introduction to *network analysis* (*ina*)

Lovro Šubelj

University of Ljubljana
Faculty of Computer and Information Science
spring 2019/20

announcements *4th week*

- *homework #1* out *today*
- *homework #1* due in *two weeks*
- *course project* details *next week*

- *grand graph* challenge in *lectures*
- again *Kahoot! quiz* in *lectures*
- bring *your laptops* to *labs*

- *feedback box* from *last week*
- posts to *feedback box* get *candy* =)

challenge *4th week*

grand graph challenge

node *centrality*

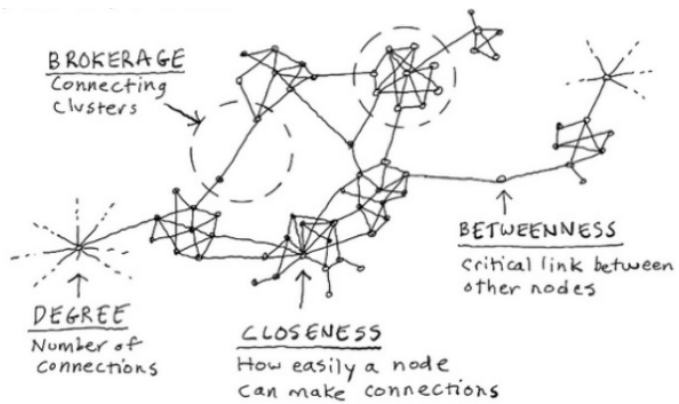
introduction to *network analysis* (*ina*)

Lovro Šubelj
University of Ljubljana
spring 2019/20

centrality *measures*

which *nodes* are most *important*?

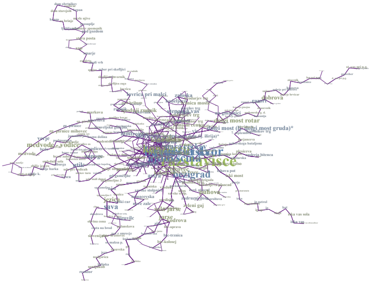
— *node centrality measures* for (*un*)*directed* networks
  - *clustering coefficients* [WS98, SV05, dNMB05]
  - *distance-based* centrality [Fre77, FBW91, New05]
  - *spectral analysis* centrality [Kat53, Bon87, BP98]
  - *fragment-based* centrality [MSOI+02, Pržˇ07, EK15]



— *link analysis algorithms* for *directed* networks

networkology *LPP*

- partial *LPP public bus transport network**
- $n = 416$ bus stops with $\langle k \rangle = 5.62$ connections
- *giant component* 95.4% nodes (6 components)
- *"small-world"* with $\langle C \rangle = 0.09$ and $\langle d \rangle = 14.26$
- *"scale-free"* with $\gamma = 2.62$ for cutoff $k_{min} = 5$



* reduced to largest connected component

centrality *clustering*

important *nodes* are *strongly embedded*

— for *undirected* G *clustering coefficient* C [WS98] of i is
  - t_i is number of *linked neighbors* or *triangles* of i

$$C_i = \frac{2t_i}{k_i(k_i-1)} \qquad C_i = 0 \text{ for } k_i \leq 1$$

— ω*-corrected clustering coefficient* C^ω [SV05] of i is
  - ω_i is *maximum possible* t_i with *respect to* $\{k\}$

$$C_i^\omega = \frac{t_i}{\omega_i} \qquad C_i^\omega = 0 \text{ for } \omega_i = 0$$

