advanced topics in *network science* (*ants*)

Lovro Šubelj & Jure Leskovec

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

announcements *F3 week*

- *low complexity* out *today*
- *low complexity* due *next week*
- *PhD presentations* later *today*

- *project & paper* details *today*
- *project presentations* in *two weeks*
- *project proposals* in *three weeks*

- think about *course project*
- keep your *reading list*!

challenge *F3 week*

low complexity challenge

node *centrality*

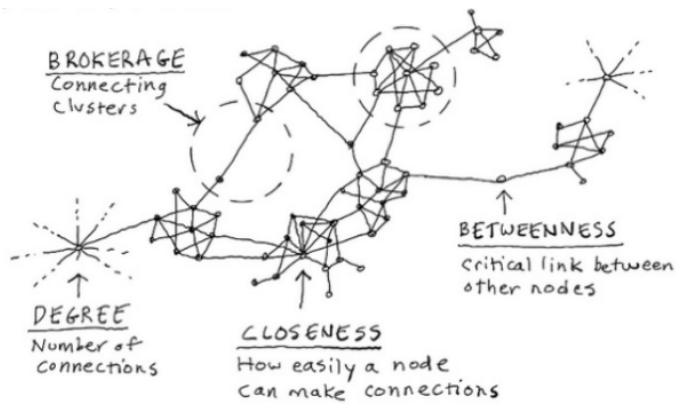
advanced topics in *network science* (*ants*)

Lovro Šubelj & Jure Leskovec
University of Ljubljana
spring 2019/20

centrality *measures*

which *nodes* are most *important*?

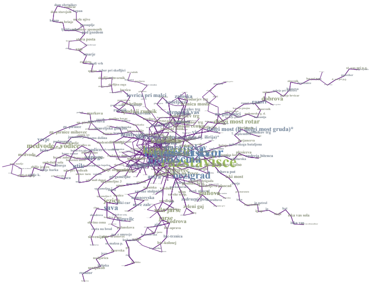
— *node centrality measures* for (*un*)*directed* networks
  - *clustering coefficients* [WS98, SV05, dNMB05]
  - *distance-based* centrality [Fre77, FBW91, New05]
  - *spectral analysis* centrality [Kat53, Bon87, BP98]
  - *fragment-based* centrality [MSOI+02, Pržˇ07, EK15]



— *link analysis algorithms* for *directed* networks

networkology *LPP*

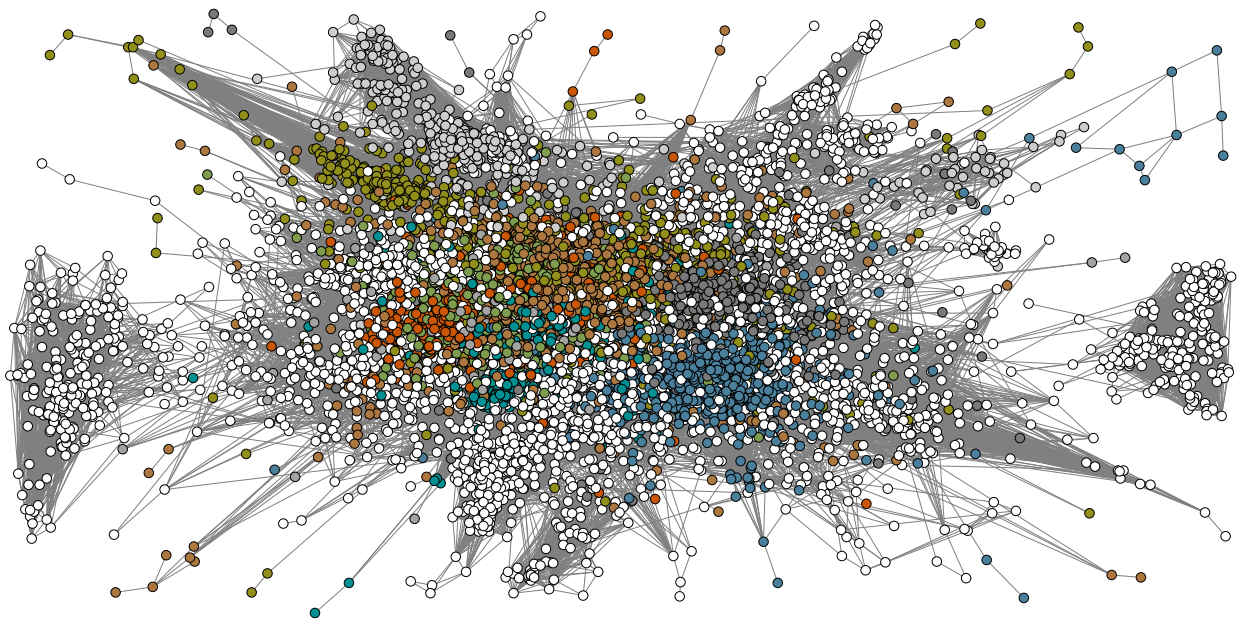
- partial *LPP* *public bus transport network**
- $n = 416$ bus stops with $\langle k \rangle = 5.62$ connections
- *giant component* 95.4% nodes (6 components)
- *"small-world"* with $\langle C \rangle = 0.09$ and $\langle d \rangle = 14.26$
- *"scale-free"* with $\gamma = 2.62$ for cutoff $k_{min}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



*reduced to largest connected component

networkology *iMDB*

- *iMDB movie actors collaboration network*
- $n = 17\,577$ actors with $\langle k \rangle = 32.7$ collaborations
- *giant component* 99.3% nodes (19 components)
- *small-world* with $\langle C \rangle = 0.34$ and $\langle d \rangle = 4.82$
- *scale-free* with $\gamma = 2.21$ for cutoff $k_{min} = 25$



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* $\{k\}$

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

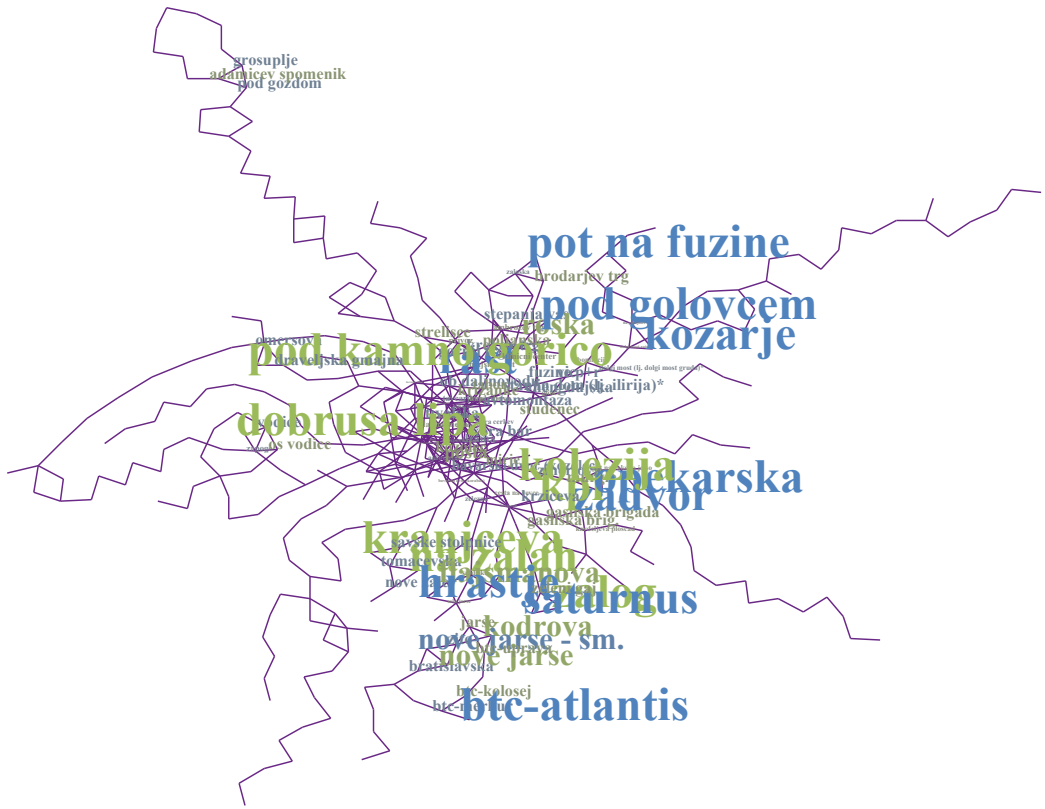
— μ-*corrected clustering coefficient* C^μ [dNMB05] of i is
  - μ is *maximum* number of *triangles* over *links*

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

networkology *LPP*

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

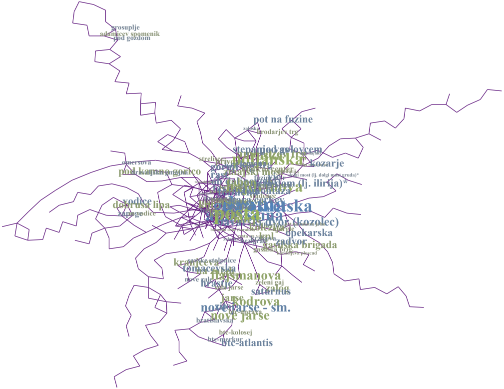
— *highest* $C_i = 1.0$ nodes are *Na Žalah etc.* with $k_i = 2$



