

# link *analysis*

introduction to *network analysis in Python* (*NetPy*)

Lovro Šubelj  
University of Ljubljana  
19th Sep 2019

# 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



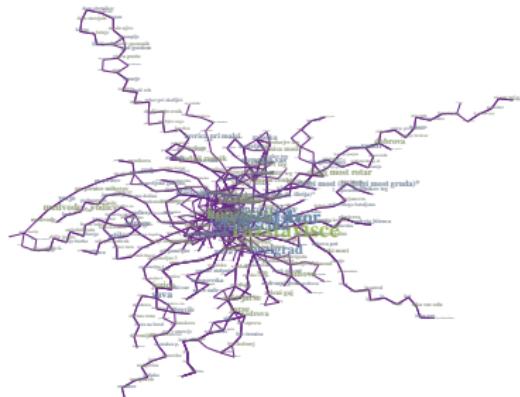
Lawrence Page



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

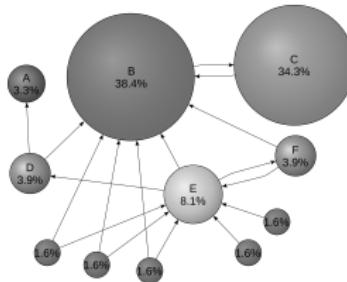
# analysis *PageRank*

*ranking algorithm for web page importance*

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

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

- $p_i$  probability *random surfer with teleports* lands on *i*



# networkology *PageRank*

- *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	Gospovshtska	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

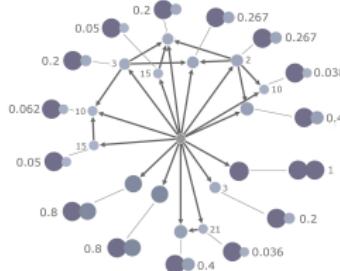
# analysis *WalkRank*

*ranking* algorithm for *web page similarity*

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

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

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



# networkology *WalkRank*

- *WalkRank ranks w* in corrected LPP network
- *highest w* nodes for *Razstavišče* and *Hajdrihova*

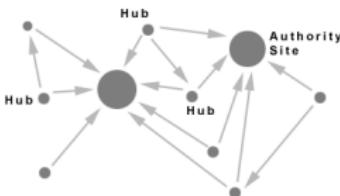
#	bus stop	$k_i$	$w_i$
1	Razstavišče	43	0.236115
2	Bavarski dvor	30	0.065124
3	Bezigrad	23	0.057260
4	Astra	16	0.047765
5	Ajdovščina	36	0.040099
6	Kozolec	10	0.038384
7	Gospovshtska	30	0.030981
8	Konzorcij	30	0.020278
9	Bavarski dvor	8	0.019262
10	Polje	10	0.014254
11	Stadion	8	0.013294
12	Topniška	8	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  $(\alpha\beta)^{-1}$
  - $\alpha$  and  $\beta$  are appropriate positive constants
- $a$  measures content and  $h$  measures table of content
- $a = 0$  for  $k^{in} = 0$  nodes and  $h = 0$  for  $k^{out} = 0$  nodes



# networkology *HITS*

- 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	Gospovetska	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	Gospovetska	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.
-  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.