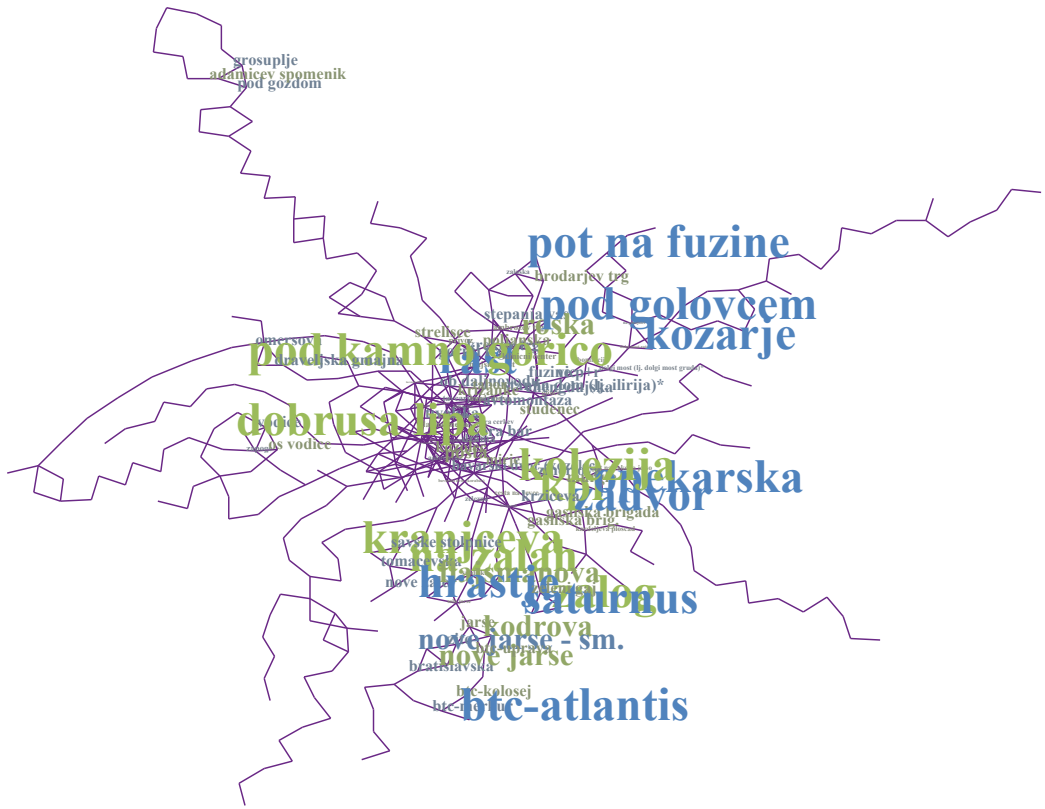
— μ*-corrected clustering coefficient* C^μ [dNMB05] of i is
  - μ is *maximum* number of *triangles* over *links*

$$C_i^\mu = \frac{2t_i}{k_i\mu} \qquad C_i^\mu = 0 \text{ for } k_i = 0$$

networkology *clustering*

— *clustering coefficient* C in partial LPP network[†]

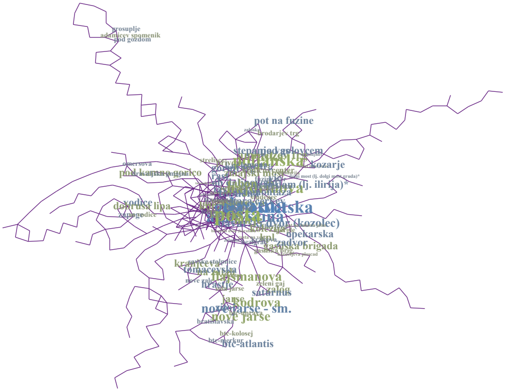
— *highest* $C_i = 1.0$ nodes are *Na Žalah etc.* with $k_i = 2$



[†] reduced to simple undirected graph

networkology *μ-clustering*

- *μ-corrected clustering* C^{μ} in partial LPP network[‡]
- *highest* $C_i^{\mu} = 0.44$ node is *Drama* with $k_i = 10$



‡ reduced to simple undirected graph

centrality *closeness*

important *nodes* are *close to other* nodes

— for (*un*)*directed* G *closeness centrality* ℓ^{-1} [New10] of i is
  - d_{ij} is (*un*)*directed distance* between i and j
  - $d_{ij} = \infty$ for nodes in *different components*

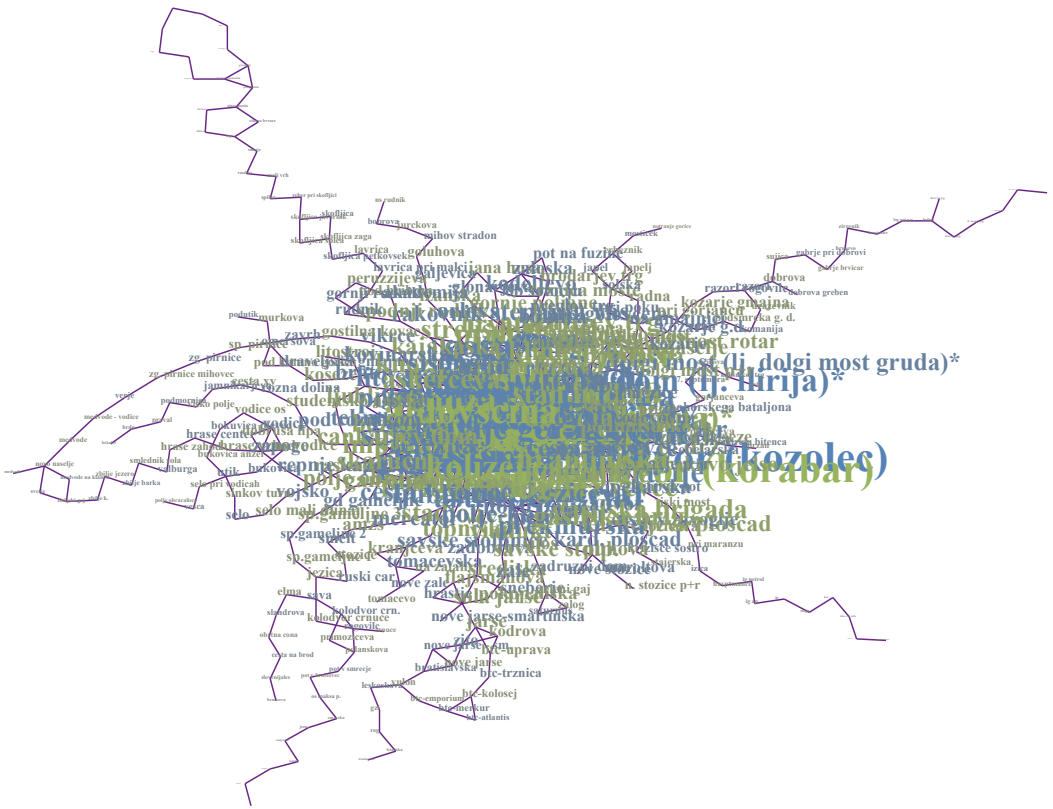
$$\ell_i^{-1} = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}$$