[†] reduced to simple undirected graph

networkology *LPP*

- 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

networkology *iMDB*

— *clustering coefficient* C in iMDB network

— *highest* C nodes are *less known actors*

#	**actor**	k	C
1	Rachner, Jonathan	46	1.0000
2	Doucette, Jeff	23	1.0000
3	Willis, Susan	20	1.0000
4	Andersson, Kris	20	1.0000
5	Kurtis, Bill	19	1.0000
6	Cantillana, Nestor	16	1.0000
7	Tolkien, J.R.R.	15	1.0000
8	Faris, Anna	15	1.0000
9	Kurata, Tetsuo	13	1.0000
10	Margera, Jess	12	1.0000
11	Rakeyohn	12	1.0000
12	Raab, Chris	12	1.0000

networkology *iMDB*

— *μ-corrected clustering C^{μ}* in iMDB network

— *highest C^{μ}* nodes are *wrestling actors/models*

#	actor	k	C^{μ}
1	Helms, Shane	216	0.3689
2	Wilson, Torrie	217	0.3688
3	Matthews, Darren	212	0.3688
4	Greenwald, Nora	215	0.3687
5	Keibler, Stacy	215	0.3687
6	Jindrak, Mark	214	0.3683
7	Wight, Paul	221	0.3677
8	Guerrero Jr., Chavo	215	0.3676
9	Bischoff, Eric	218	0.3675
10	Hugger, John	212	0.3675
11	Flair, Ric	219	0.3672
12	Layfield, John	192	0.3667

centrality *closeness*

important *nodes* are *close to other* nodes

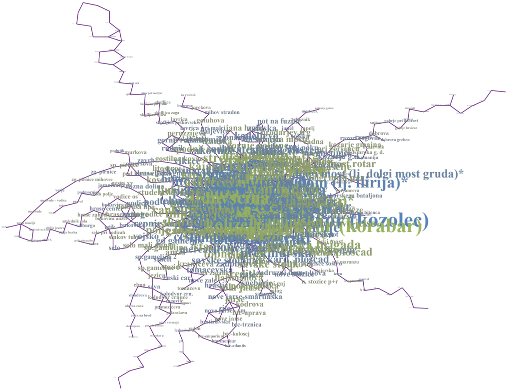
— for (*un*)*directed* G *closeness centrality* ℓ^{-1} [New10] of i is
  - d_{ij} is (*un*)*directed distance* between i and j
  - $d_{ij} = \infty$ for nodes in *different components*

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

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

networkology *LPP*

- *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

— *closeness centrality* ℓ^{-1} in iMDB network[¶]

— *highest* ℓ^{-1} nodes are *Hollywood actors*

#	actor	k	ℓ^{-1}
1	Goldberg, Whoopi	398	0.3506
2	Hanks, Tom	457	0.3500
3	Jackson, Samuel L.	427	0.3497
4	Berry, Halle	376	0.3468
5	Diaz, Cameron	361	0.3459
6	Stiller, Ben	382	0.3452
7	Lopez, Jennifer	410	0.3427
8	Myers, Mike	345	0.3409
9	Douglas, Michael	263	0.3403
10	Cruise, Tom	336	0.3401
11	Travolta, John	335	0.3388
12	Schwarzenegger, Arnold	333	0.3385

[¶] reduced to largest connected component

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

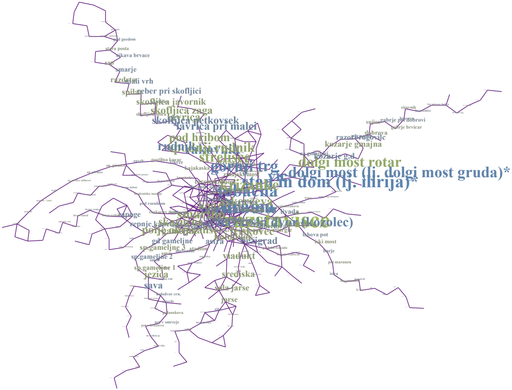
$$\sigma_i = \frac{1}{n^2} \sum_{st} \frac{g_{st}^{i}}{g_{st}}$$

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

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

networkology *LPP*

- *betweenness centrality* σ in partial LPP network[‖]
- *highest* $\sigma_i = 0.235$ node is *Razstavišče* with $k_i = 11$



‖ reduced to simple undirected graph

— *betweenness centrality* σ in iMDB network**

— *highest* σ nodes are *international actors*

#	actor	k	σ
1	Jeremy, Ron	471	0.0640
2	Chan, Jackie	135	0.0310
3	Cruz, Penelope	182	0.0284
4	Shahlavi, Darren	8	0.0282
5	Del Rosario, Monsour	6	0.0280
6	Depardieu, Gerard	159	0.0265
7	Bachchan, Amitabh	66	0.0169
8	Jackson, Samuel L.	427	0.0167
9	Soualem, Zinedine	121	0.0155
10	Del Rio, Olivia	168	0.0152
11	Jaenicke, Hannes	73	0.0140
12	Hayek, Salma	185	0.0139

** reduced to largest connected component

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

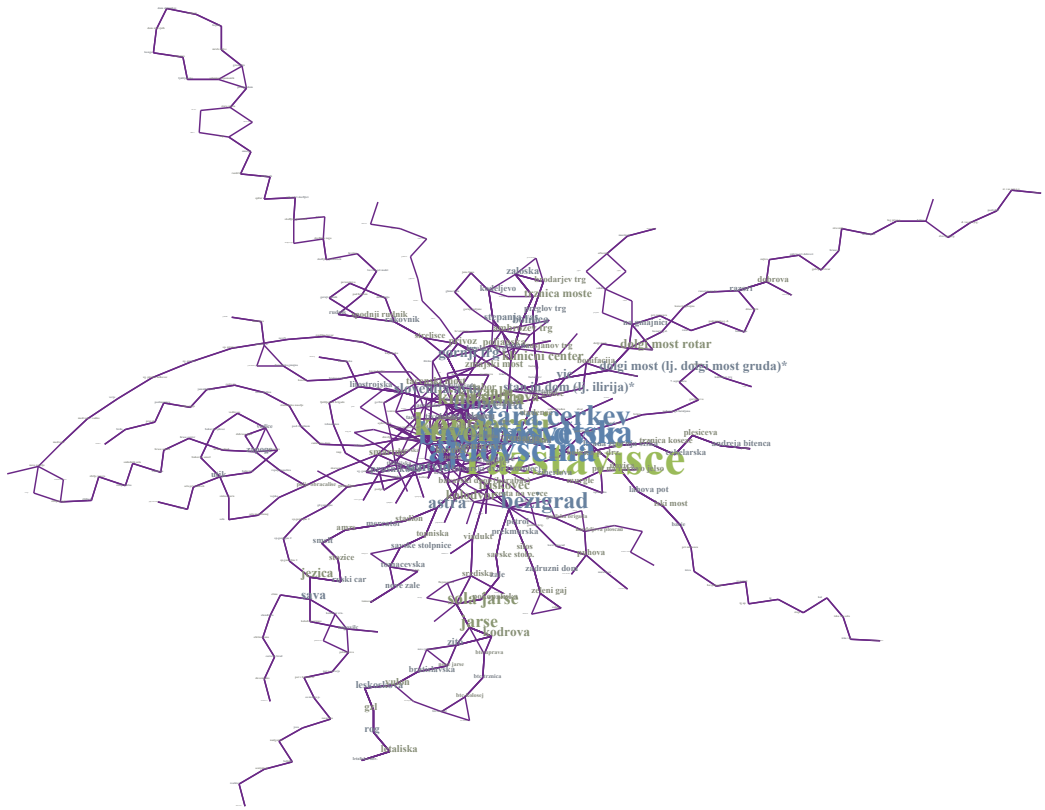
$$d_i^{in} = \frac{1}{n-1} \sum_{j \neq i} A_{ij} = \frac{k_i^{in}}{n-1}$$

— 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 *LPP*

- *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 $k_i^{out} = 21$



— *degree centrality d* in iMDB network

— *highest d* nodes are *pornographic actors*

#	**actor**	*k*	*d*
1	Davis, Mark	784	0.0446
2	Sanders, Alex	610	0.0347
3	North, Peter	599	0.0341
4	Marcus, Mr.	584	0.0332
5	Tedeschi, Tony	561	0.0319
6	Dough, Jon	555	0.0316
7	Stone, Lee	545	0.0310
8	Voyeur, Vince	533	0.0303
9	Lawrence, Joel	500	0.0284
10	Steele, Lexington	493	0.0280
11	Ashley, Jay	490	0.0279
12	Boy, T.T.	475	0.0270

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

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

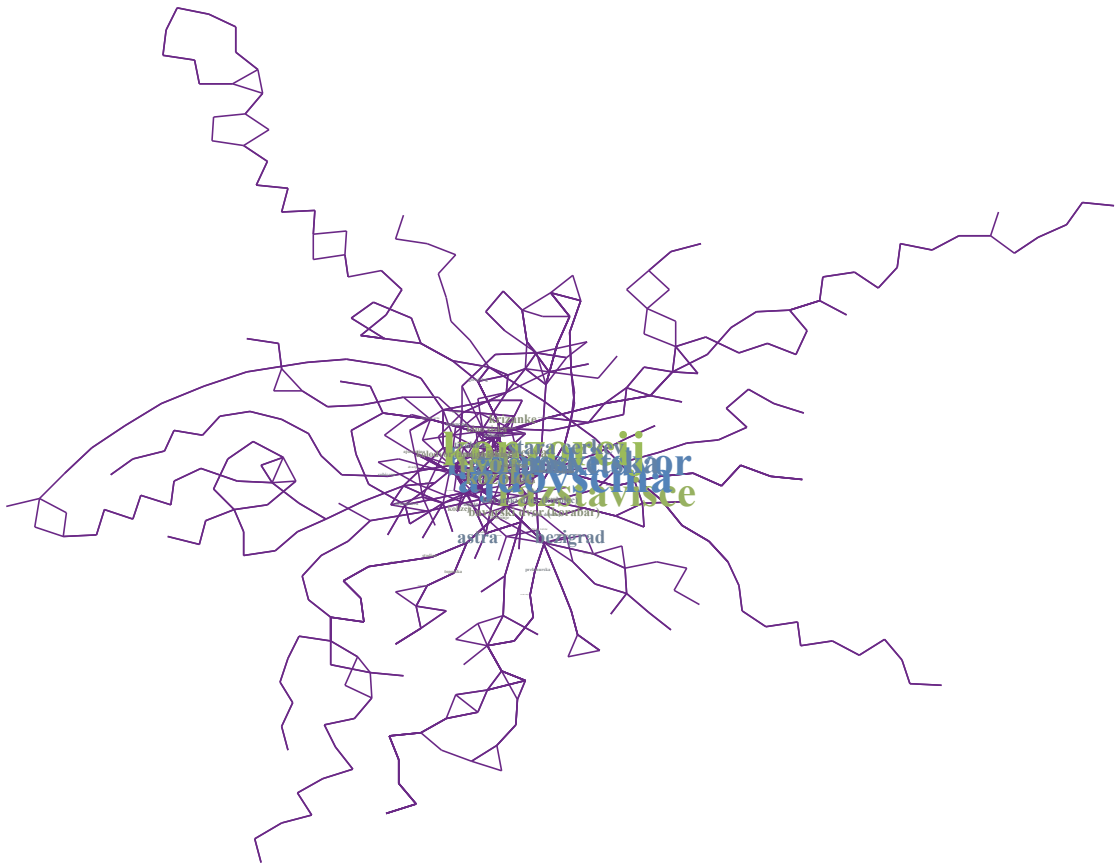
$$e_i = \lambda_1^{-1} \sum_j A_{ij} e_j$$

