Centralita vrcholov v sociálnych sieťach

Beáta Stehlíková

2-EFM-155 Analýza sociálnych sietí

Fakulta matematiky, fyziky a informatiky, UK v Bratislave, 2019

Tri základné centrality: stupňa, blízkosti a medzipolohy

Pre neorientované nevážené siete

- **centralita vrchola**: počet hrán, ktoré z vrchola vychádzajú
- **centralita blízkosti**: nepriamo úmerná $\sum_{j \neq i} d(i,j)$, konkrétne pre `normalized = FALSE` (default), resp. `normalized = TRUE`

$$\frac{1}{\sum_{j \neq i} d(i,j)}, \text{ resp. } \frac{n-1}{\sum_{j \neq i} d(i,j)}$$

- **centralita medzipolohy**: úmerná $\sum_{i,j \neq k} P_k(i,j)/P(i,j)$, konkrétne pre `normalized = FALSE` (default), resp. `normalized = TRUE`

$$\sum_{i,j \neq k} P_k(i,j)/P(i,j), \text{ resp. } \frac{2}{(n-1)(n-2)} \sum_{i,j \neq k} P_k(i,j)/P(i,j)$$

kde n je počet vrcholov, $d(i,j)$ je dĺžka najkratšej cesty z i do j, $P(i,j)$ je počet takých najkratších ciest, $P_k(i,j)$ je počet takých, ktoré obsahujú vrchol k

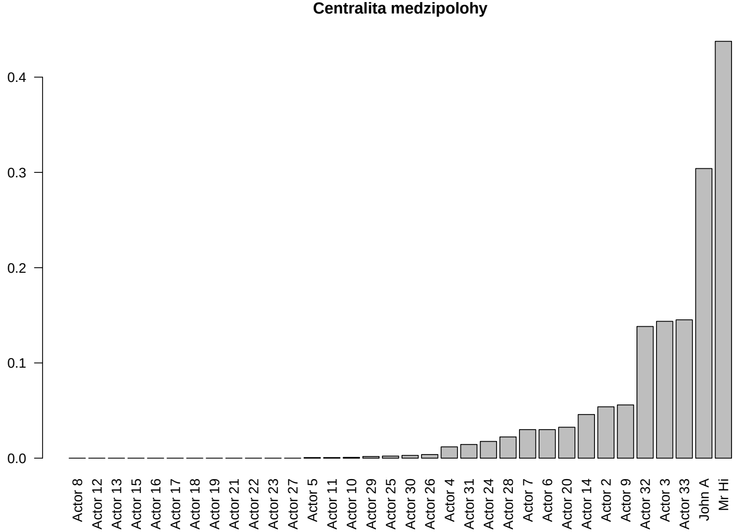
Príklad 1: Zacharyho karate klub

- Zacharyho karate klub
- Hranám odoberieme váhy
  - teda existencia hrany bude znamenať, že vrcholy (t. .j. príslušní ľudia) majú nejaký kontakt
  - ale počet rôznych vzťahov medzi nimi neberieme do úvahy

```
library(igraphdata)
data(karate)

# MOZNOST 1
g1 <- set_edge_attr(karate, "weight", value = 1)
is_weighted(g1)  # graf je vazeny, ale vahy su rovnake

# MOZNOST 2
g2 <- delete_edge_attr(karate, "weight")
is_weighted(g2)  # graf uz nie je vazeny
```

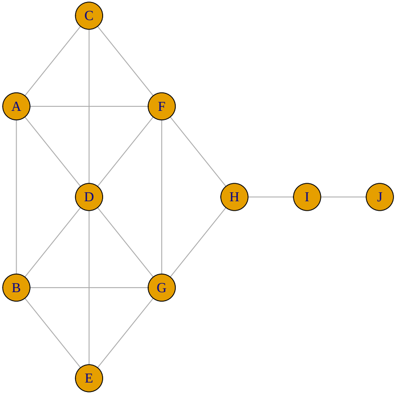
Príklad 1: Zacharyho karate klub

- Zoraďte vrcholy podľa centrality a graficky znázornite, napr.:



Príklad 2: Sieť "kite" a najcentrálnejší vrchol

Nájdite najcentrálnejší vrchol podľa každého z týchto troch kritérií pre sieť `kite` knižnice `igraphdata`



Príklad 3: Hierarchická štruktúra mafie

Calderoni, F. (2014). Identifying mafia bosses from meeting attendance. In Networks and network analysis for defence and security (pp. 27-48). Springer, Cham.

https://link.springer.com/chapter/10.1007/978-3-319-04147-6_2

Z abstraktu:

- The analysis relies on data from a large-scale investigation on the presence of the 'Ndrangheta, a mafia from the Southern Italian region of Calabria.
- Operation Infinito identified several mafia families and tracked a number of mafia meetings.
- The results show that betweenness centrality is the most significant predictor of leadership in the mafia.
- A logistic regression model, using network measures as predictors, is able successfully to predict the position (boss or other) of 92 % of the individuals in the network.

Príklad 3: Hierarchická štruktúra mafie

Z článku:

- Participants and meetings were entered into a two-mode, affiliation matrix (left graph in Fig. 1).
- From this, a one-mode, valued and undirected adjacency matrix was computed (right graph in Fig. 1), recording each individual's co-participation in meetings with any other individual in the network.
- The affiliation matrix enabled the calculation of a number of social network measures (all based on a binary matrix, except for valued degree centrality).

Príklad 3: Hierarchická štruktúra mafie

Siete:

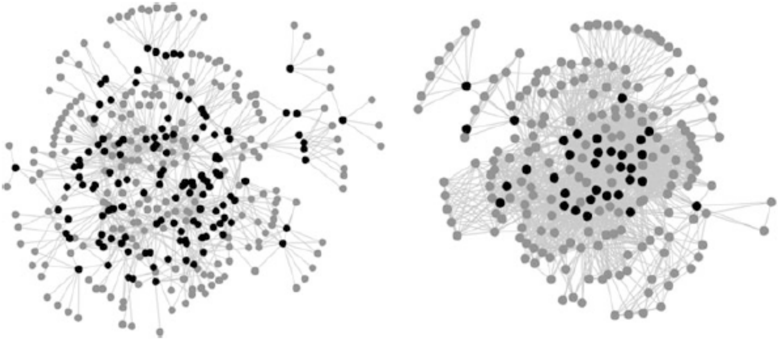


Fig. 1 The infinito 1-mode and 2-mode networks. (In the *left* graph, *black* nodes are meetings, *grey* nodes are participants. In the *right* graph, *black* nodes are the bosses, *grey* nodes are other individuals)

Príklad 3: Hierarchická štruktúra mafie

Priemerné centrality:

Table 3 Mean number of meetings and centrality measures by hierarchy

	Bosses	Others	Ratio bosses/others
Meetings attended	11.7	2.6	4.5
Degree (normalised)	20.0	6.7	3.0
Valued degree	90.7	20.3	4.5
Betweenness (normalised)	3.5	0.2	14.7
Closeness (normalised)	48.1	38.3	1.3
Eigenvector (normalised)	16.2	4.8	3.4
Clustering coefficient	0.5	0.9	0.6
Number of pairs	1081.5	151.6	7.1

The differences of the means are statistically significant at $p < 0.001$ (Mann–Whitney U test)

Príklad 3: Hierarchická štruktúra mafie

- Podobné dáta:
  - Na stránke https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/ndranghetamafia2
  - Popis dát: *Data is on attendance of suspected members of the Ndrangheta criminal organization at summits (meetings whose purpose is to make important decisions and/or affiliations, but also to solve internal problems and to establish roles and powers) taking place between 2007 and 2009.*
  - Štruktúra dát: *2-mode matrix 156 x 47 persons by events (summits), undirected binary ties.*
- Ktoré vrcholy považujete za podozrivé?