— ℓ^{-1} spans *small range* in *small-world* networks

networkology *closeness*

— *closeness centrality* ℓ^{-1} in partial LPP network[§]

— *highest* $\ell_i^{-1} = 0.208$ node is *Gosposvetska* with $k_i = 14$



[§] reduced to simple undirected graph

centrality *betweenness*

important *nodes* are *bridges for other* nodes

— for (*un*)*directed* G *betweenness centrality* σ [Fre77] of i is
  - g_{st} is number of *geodesic paths between* s and t
  - g_{st}^{i} is number of *such geodesic paths through* i

$$\sigma_i = \frac{1}{n^2} \sum_{st} \frac{g_{st}^{i}}{g_{st}}$$

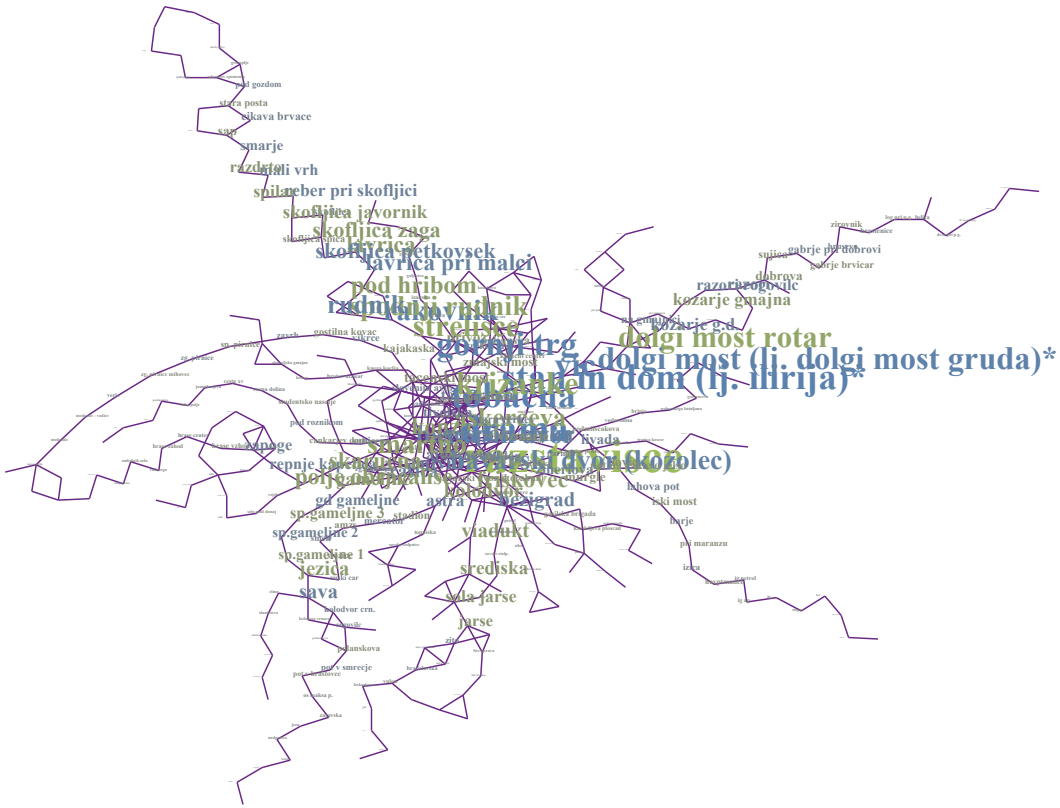
— σ considers *only geodesic paths* [FBW91, New05]

— σ mixes *local centers* with *global bridges* [JMK+16]

networkology *betweenness*

— *betweenness centrality* σ in partial LPP network[¶]