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

networkology *LPP*

— *eigenvector centrality* e in partial LPP network

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



— *eigenvector centrality* e in iMDB network

— *highest* e nodes are *wrestling actors*

#	**actor**	k	e
1	Benoit, Chris	225	0.005261
2	Guerrero, Eddie	225	0.005261
3	Storm, Lance	225	0.005236
4	Wight, Paul	221	0.005231
5	Jericho, Chris	253	0.005230
6	Runnels, Dustin	220	0.005206
7	Flair, Ric	219	0.005203
8	Huffman, Booker	219	0.005199
9	Bischoff, Eric	218	0.005192
10	Wilson, Torrie	217	0.005192
11	Gruner, Peter	218	0.005183
12	Levy, Scott	226	0.005181

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*

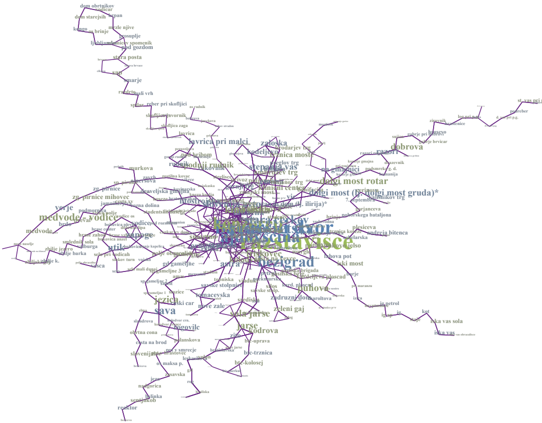
$$p_i = \alpha \sum_j A_{ij} \frac{p_j}{k_j^{out}} + \beta_j$$

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

see PageRank algorithm NetLogo demo

networkology *LPP*

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



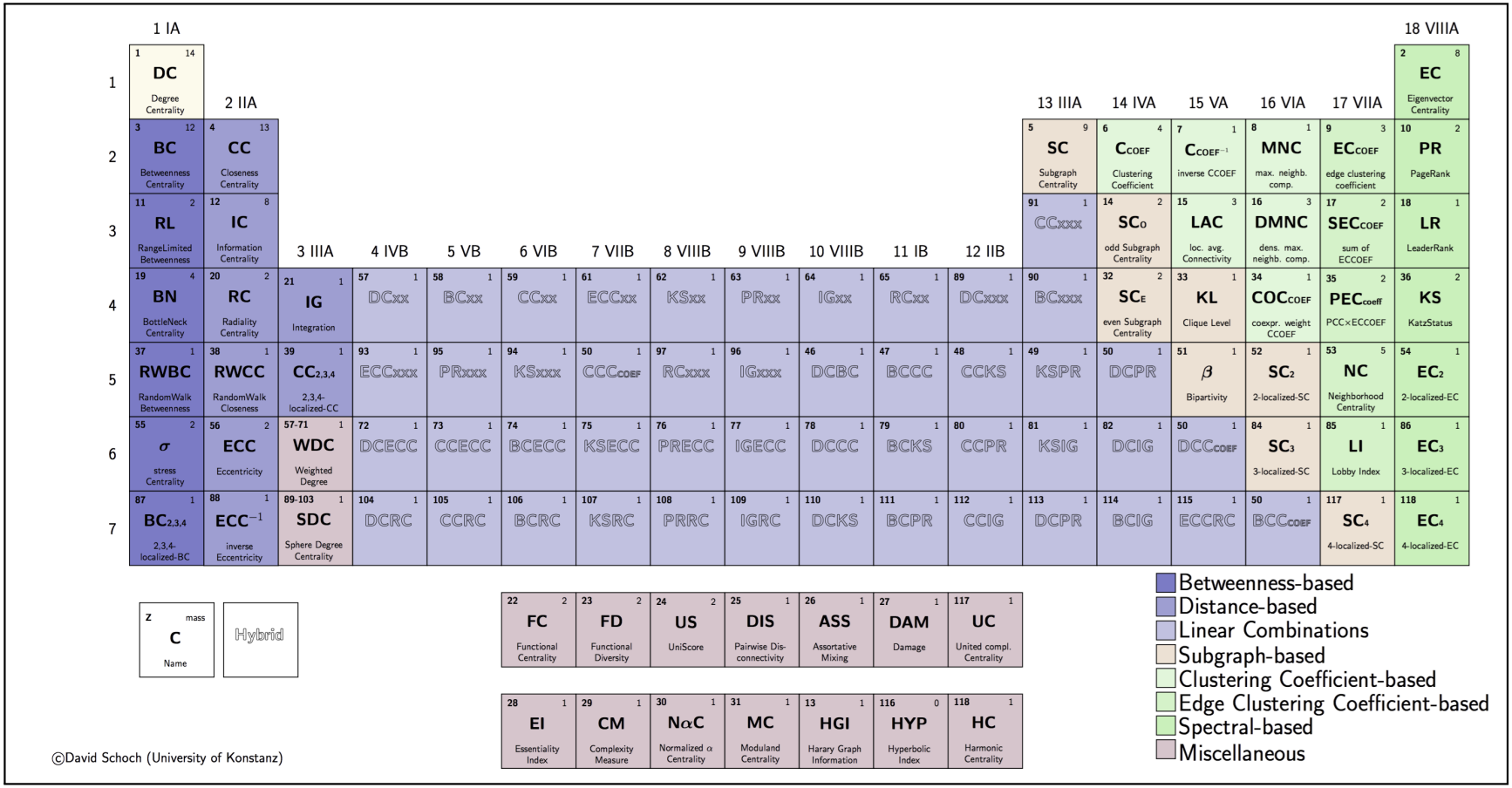
— *PageRank centrality p* in iMDB network

— *highest p* nodes are *Hollywood actors*

#	**actor**	*k*	*p*
1	Hanks, Tom	457	0.000660
2	Jackson, Samuel L.	427	0.000642
3	Goldberg, Whoopi	398	0.000607
4	Stiller, Ben	382	0.000561
5	Davis, Mark	784	0.000547
6	Lopez, Jennifer	410	0.000542
7	Berland, Francois	194	0.000540
8	Berry, Halle	376	0.000536
9	Diaz, Cameron	361	0.000514
10	Travolta, John	335	0.000486
11	Jeremy, Ron	471	0.000476
12	Myers, Mike	345	0.000468

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.

centrality *references*

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

link *analysis*

advanced topics in *network science* (*ants*)

Lovro Šubelj & Jure Leskovec
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]

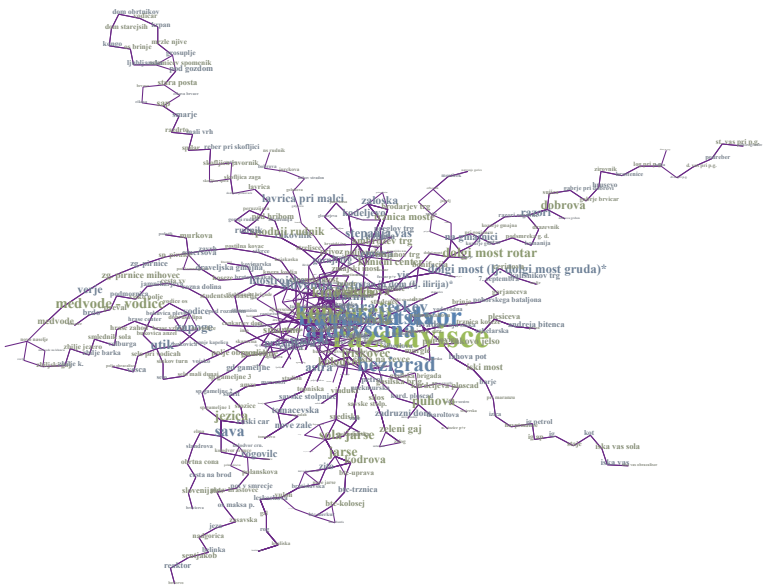
Sergey Brin

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*

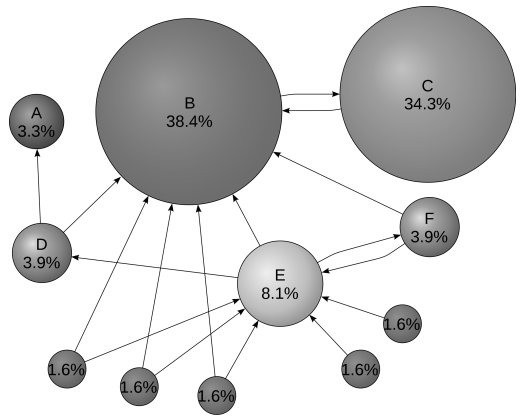
ranking algorithm for *web page importance*

— 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



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	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*

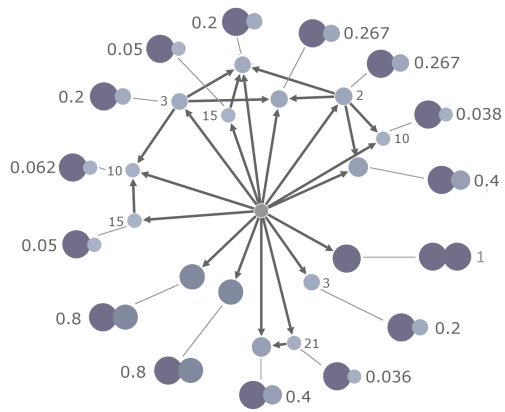
ranking algorithm for *web page similarity*

— 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*

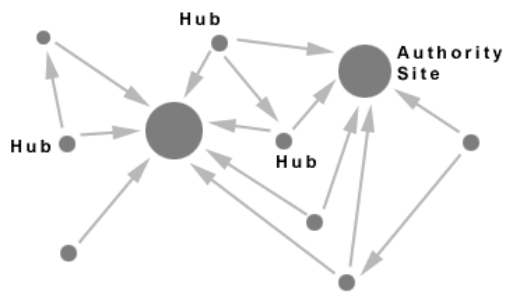
ranking algorithm for *web hubs & authorities*

— 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.

