.$$

(Edges) betweenness centrality

What changes if one ask about the mose important edge?

We are looking for the most important *line*.

Notation:

- $\delta_{jk}^{(e)}$ is the number of shortest paths connecting the nodes j and k through the edge e.

Definition

$$B_i = \frac{2}{N(N-1)} \sum_k \sum_{j>k} \frac{\delta_{jk}^{(e)}}{\delta_{jk}}.$$

Do we need more metrics?

It is one of the goals in complex network science:

- the whole network is too complex so we need a simplification,

Do we need more metrics?

It is one of the goals in complex network science:

- the whole network is too complex so we need a simplification,
- different people are interested in different networks features,

Do we need more metrics?

It is one of the goals in complex network science:

- the whole network is too complex so we need a simplification,
- different people are interested in different networks features,
- often certain specific measures are needed to describe certain particular types of networks...

Specific metrics

- Hirsch index – in citation networks,

Specific metrics

- Hirsch index – in citation networks,
- Erdős number – in citation networks,

Specific metrics

- Hirsch index – in citation networks,
- Erdős number – in citation networks,
- Bacon number – in actor networks,

Specific metrics

- Hirsch index – in citation networks,
- Erdős number – in citation networks,
- Bacon number – in actor networks,
- PageRank – in the www network,

Specific metrics

- Hirsch index – in citation networks,
- Erdős number – in citation networks,
- Bacon number – in actor networks,
- PageRank – in the www network,
- epidemic threshold – in epidemiology,

Specific metrics

- Hirsch index – in citation networks,
- Erdős number – in citation networks,
- Bacon number – in actor networks,
- PageRank – in the www network,
- epidemic threshold – in epidemiology,
- immunity to attacks/failures – in engineering,

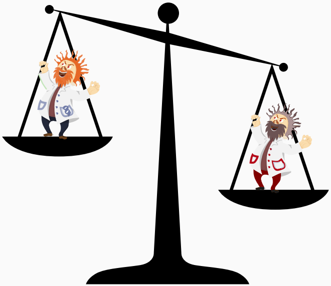
Specific metrics

- Hirsch index – in citation networks,
- Erdős number – in citation networks,
- Bacon number – in actor networks,
- PageRank – in the www network,
- epidemic threshold – in epidemiology,
- immunity to attacks/failures – in engineering,
- community detection,

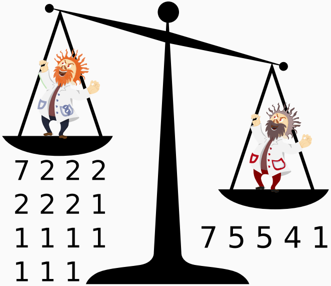
Specific metrics

- Hirsch index – in citation networks,
- Erdős number – in citation networks,
- Bacon number – in actor networks,
- PageRank – in the www network,
- epidemic threshold – in epidemiology,
- immunity to attacks/failures – in engineering,
- community detection,
- and many others...

Hirsch index



How to measure scientific success?



Let us count number of citations (vertices degrees).

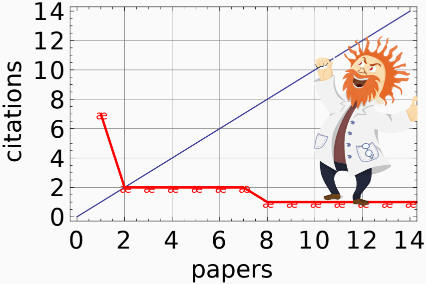
Hirsch index

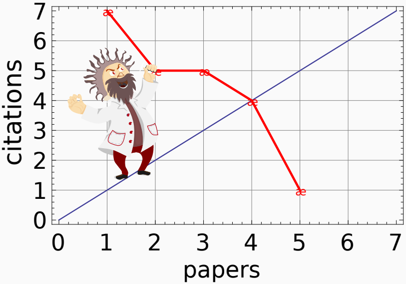
J.E. Hirsch, PNAS **102**, (2005).

$$h\text{-index} = \max\left\{h = 1, \ldots, n:\ X_{(n-h+1)} \geq h\right\},$$

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}.$$

Hirsch index





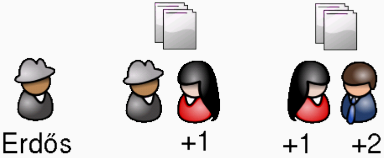
Erdős number

Source: wikipedia

Paul Erdős 1913-1996

- hungarian mathematician,
- in the next class we will learn about Erdős-Rényi graphs.