— *highest* $\sigma_i = 0.235$ node is *Razstavišče* with $k_i = 11$



[¶] reduced to simple undirected graph

centrality *degrees*

important *nodes* are *linked by many* nodes

— for *undirected* G *degree centrality* d of i is

$$d_i = \frac{1}{n-1} \sum_{j \neq i} A_{ij} = \frac{k_i}{n-1}$$

— in *directed* G *in-degree centrality* d^{in} of i is

$$d_i^{in} = \frac{1}{n-1} \sum_{j \neq i} A_{ij} = \frac{k_i^{in}}{n-1}$$

— in *directed* G *out-degree centrality* d^{out} of i is

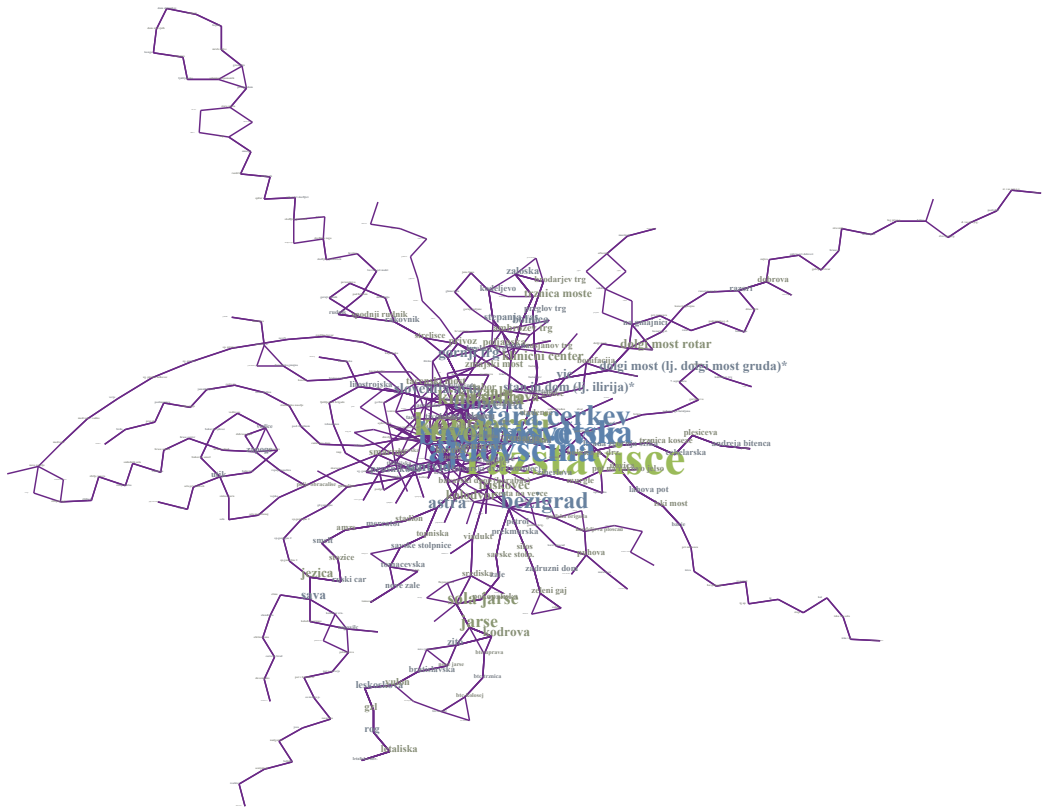
$$d_i^{out} = \frac{1}{n-1} \sum_{j \neq i} A_{ji} = \frac{k_i^{out}}{n-1}$$

networkology *degrees*

— *degree centrality* d in partial LPP network

— *highest* $d_i = 0.099$ node is *Razstavišče* with $k_i = 41$

— *highest* $d_i^{\cdot}$ node is *Razstavišče* with $k_i^{in} = 20$ and $k_i^{out} = 21$



centrality *eigenvector*

important *nodes* are *linked by* *important* nodes

— for (*un*)*directed* G *eigenvector centrality* e [Bon87] of i is
  - v and λ are *eigenvectors* and *eigenvalues* of A
  - e is *proportional* to *leading eigenvector* v_1

$$e(t) = A^t e(0) = A^t \sum_i C_i v_i = \sum_i C_i \lambda_i^t v_i = \lambda_1^t \sum_i C_i \left[\frac{\lambda_i}{\lambda_1}\right]^t v_i \to C_1 \lambda_1^t v_1$$