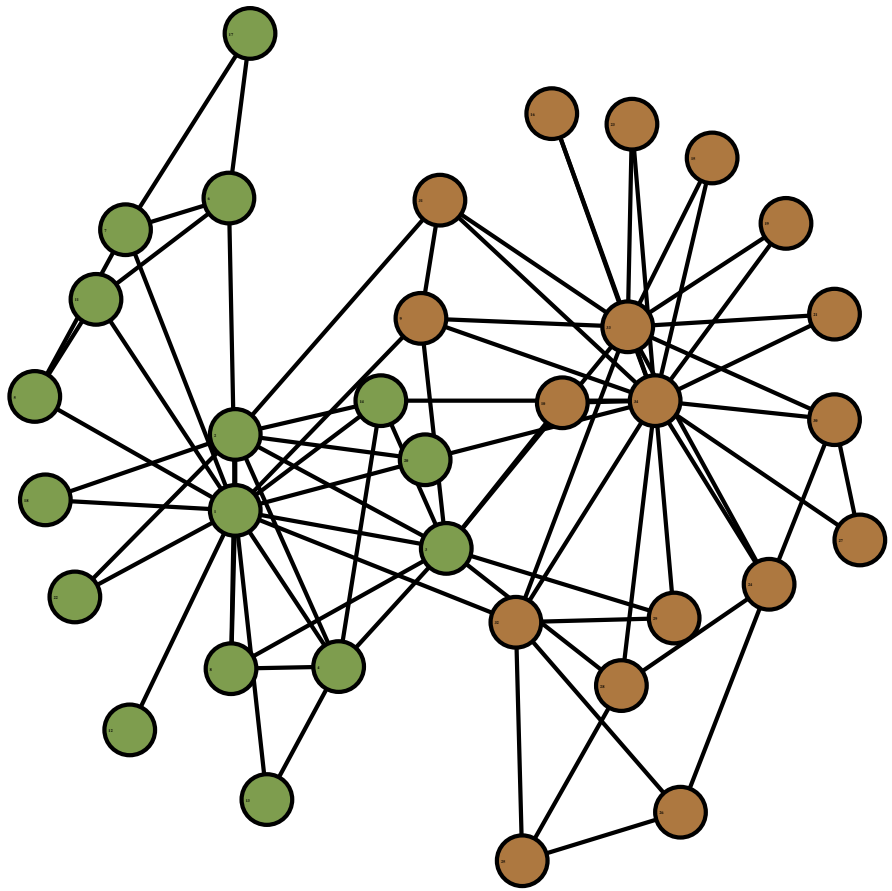
community structure

advanced topics in *network science* (*ants*)

Lovro Šubelj & Jure Leskovec
University of Ljubljana
spring 2019/20

community *structure*

karate club *network split* [Zac77]

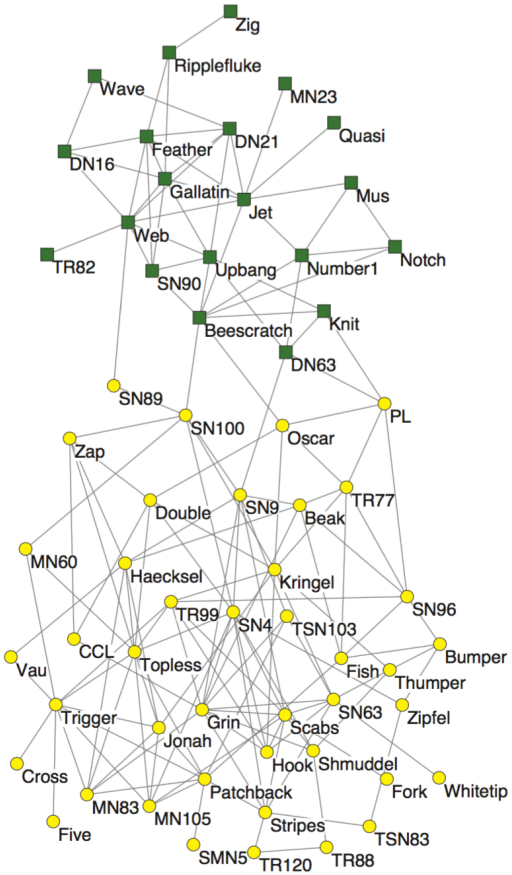


community *detection*

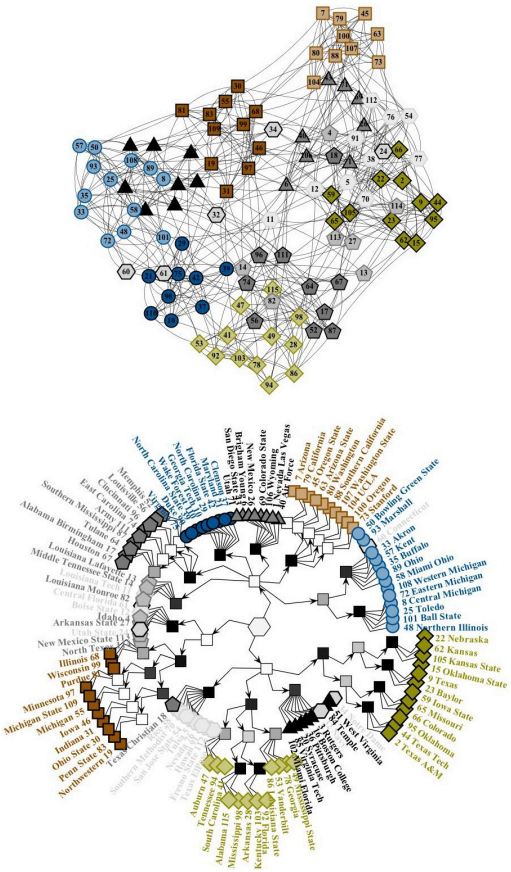
karate club *split detection* [RAK07]

community *examples*

most social networks contain *communities* [GN02]



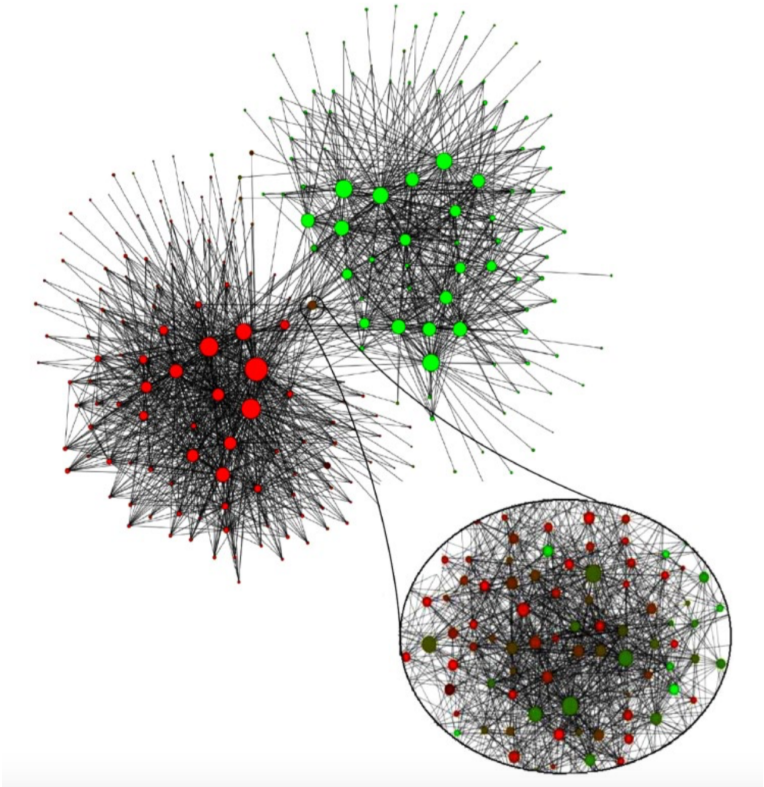
bottlenose dolphins [LSB+03]



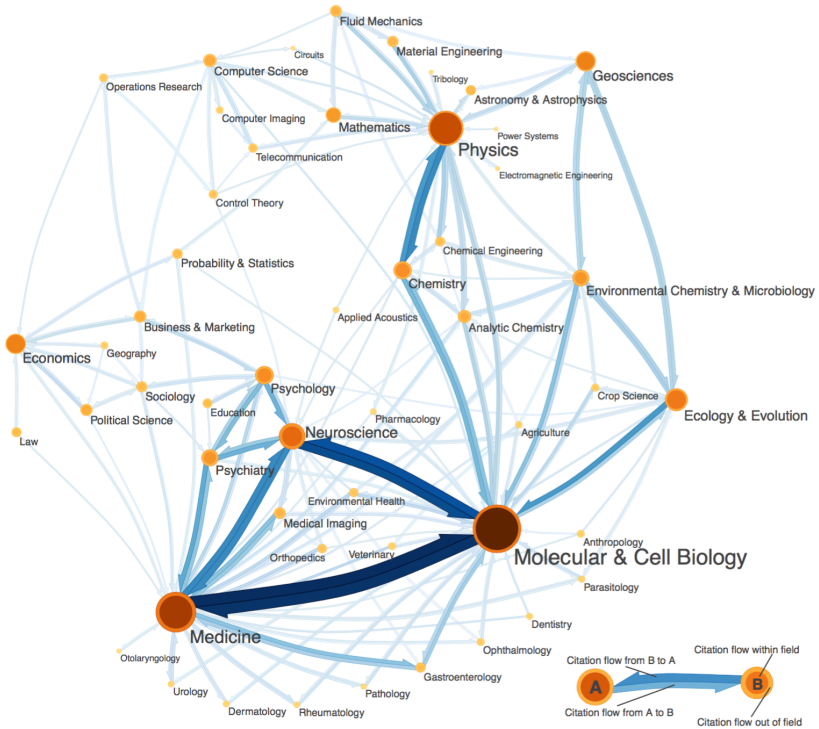
college football [GN02]

community *examples*

many information networks contain *communities* [FLG00]