Ako skombinovať rôzne miery centrality

Námet na projekt - použiť jeden z prístupov:

Mesgari, I., Kermani, M. A. M. A., Hanneman, R., & Aliahmadi, A. (2015). Identifying key nodes in social networks using multi-criteria decision-making tools. In Mathematical Technology of Networks (pp. 137-150). Springer, Cham.

https://link.springer.com/chapter/10.1007/978-3-319-16619-3_10

Vysvetliť postup, zopakovať výpočet z článku (napr. starý známy karate klub), použiť na nové originálne a zaujímavé dáta

Neorientované vážené siete

Centralita stupňa - rovnako ako pri nevážených - je to počet hrán vychádzajúcich z daného vrchola

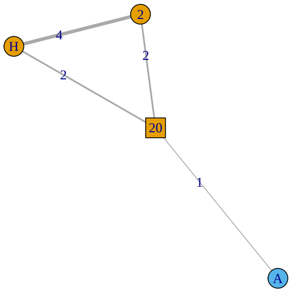
Môžeme ale chcieť zobrať do úvahy aj váhy:

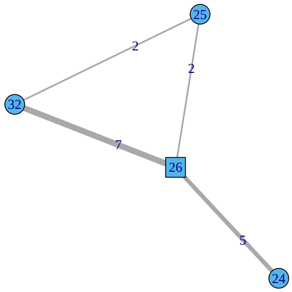
```
# povodne Zacharyho data s vahami
g <- make_ego_graph(karate,
                    nodes = c("Actor 20", "Actor 26"))
    # ALEBO NAPR.:  nodes = c(1, 5, 12)

g20 <- g[[1]]  # g je typu 'list'
plot(g20, edge.label=E(g20)$weight)
```

`make_ego_graph` zoberie zadaný graf a zoberie z neho podgraf obsahujúci zadaný vrchol a vrcholy s ním susediace, spolu so všetkými hranami medzi týmito vrcholmi, napr. `make_ego_network(karate, nodes = "Mr Hi")` alebo `make_ego_network(karate, nodes = 5)`

Neorientované vážené siete

- Vrcholy majú rovnaký počet hrán, ale s rôznymi váhami.
- Porovnajte `degree(karate)` a `strength(karate)`.





Neorientované vážené siete

Centralita blízkosti a centralita medzipolohy:

- treba špecifikovať "vzdialenosti" vrcholov, ktoré sa budú používať pri výpočte dĺžky ciest medzi vrcholmi
- vo funkciách `closeness` a `betweenness` je to parameter `weights`, ak majú hrany parameter `weight`, defaultne sa zoberú tieto

Cvičenie:

- V našom príklade karate klubu nám toto defaultné nastavenie nevyhovuje, vysoká hodnota `weights` hovorí o veľkom počte kontaktov $\rightarrow$ vzdialenosť by mala byť malá.
- Zvoľte vhodnú transformáciu (nie je daná jednoznačne) a vypočítajte centrality

Orientované siete

Centralita stupňa - rozlišuje sa

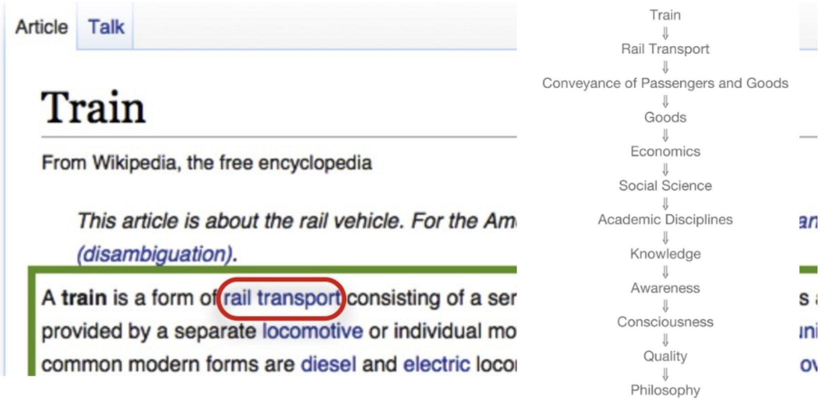
- *in-degree*: počet hrán, ktoré z vrchola vychádzajú
- *out-degree*: počet hrán, ktoré do vrchola prichádzajú
- vo funkcii `degree` parameter `mode` s hodnotou `"in"`, resp. `"out"` (ak chceme všetky, tak `"all"` alebo `"total"`)