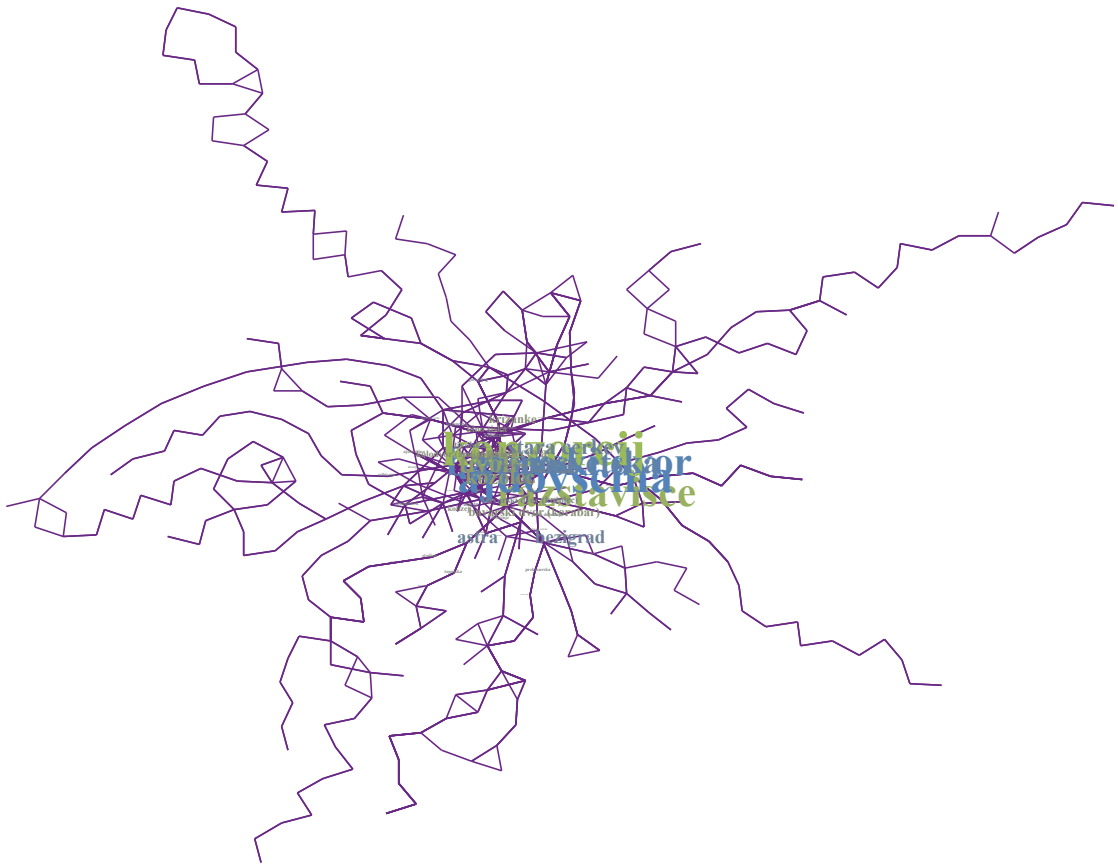
$$e_i = \lambda_1^{-1} \sum_j A_{ij} e_j$$

— in *directed* G $e = 0$ for $k^{in} = 0$ *nodes etc.*

networkology *eigenvector*

— *eigenvector centrality* e in partial LPP network

— *highest* $e_i = 0.082$ node is *Konzorcij* with $k_i = 30$



centrality *Katz*

nodes get small amount of *importance for free*

— for (*un*)*directed* G *Katz centrality* z [Kat53] of i is
  - α and β are some *positive constants*

$$z_i = \alpha \sum_j A_{ij} z_j + \beta_i$$

— for *convenience* $\beta = 1$ whereas $\alpha < \lambda_1^{-1}$
  - λ_1 is *leading eigenvalue* of A

centrality *PageRank*

nodes distribute equal amount of *importance*

— for (*un*)*directed* G *PageRank centrality* p [BP98] of i is
  - α and β are some *positive constants*

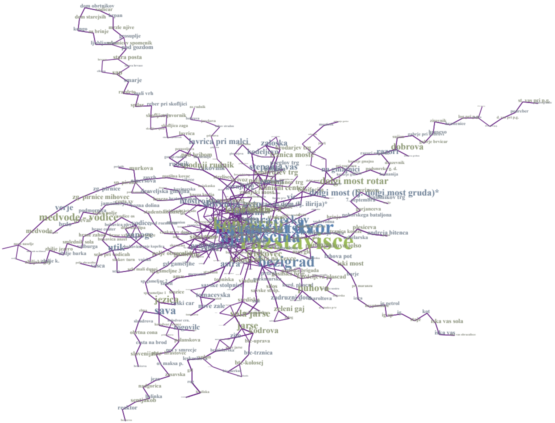
$$p_i = \alpha \sum_j A_{ij} \frac{p_j}{k_j^{out}} + \beta_j$$