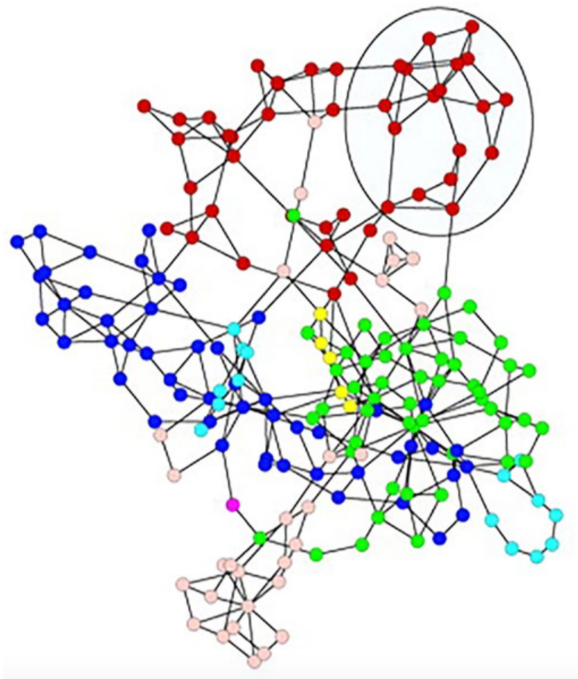
mobile communications [BGLL08]



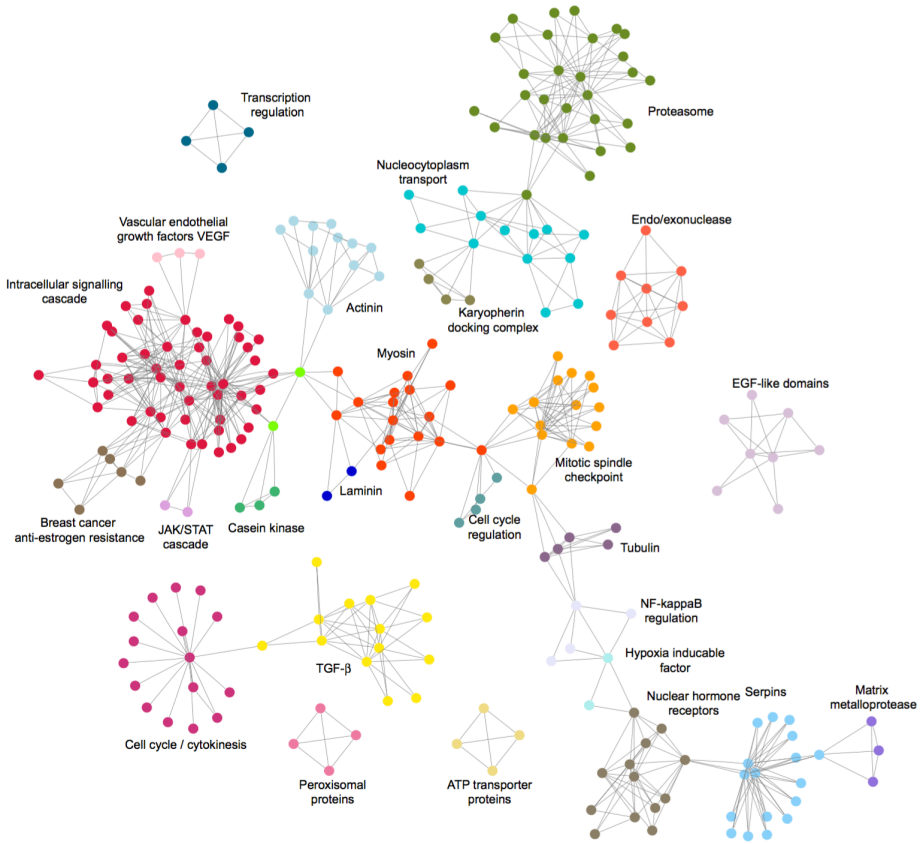
journal citations [RB08]

community *examples*

many biological networks contain *communities* [RSM+02]



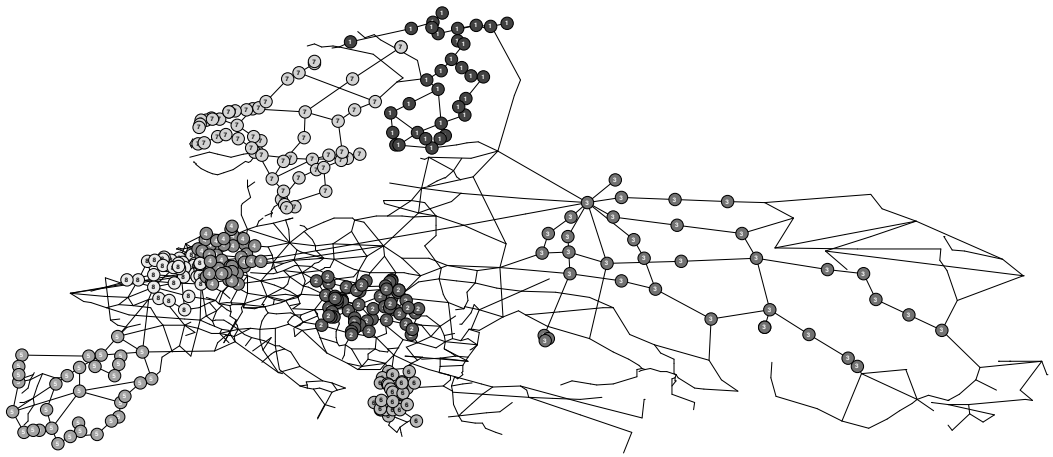
E. coli metabolism [RSM+02]



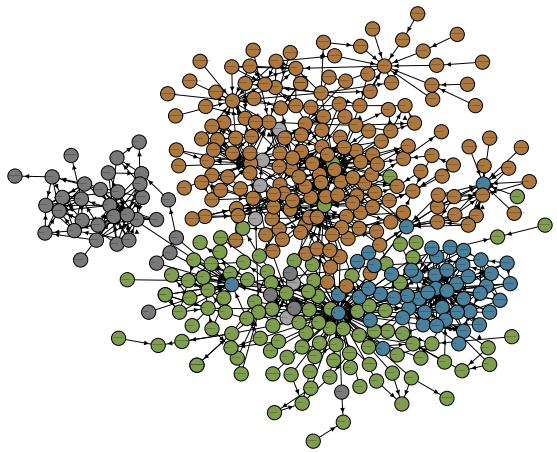
protein interactions [JCZB06]

community *examples*

some technological networks contain *communities* [ŠB11a]



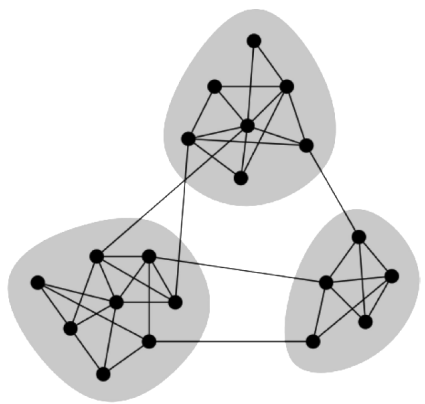
European highways [ŠB11b]



JUNG dependencies [ŠB11a]

community *explanation*

- *weak* & *strong ties* according to *information flow*
- *bridges* & *embedded ties* according to *network span*
  - removal of *local bridge* $\{i, j\}$ causes $d_{ij} > 2$
  - removal of *bridge* $\{i, j\}$ causes $d_{ij} = \infty$
  - *embedded tie* $\{i, j\}$ has $C_{ij} > 0$



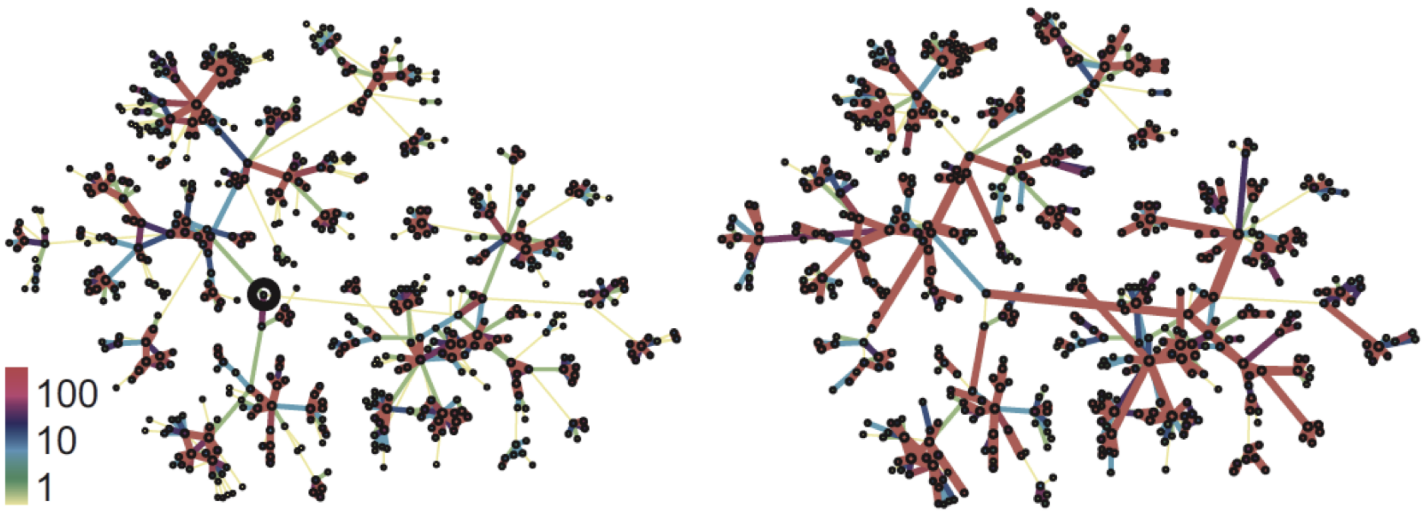
strength of weak ties or weakness of strong ties