Príklad: *Ibrahim, M., Danforth, C. M., & Dodds, P. S. (2017). Connecting every bit of knowledge: The structure of Wikipedia's First Link Network. Journal of Computational Science, 19, 21-30.*

https://www.sciencedirect.com/science/article/pii/S1877750316304471

Z abstraktu: *By following the first link in each article, we algorithmically construct a directed network of all 4.7 million articles: Wikipedia's First Link Network.*

Orientované siete



Orientované siete

- *Out-degree* každého vrchola je 1
- *In-degree*:

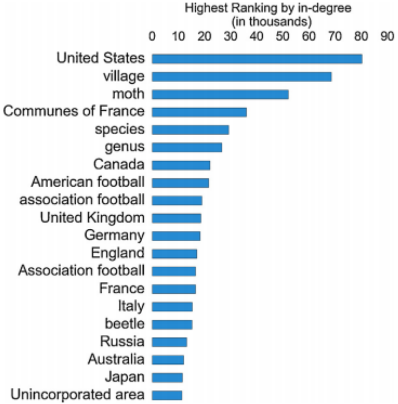


Fig. 5. Highest Ranking Articles by in-degree. We rank each article by the number of direct first links to the article (in-degree). The highest-ranking articles tend to represent geographical and biological abstractions. A full online appendix of the results and data is available here.

Orientované siete

Centralita blízkosti - vo funkcii `closeness` treba určiť (parameter `mode` s hodnotou `"in"`, resp. `"out"`), či nás zaujímajú najkratšie cesty

- z ostatných vrcholov do daného vrchola `"in"`
- z daného vrchola do ostatných (`"out"`)

Centralita medzipolohy - nemení sa nič, nakoľko sa uvažujú všetky najkratšie cesty v sieti (medzi každými dvoma vrcholmi)

Siete musia byť silno súvislé[1], aby sa tieto výpočty dal robiť.

[1] Z každého vrcholu A sa vieme v smere šípok dostať do ľubovoľného vrchola Ba

Príklad 4: Odkazy na webstránkach

Dáta:

- https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/domesticterroristweblinks
- Popis: 1-mode matrix 32 x 32 website by website. Directed binary ties are based on analysis of hyperlinks between sites.

Zadanie:

Zoberte najväčší súvislý komponent, na ktorý sa sieť rozpadne a analyzujte centralitu webstránok (vynechajte pritom odkaz, ktorým stránka odkazuje sama na seba).

Poznámky:

- Vynechanie odkazovania na seba: vo funkcii `graph_from_adjacency_matrix` zvolíme navyše `diag = FALSE` (diagonála matice susednosti sa zoberie nulová)
- Pri hľadaní komponentov musíme zadať parameter `mode = "strong"` (zaujíma nás silná súvislosť)

Centralita vlastného vektora

Základná myšlienka

- Centralita vrchola by mala závisieť od centralít vrcholov, ktoré s ním susedia

Pre neorientované siete (môžu byť vážené)