Erdős number

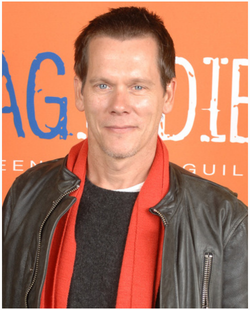


Source: wikipedia

Definition

- Paul Erdős has Erdős number equal to 0.
- Erdős number of every scientists is equal to minimum of the Erdős numbers of his/her coauthors $+1$.

Bacon number



Source: wikipedia

Kevin Bacon ur. 1958

- american actor, director and movie producer.

Bacon number

Equivalent of the Erős number in actor network.

Examples:

- Elvis Preasley: 2,
- Ronald Reagan: 2,
- Andrzej Grabowski: 3,
- Andrzej Lepper: 3,
- Zdzisław Maklakiewicz: 3,
- Jan Himilsbach: 3,

Erdős-Bacon number

The sum of Erdősa and Bacon numbers:

- Steven Strogatz $E = 3$ $B = 1 \Rightarrow EB = 4$,
- Richard Feynman $E = 3$ $B = 3 \Rightarrow EB = 6$,
- Stephen Hawking $E = 4$ $B = 2 \Rightarrow EB = 6$,
- Natalie Portman $E = 5$ $B = 2 \Rightarrow EB = 7$,
- Colin Firth $E = 6$ $B = 1 \Rightarrow EB = 7$,
- Kristen Stewart $E = 5$ $B = 2 \Rightarrow EB = 7$,
- Mayim Bialik $E = 5$ $B = 2 \Rightarrow EB = 7$.

The sum of Erdősa and Bacon numbers:

- Steven Strogatz $E = 3$ $B = 1 \Rightarrow EB = 4$,
- Richard Feynman $E = 3$ $B = 3 \Rightarrow EB = 6$,
- Stephen Hawking $E = 4$ $B = 2 \Rightarrow EB = 6$,
- Natalie Portman $E = 5$ $B = 2 \Rightarrow EB = 7$,
- Colin Firth $E = 6$ $B = 1 \Rightarrow EB = 7$,
- Kristen Stewart $E = 5$ $B = 2 \Rightarrow EB = 7$,
- Mayim Bialik $E = 5$ $B = 2 \Rightarrow EB = 7$.

Fun fact

Read about Erdős-Bacon-Black Sabath number...

Other metrics

PageRank

- method of selecting the most important websites,
- we will deal with it during the 11th class.

Other metrics

PageRank

- method of selecting the most important websites,
- we will deal with it during the 11th class.

Epidemic threshold

- the minimum number of infected people in the social network that results in an epidemic,
- we will deal with it during the 12th class.

Other metrics

PageRank

- method of selecting the most important websites,
- we will deal with it during the 11th class.

Epidemic threshold

- the minimum number of infected people in the social network that results in an epidemic,
- we will deal with it during the 12th class.

Resistance to accidental failures and intentional attacks

- we will deal with it during the 7th class.

Other metrics

PageRank

- method of selecting the most important websites,
- we will deal with it during the 11th class.

Epidemic threshold

- the minimum number of infected people in the social network that results in an epidemic,
- we will deal with it during the 12th class.

Resistance to accidental failures and intentional attacks

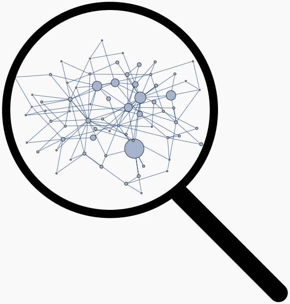
- we will deal with it during the 7th class.

Community detection

- we will deal with it during the 8th class.

Summary

Summary



European Funds
Knowledge Education Development

Warsaw University of Technology

European Union
European Social Fund

MSc program in Data Science has been developed
as a part of task 10 of the project
„NERW PW. Science - Education - Development - Cooperation"
co-funded by European Union from European Social Fund.

Thank you for your attention!