- *weak ties are* (*local*) *bridges* under *triadic closure* [Gra73]
- *assortative mixing* and *homophily* in (social) networks [NG03]

community *experiment*

— *tie strength* in mobile communications [OSH+07]

— *weak ties are* (*local*) *bridges* in real networks



real strengths *shuffled strengths*

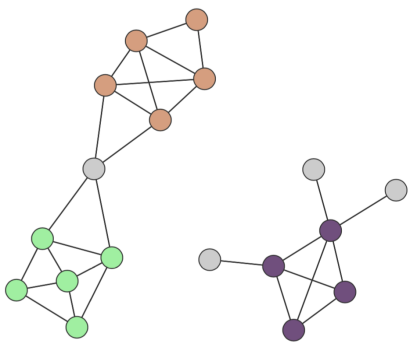
strength of weak ties or weakness of strong ties

community *definition*

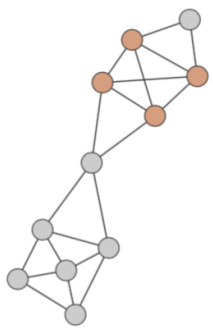
- *clique* is *complete subgraph of* some *graph*
  - also *k-plexes*, *k-cores*, *k-cliques*, *k-clubs*, *k-clans*
- *community* is *dense subgraph of sparse network* [GN02]
- *strong* and *weak community* C [FLG00, RCC+04] defined as
  - k_i^{int} and k_i^{ext} are *internal* and *external degree* of i

$$\forall i \in C : k_i^{int} > k_i^{ext} \qquad \sum_{i \in C} k_i^{int} > \sum_{i \in C} k_i^{ext}$$

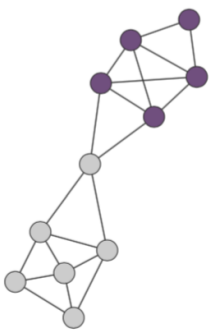
- *community detection* is *not graph partitioning* [For10]



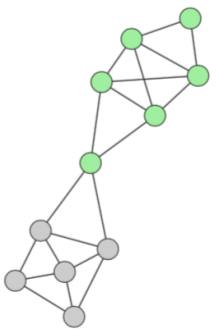
connected communities



maximum clique



strong community



weak community

community *modularity*

— *random graph lacks community structure*

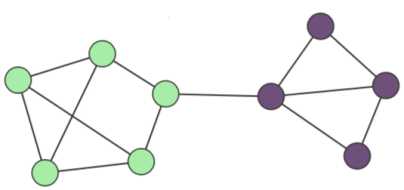
— *modularity* Q [GN02] of *communities* $\{C\}$ defined as

– $k_c = \sum_{i \in C} k_i$ is *total degree* and m_c is *number of links* in C

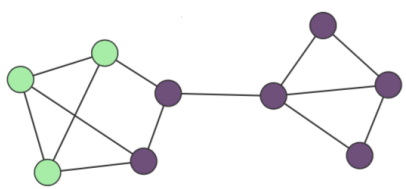
$$\frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{c_i c_j} = \frac{1}{2m} \sum_{C} \sum_{ij \in C} \left(A_{ij} - \frac{k_i k_j}{2m} \right) = \sum_{C} \frac{m_c}{m} - \left(\frac{k_c}{2m} \right)^2$$

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{c_i c_j} = \sum_{C} \frac{m_c}{m} - \left(\frac{k_c}{2m} \right)^2$$

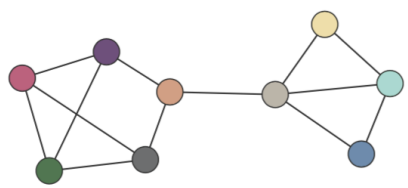
— *modularity* Q popular *quality*/*optimization function* [For10]



optimal $Q = 0.41$



suboptimal $Q = 0.22$



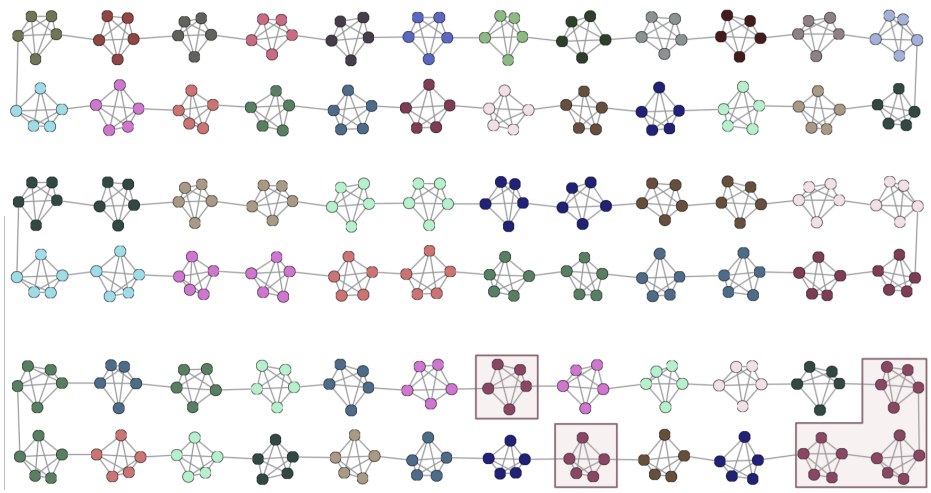
isolates $Q = -0.12$

community ¬*modularity*

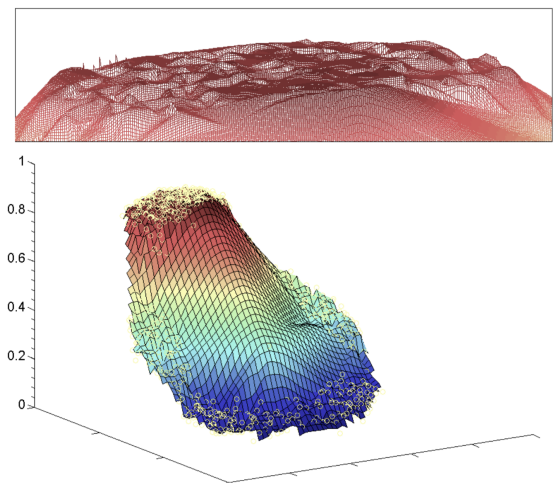
— *modularity* $Q \gg 0$ also in *random graphs* [GSPA04]

— *modularity* Q has *resolution limit* at $k_c \leq \sqrt{2m}$ [FB07]

— *modularity* Q *lacks clear optimum* in real networks [GdMC10]

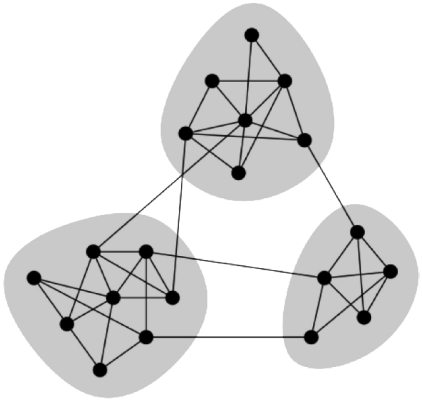


intuitive $Q = 0.867$, *optimal* $Q = 0.871$ and random $Q = 0.8$

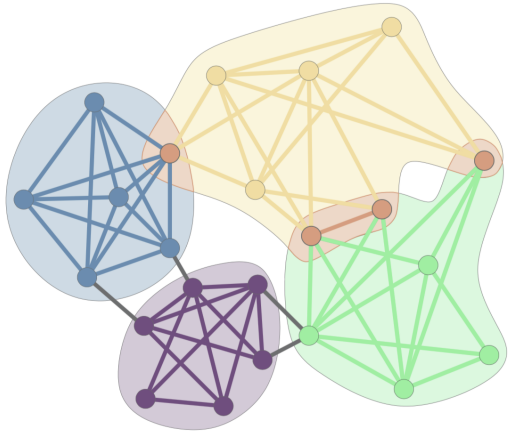


Q *plateau* and maxima

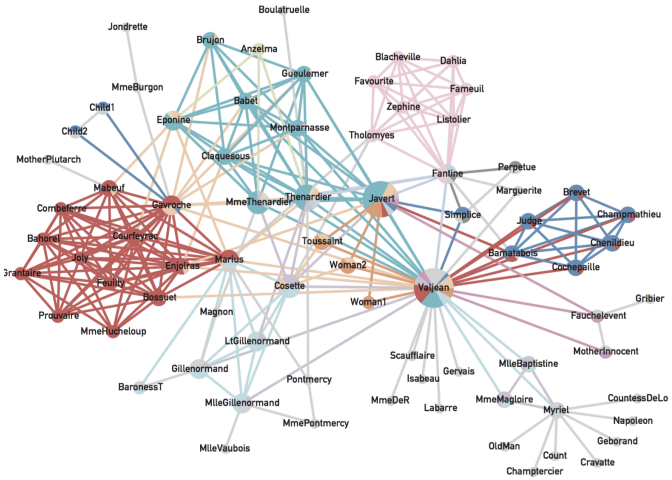
community *overview*



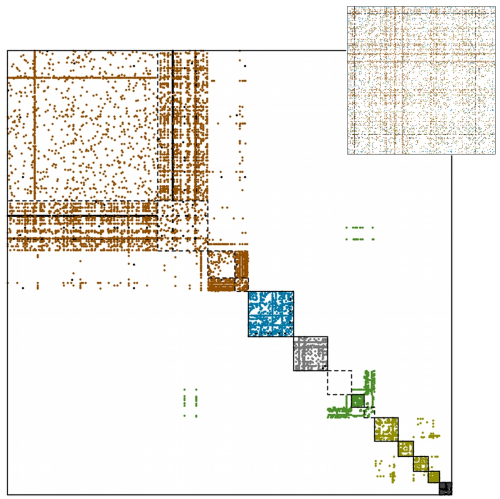
communities [GN02]



overlapping communities [PDFV05]