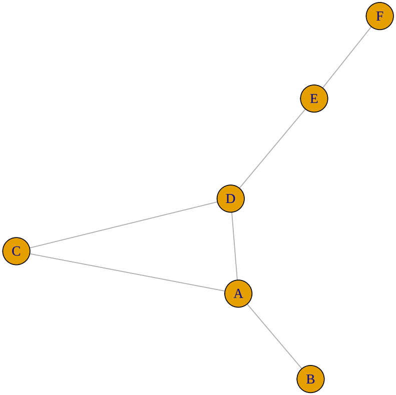
- Ak A je matica susednosti (prípadne aj s váhami) a v je vektor centralít, tak by malo pre najaké kladné číslo c platiť

$$v_i = c \sum_j A_{ij} v_j$$

- Maticovo: $v = \lambda A v$, kde $\lambda = 1/c$
- Teda v (vektor centralít) je vlastným vektorom $A \rightarrow$ preto názov *centralita vlastného vektora*, resp. *eigenvalue centrality*

Výpočet v R-ku

```
g <- graph_from_literal(A-B, A-C, A-D, C-D, D-E,
                        E-F)
plot(g)
```



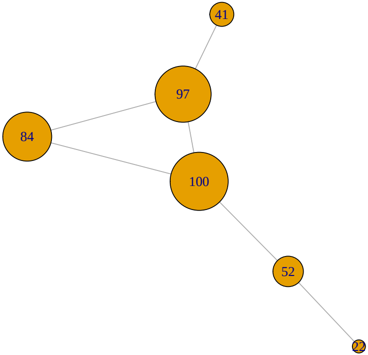
Výpočet v R-ku

```r
centralita <- eigen_centrality(g)$vector
round(centralita, 4) # zaokruhlene, 4 desatinne miesta
```

```
##      A      B      C      D      E      F
## 0.9669 0.4142 0.8426 1.0000 0.5247 0.2248
```

Príklad

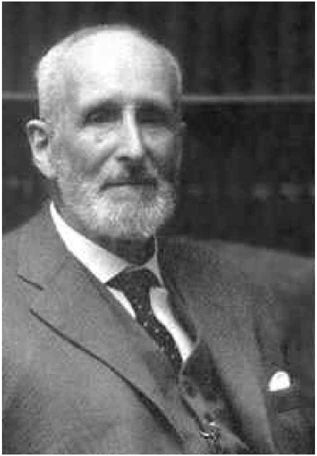
```
plot(g, vertex.label = round(centralita * 100, 0),
        vertex.size = centralita * 35)
```



Perronova-Frobeniova veta

Otázka: *Ktoré vlastné číslo a vlastný vektor zobrať?* Matica A má n vlastných čísel.

Odpoveď súvisí s Perronovou-Frobeniovou vetou z teórie matíc.



Oskar Perron (1880 - 1975), Ferdinand Georg Frobenius (1849 - 1917)

Perronova-Frobeniova veta

Nech A je matica susednosti súvislého[2] grafu. Označme λ_i jej vlastné hodnoty a $\rho = \max |\lambda_i|$. Z Perronovej-Frobeniovej vety vyplýva, že:

- ρ je vlastná hodnota matice A
- existuje (až na násobok) jediný vektor x, pre ktorý $Ax = \rho x$
- existuje (až na násobok) jediný vektor y, pre ktorý $y'A = \rho y'$
- tieto vektory majú zložky s rovnakým znamienkom

[2] Z každého vrcholu A sa vieme dostať do ľubovoľného vrchola B, v prípade orientovaného grafu iba v smere šípok

Príklad

```r
A <- as_adjacency_matrix(g, sparse = FALSE)
A
```

```
##   A B C D E F
## A 0 1 1 1 0 0
## B 1 0 0 0 0 0
## C 1 0 0 1 0 0
## D 1 0 1 0 1 0
## E 0 0 0 1 0 1
## F 0 0 0 0 1 0
```

```r
lambda <- eigen(A)$values

# o funkcii `eigen` a vystupe `values`:
# sorted in decreasing order, according to Mod(values)
# preto:
rho <- lambda[1]
```

Príklad

```r
# o funkcii `eigen` a vystupe `vector`:
# matrix whose columns contain the eigenvectors
# preto:
v <- eigen(A)$vector[,1]
```

Prenásobne vektor `v` konštantou (zostane vlastným vektorom) tak, aby mal rovnaké škálovanie ako centralita.