— for *convenience* $\beta = \frac{1-\alpha}{n}$ whereas $\alpha = 0.85$

see PageRank algorithm NetLogo demo

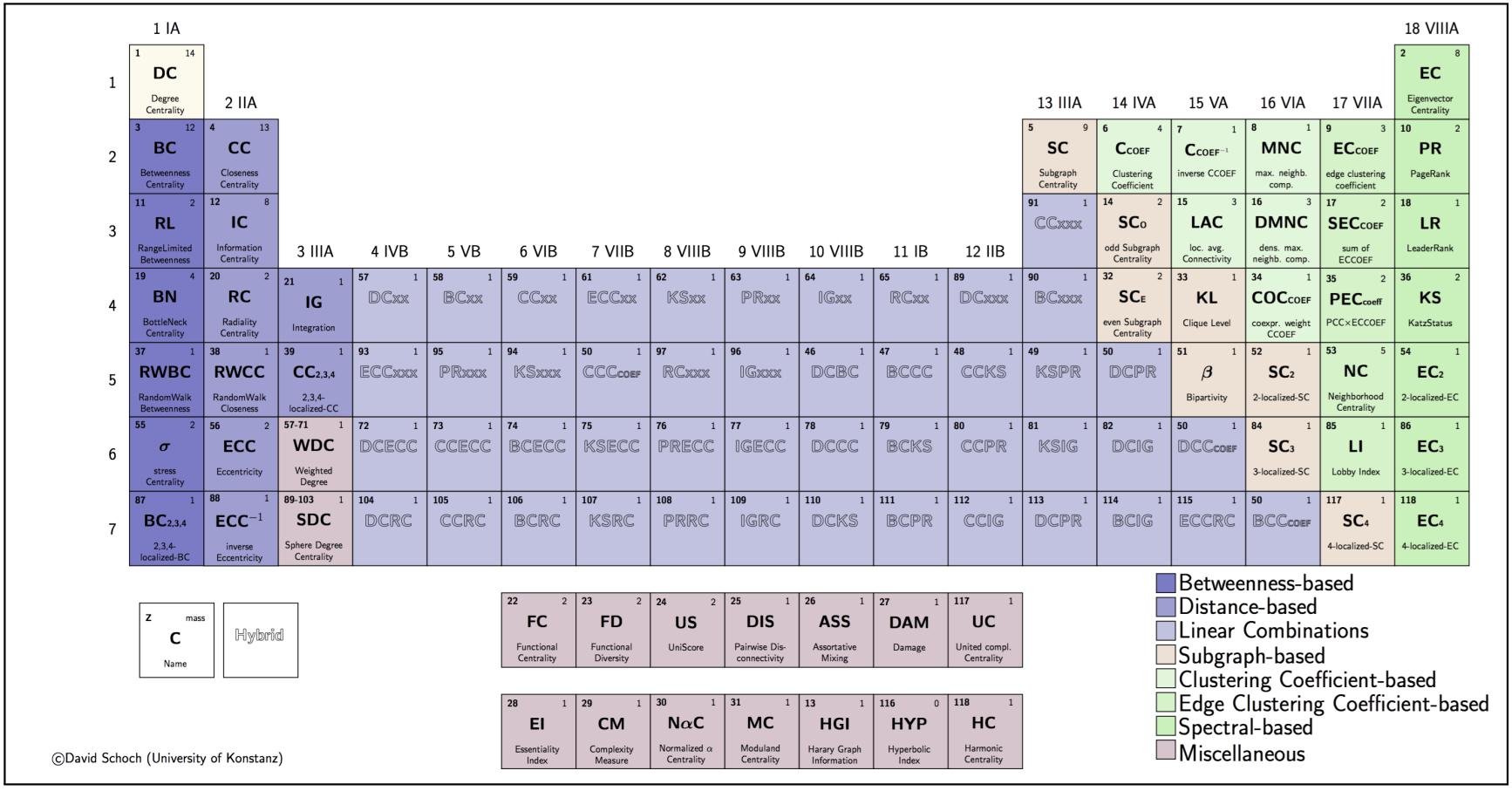
networkology *PageRank*

- *PageRank centrality* p in partial LPP network
- *highest* $p_i = 0.011$ node is *Razstavišče* with $k_i = 41$



centrality *overview*

which *nodes* are most *important*?



centrality *references*

Phillip Bonacich.

Power and centrality: A family of measures.
American Journal of Sociology, 92(5):1170–1182, 1987.

S. Brin and L. Page.

The anatomy of a large-scale hypertextual Web search engine.
Comput. Networks ISDN, 30(1-7):107–117, 1998.

Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj.

Exploratory Social Network Analysis with Pajek.
Cambridge University Press, Cambridge, 2005.

David Easley and Jon Kleinberg.

Networks, Crowds, and Markets: Reasoning About a Highly Connected World.
Cambridge University Press, Cambridge, 2010.

Ernesto Estrada and Philip A. Knight.

A First Course in Network Theory.
Oxford University Press, 2015.

Linton C. Freeman, Stephen P. Borgatti, and Douglas R. White.

Centrality in valued graphs: A measure of betweenness based on network flow.
Soc. Networks, 13(2):141–154, 1991.

L. Freeman.

A set of measures of centrality based on betweenness.
Sociometry, 40(1):35–41, 1977.

centrality *references*

Pablo Jensen, Matteo Morini, Marton Karsai, Tommaso Venturini, Alessandro Vespignani, Mathieu Jacomy, Jean-Philippe Cointet, Pierre Merckle, and Eric Fleury.
Detecting global bridges in networks.
J. Complex Netw., 4(3):319–329, 2016.