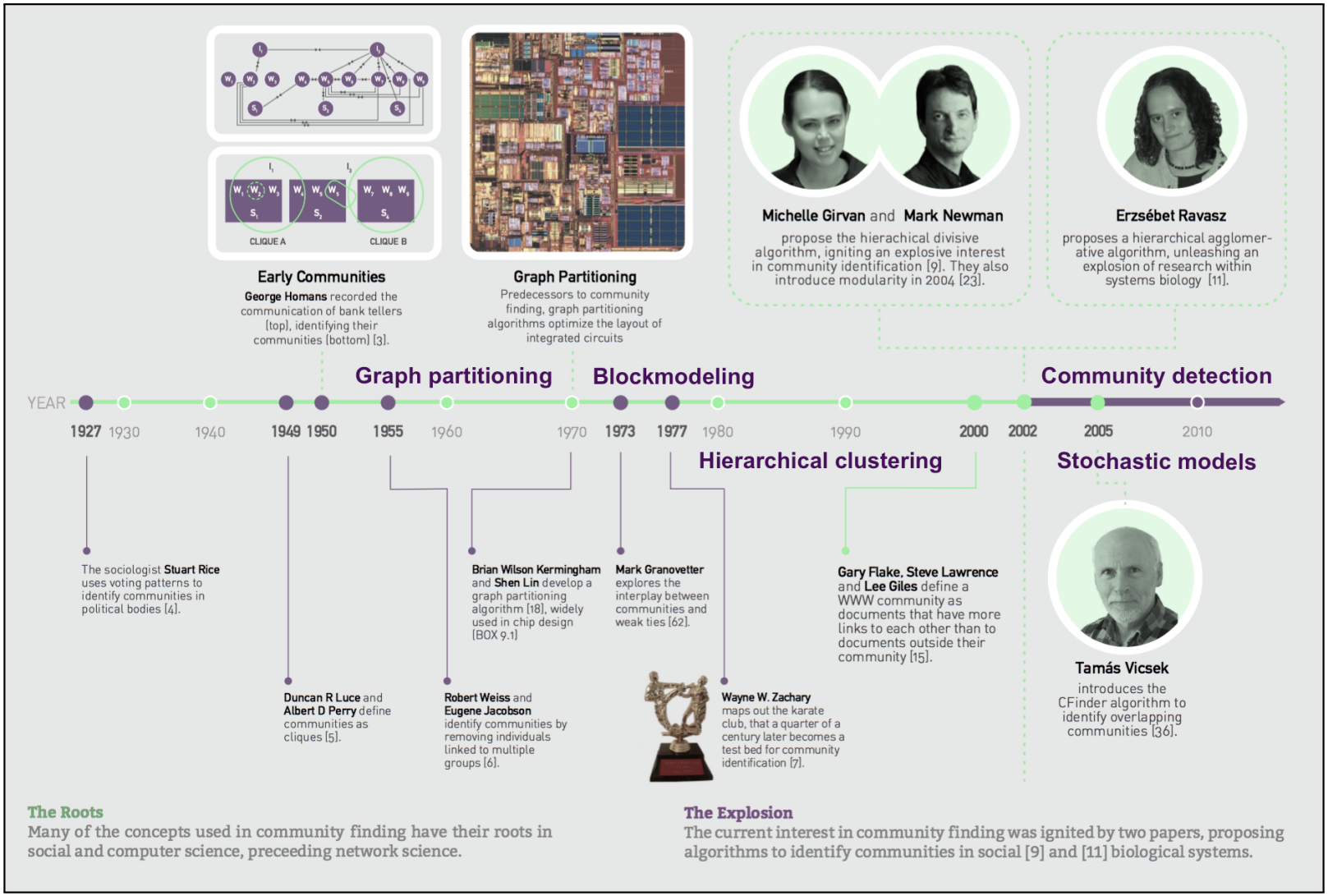
link communities [EL09, ABL10]



block models, *blockmodeling* etc.

`javax.swing`, `javax.management`, `javax.xml`, `javax.print`, `javax.naming`, `javax.lang`

community *history*



community *references*

Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann.
Link communities reveal multiscale complexity in networks.
Nature, 466(7307):761–764, 2010.

A.-L. Barabási.
Network Science.
Cambridge University Press, Cambridge, 2016.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre.
Fast unfolding of communities in large networks.
J. Stat. Mech., P10008, 2008.

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.

T. S. Evans and R. Lambiotte.
Line graphs, link partitions and overlapping communities.
Phys. Rev. E, 80(1):016105, 2009.

Santo Fortunato and Marc Barthelemy.
Resolution limit in community detection.
P. Natl. Acad. Sci. USA, 104(1):36–41, 2007.

community *references*

Gary William Flake, Steve Lawrence, and C. Lee Giles.

Efficient identification of web communities.
In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, USA, 2000.

Santo Fortunato.

Community detection in graphs.
Phys. Rep., 486(3-5):75–174, 2010.

Benjamin H. Good, Yves Alexandre de Montjoye, and Aaron Clauset.

Performance of modularity maximization in practical contexts.
Phys. Rev. E, 81(4):046106, 2010.

M. Girvan and M. E. J Newman.

Community structure in social and biological networks.
P. Natl. Acad. Sci. USA, 99(12):7821–7826, 2002.

Mark S. Granovetter.

The strength of weak ties.
Am. J. Sociol., 78(6):1360–1380, 1973.

Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral.

Modularity from fluctuations in random graphs and complex networks.
Phys. Rev. E, 70(2):025101, 2004.

Pall F. Jonsson, Tamara Cavanna, Daniel Zicha, and Paul A. Bates.

Cluster analysis of networks generated through homology: Automatic identification of important protein communities involved in cancer metastasis.
BMC Bioinformatics, 7:2, 2006.

community *references*

D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson.

The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait?
Behav. Ecol. Sociobiol., 54(4):396–405, 2003.

Mark E. J. Newman.

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

M. E. J. Newman and M. Girvan.

Mixing patterns and community structure in networks.
Phys. Rev. E, 67(2):026126, 2003.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási.

Structure and tie strengths in mobile communication networks.
P. Natl. Acad. Sci. USA, 104(18):7332–7336, 2007.

Gergely Palla, Imre Derényi, Illes Farkas, and Tamas Vicsek.

Uncovering the overlapping community structure of complex networks in nature and society.
Nature, 435(7043):814–818, 2005.

Usha Nandini Raghavan, Reka Albert, and Soundar Kumara.

Near linear time algorithm to detect community structures in large-scale networks.
Phys. Rev. E, 76(3):036106, 2007.

M. Rosvall and C. T. Bergstrom.

Maps of random walks on complex networks reveal community structure.
P. Natl. Acad. Sci. USA, 105(4):1118–1123, 2008.

community *references*

Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi.
Defining and identifying communities in networks.
P. Natl. Acad. Sci. USA, 101(9):2658–2663, 2004.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and Albert László Barabási.
Hierarchical organization of modularity in metabolic networks.
Science, 297(5586):1551–1555, 2002.

Lovro Šubelj and Marko Bajec.
Community structure of complex software systems: Analysis and applications.
Physica A, 390(16):2968–2975, 2011.

Lovro Šubelj and Marko Bajec.
Robust network community detection using balanced propagation.
Eur. Phys. J. B, 81(3):353–362, 2011.

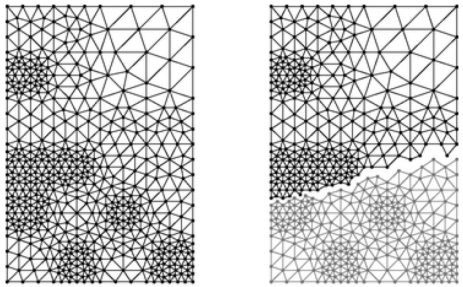
Wayne W. Zachary.
An information flow model for conflict and fission in small groups.
J. Anthropol. Res., 33(4):452–473, 1977.

network *clustering*

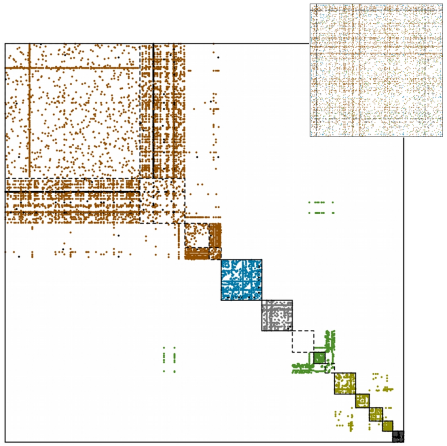
advanced topics in *network science* (*ants*)

Lovro Šubelj & Jure Leskovec
University of Ljubljana
spring 2019/20

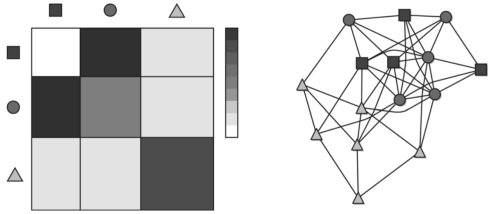
clustering *overview*



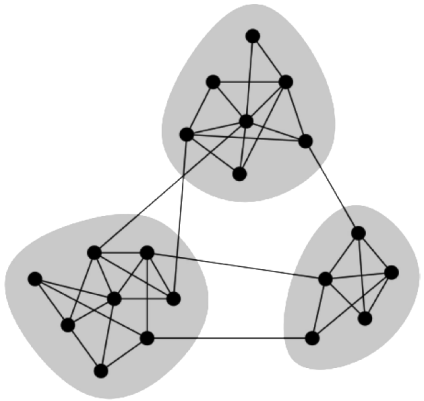
graph partitioning [KL70, Fie73]