Porovnajte výsledok s výstupom z `eigen_centrality`

Túto mieru centrality sme už videli v článku o mafii:

Table 3 Mean number of meetings and centrality measures by hierarchy

	Bosses	Others	Ratio bosses/others
Meetings attended	11.7	2.6	4.5
Degree (normalised)	20.0	6.7	3.0
Valued degree	90.7	20.3	4.5
Betweenness (normalised)	3.5	0.2	14.7
Closeness (normalised)	48.1	38.3	1.3
Eigenvector (normalised)	16.2	4.8	3.4
Clustering coefficient	0.5	0.9	0.6
Number of pairs	1081.5	151.6	7.1

The differences of the means are statistically significant at $p < 0.001$ (Mann–Whitney U test)

Orientované siete

Námet na projekt:

Ak by sme chceli postupovať analogicky:

- Ak A je matica susednosti (prípadne aj s váhami) a v je vektor centralít, tak by malo pre najaké kladné číslo c platiť

$$v_i = c \sum_j A_{ji} v_j$$

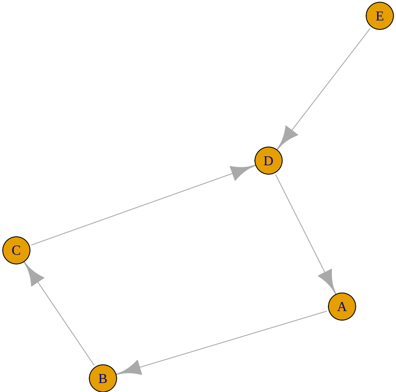
- Maticovo: $v^T = v^T \lambda A$, kde $\lambda = 1/c$
- Teda v (vektor centralít) je ľavým vlastným vektorom matice A
- Perronovu-Frobeniovu vetu môžeme použiť aj teraz
- `eigen_centrality` s parametrom `directed = TRUE`
- Môžeme skontrolovať výpočtom vlastných hodnôt a vektorov transponovanej matice A

Prečo ale v R-ku upozorňujú: *Don't send this routine asymmetric matrices unless you really mean to do so* ?

Orientované siete

Príklad 1

```
g <- graph_from_literal(A-+B, B-+C, C-+D, D+-E, D-+A)
plot(g, edge.arrow.size=1.5)
```



Orientované siete

Výpočet centrality vlastného vektora:

```
centralita <- eigen_centrality(g, directed=TRUE)$vector
round(centralita, 4) # zaokruhlene
```

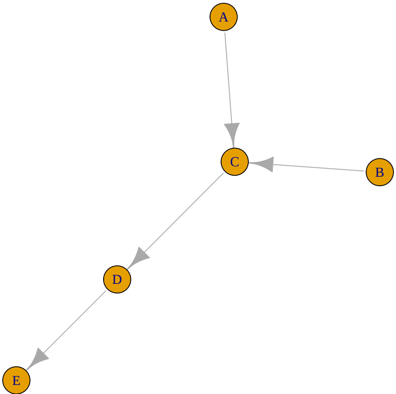
```
## A B C D E
## 1 1 1 1 0
```

Vrchol E nijako nezvyšuje centralitu vrcholu D.

Orientované siete

Príklad 2

```r
g <- graph_from_literal(A-+C, B-+C, C-+D, D-+E)
plot(g, edge.arrow.size=1.5)
```



Orientované siete

Výpočet centrality vlastného vektora:

```
centralita <- eigen_centrality(g, directed=TRUE)$vector
```

```
## Warning in eigen_centrality(g, directed = TRUE): At cent
## is directed and acyclic; eigenvector centralities will b
```

```
centralita
```

```
## A C B D E
## 0 0 0 0 0
```

Všetky centrality sú nulové.

Orientované siete

V projekte:

- iné miery centrality s podobnou myšlienkou: Katzova centralita, Page Rank
- vysvetliť postup, spraviť jednoduchý príklad
- použiť na nové originálne a zaujímavé dáta