Leo Katz.
A new status index derived from sociometric analysis.
Psychometrika, 18(1):39–43, 1953.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon.
Network motifs: Simple building blocks of complex networks.
Science, 298(5594):824–827, 2002.

M. E. J. Newman.
A measure of betweenness centrality based on random walks.
Soc. Networks, 27(1):39–54, 2005.

Mark E. J. Newman.
Networks: An Introduction.
Oxford University Press, Oxford, 2010.

Nataša Pržulj.
Biological network comparison using graphlet degree distribution.
Bioinformatics, 23(2):e177–e183, 2007.

Sara Nadiv Soffer and Alexei Vázquez.
Network clustering coefficient without degree-correlation biases.
Phys. Rev. E, 71(5):057101, 2005.

D. J. Watts and S. H. Strogatz.
Collective dynamics of 'small-world' networks.
Nature, 393(6684):440–442, 1998.

link *analysis*

introduction to *network analysis* (*ina*)

Lovro Šubelj
University of Ljubljana
spring 2019/20

link *analysis*

which *web pages* are most *important*?

— *node centrality measures* for (*un*)*directed* networks

— *link analysis algorithms* primarily for *directed web graphs*

- Google *search ranking PageRank* [BP98, PBMW99]
- hyperlink-induced *topic search HITS* [Kle99]

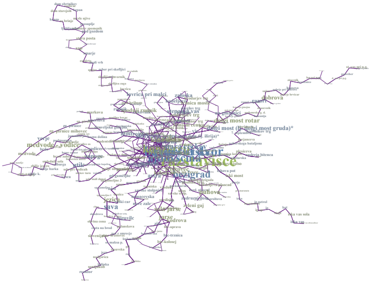
Sergey Brin

Lawrence Page

Jon Kleinberg

networkology *LPP*

- corrected *LPP public bus transport network**
- $n = 408$ bus stops with $\langle k \rangle = 5.73$ connections
- *giant component* 95.3% nodes (6 components)
- *"small-world"* with $\langle C \rangle = 0.10$ and $\langle d \rangle = 14.43$
- *"scale-free"* with $\gamma = 2.60$ for cutoff $k_{min} = 5$



*reduced to largest connected component

analysis *PageRank*

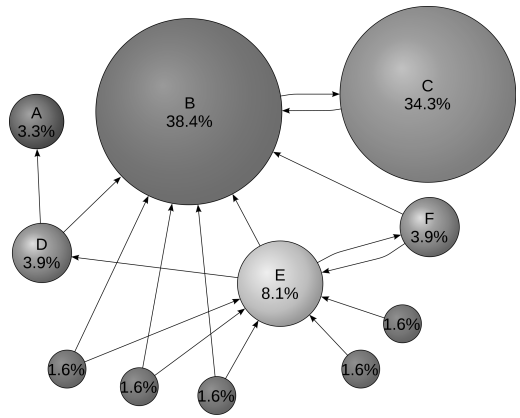
ranking algorithm for *web page importance*

— for *directed* G *PageRank rank* p [BP98] of i is
  – α is *positive constant* traditionally $\alpha = 0.85$

$$p_i = \alpha \sum_j A_{ij} \frac{p_j}{k_j^{out}} + \frac{1-\alpha}{n}$$

— p *oscillates* in *spider traps* and *leaks out* of *dead ends*

— p_i probability *random surfer with teleports* lands on i



— *PageRank ranks p* in corrected LPP network

— *highest p* nodes are *Razstavišče* and *Ajdovščina*

#	**bus stop**	k_i	p_i
1	Razstavišče	43	0.010601
2	Ajdovščina	36	0.007694
3	Bežigrad	23	0.007161
4	Bavarski dvor	30	0.007013
5	Konzorcij	30	0.006884
6	Gosposvetska	30	0.006527
7	Stara cerkev	26	0.005485
8	Sava	12	0.005165
9	Tobačna	22	0.005136
10	Kino Šiška	18	0.004907
11	Medvode	4	0.004853
12	Tivoli	26	0.004838

analysis *WalkRank*

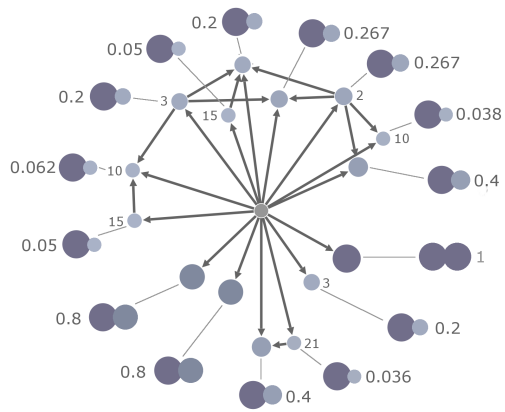
ranking algorithm for *web page similarity*

— for *directed* G *WalkRank rank* w [TFP06] for t of i is
  – α is *positive constant* traditionally $\alpha = 0.85$

$$w_i^t = \alpha \sum_j A_{ij} \frac{w_j^t}{k_j^{out}} + (1 - \alpha)\delta_{it}$$

— w_i^t probability *random surfer with teleport* t lands on i

— *personalized PageRank* and *SimRank* [PBMW99, JW02]



— *WalkRank ranks w* in corrected LPP network

— *highest w* nodes for *Razstavišče* and *Hajdrihova*

#	bus stop	k_i	w_i
1	Razstavišče	43	0.236115
2	Bavarski dvor	30	0.065124
3	Bezigrad	23	0.057260
4	Astra	16	0.047765
5	Ajdovščina	36	0.040099
6	Kozolec	10	0.038384
7	Gosposvetska	30	0.030981
8	Konzorcij	30	0.020278
9	Bavarski dvor	8	0.019262
10	Polje	10	0.014254
11	Stadion	8	0.013294
12	Topniška	8	0.013235

#	bus stop	k_i	w_i
1	Hajdrihova	14	0.201318
2	Tobačna	22	0.091186
3	Ilirija	12	0.051714
4	Stara cerkev	26	0.046825
5	Tabor	10	0.038395
6	Vič	16	0.034478
7	Avtomontaža	6	0.030372
8	Stan in dom	4	0.030296
9	Kino Šiška	18	0.028569
10	Tivoli	26	0.028180
11	Glince	8	0.027528
12	Na klancu	10	0.023836

analysis *HITS*

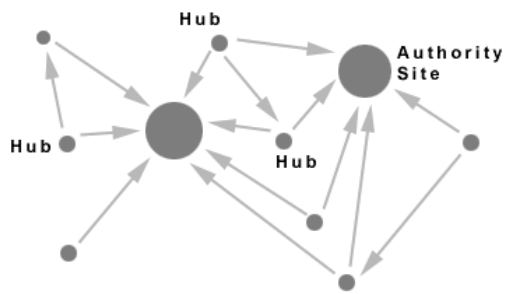
ranking algorithm for *web hubs & authorities*

— for *directed* G *hub & authority ranks* h & a [Kle99] of i
  - h is *eigenvector* of $A^T A$ with *eigenvalue* $(\alpha\beta)^{-1}$
  - a is *eigenvector* of AA^T with *eigenvalue* $(\alpha\beta)^{-1}$
  - α and β are some *positive constants*

$$h_i = \alpha \sum_j A_{ji} a_j \qquad a_i = \beta \sum_j A_{ij} h_j$$

— a measures *content* and h measures *table of content*

— $a = 0$ for $k^{in} = 0$ *nodes* and $h = 0$ for $k^{out} = 0$ *nodes*



— *hub & authority ranks h & a* in corrected LPP network

— *highest h* node is *Ajdovščina* and *highest a* node is *Konzorcij*

#	**bus stop**	k_i	h_i
1	Ajdovščina	36	0.715370
2	Razstavišče	43	0.455771
3	Tivoli	26	0.286178
4	Drama	23	0.256027
5	Gosposvetska	30	0.175142
6	Bavarski dvor	30	0.129155
7	Pošta	9	0.111497
8	Kolodvor	4	0.090644
9	Konzorcij	30	0.083028
10	Tavčarjeva	7	0.069477
11	Kozolec	10	0.068749
12	Stara cerkev	26	0.064760

#	**bus stop**	k_i	a_i
1	Konzorcij	30	0.656745
2	Bavarski dvor	30	0.512119
3	Gosposvetska	30	0.235790
4	Kozolec	10	0.224651
5	Bežigrad	23	0.176839
6	Astra	16	0.172509
7	Stara cerkev	26	0.172482
8	Ajdovščina	36	0.161840
9	Razstavišče	43	0.110391
10	Tivoli	26	0.106024
11	Bavarski dvor	8	0.096486
12	Kolizej	4	0.088636

analysis *references*

S. Brin and L. Page.

The anatomy of a large-scale hypertextual Web search engine.
Comput. Networks ISDN, 30(1-7):107–117, 1998.

David Easley and Jon Kleinberg.

Networks, Crowds, and Markets: Reasoning About a Highly Connected World.
Cambridge University Press, Cambridge, 2010.

G. Jeh and J. Widom.

SimRank: A measure of structural-context similarity.
In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543, 2002.

J. M. Kleinberg.

Authoritative sources in a hyperlinked environment.
J. ACM, 46(5):604–632, 1999.

Mark E. J. Newman.

Networks: An Introduction.
Oxford University Press, Oxford, 2010.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd.

The PageRank citation ranking: Bringing order to the Web.
Technical report, Stanford University, 1999.

H. Tong, Christos Faloutsos, and Jia-Yu Pan.

Fast random walk with restart and its applications.
In *Proceedings of the IEEE International Conference on Data Mining*, pages 613–622, Washington, DC, USA, 2006.