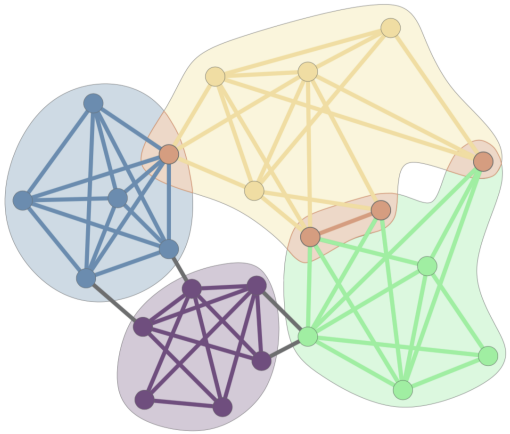
blockmodeling [LW71, WR83]



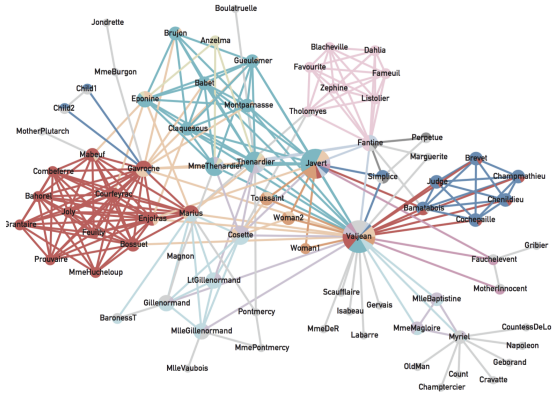
stochastic block models [Pei15]



communities [GN02]

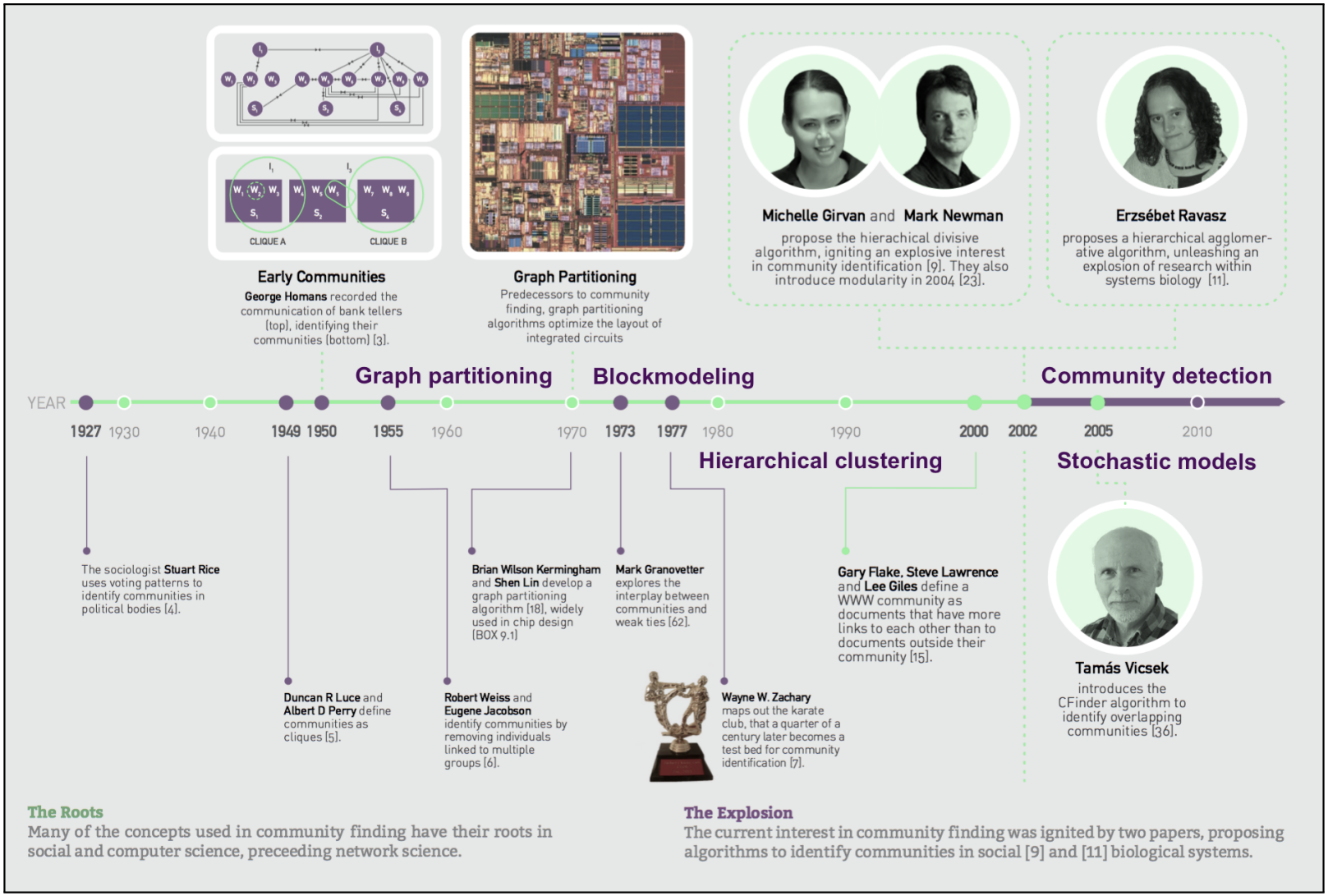


overlapping communities [PDFV05]



link communities [EL09, ABL10]

clustering *history*



graph *partitioning*

advanced topics in *network science* (*ants*)

Lovro Šubelj & Jure Leskovec
University of Ljubljana
spring 2019/20

partitioning *bisection*

— Kernighan-Lin *graph bisection* [KL70]

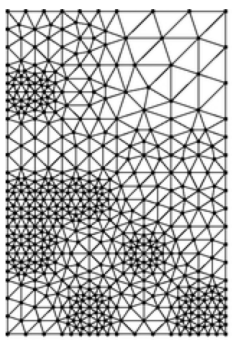
– define *bisection quality* as *cut size*

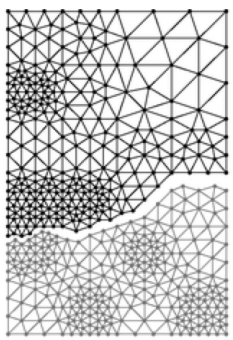
$$R = \frac{1}{2}\sum_{ij} A_{ij}(1 - \delta_{c_i c_j}) \qquad \forall i : c_i = \pm 1$$

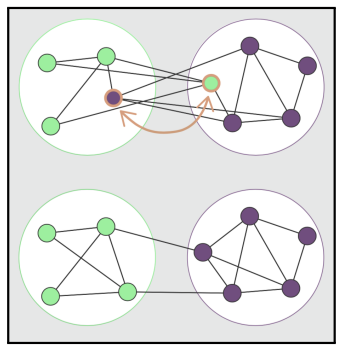
1. *swap nodes* by *minimizing cut size* $\mathcal{O}(cn^2m)$

$$\Delta R_{ij} = k_i^{ext} - k_i^{in} + k_j^{ext} - k_j^{in} - 2A_{ij}$$

2. repeat 1. *until* $\min(n_1, n_2)$ *nodes swapped*
3. return *bisection minimizing cut size*







*example mesh bisection with cut size equal to 40

partitioning *spectral*

— Fiedler *graph bisection* [Fie73]

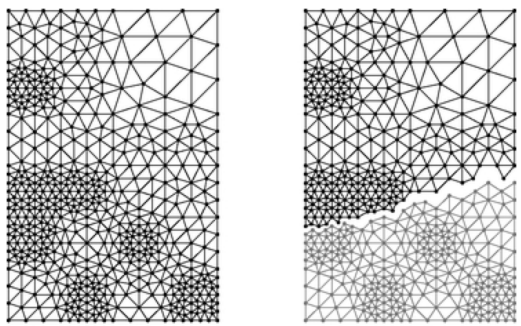
– define *bisection quality* as *cut size*

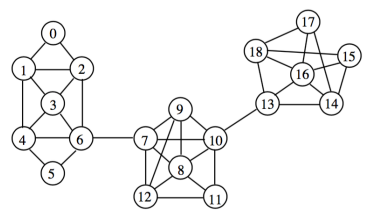
$$R = \frac{1}{4}\sum_{ij} A_{ij}(1 - s_i s_j) \qquad \forall i : s_i = \delta_{c_i c_1} - \delta_{c_i c_2}$$

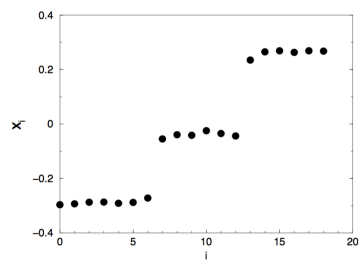
– formulate *eigenvector problem* of *graph Laplacian*

$$R = \frac{1}{4}\sum_i k_i s_i^2 - \frac{1}{4}\sum_{ij} A_{ij} s_i s_j = \frac{1}{4}\sum_{ij}(k_i\delta_{ij} - A_{ij})s_i s_j = \frac{1}{4}s^T L s \simeq \frac{1}{4}v^T L v = \frac{n_1 n_2}{n}\lambda$$

1. find *eigenvector* v_2 with *algebraic connectivity* λ_2 $\mathcal{O}(nm)$
2. assign n_1 *nodes* with *largest*/*smallest* v_2 to C_1
3. return *bisection minimizing cut size*







see graclus and metis implementations

† example mesh bisection with cut size equal to 46

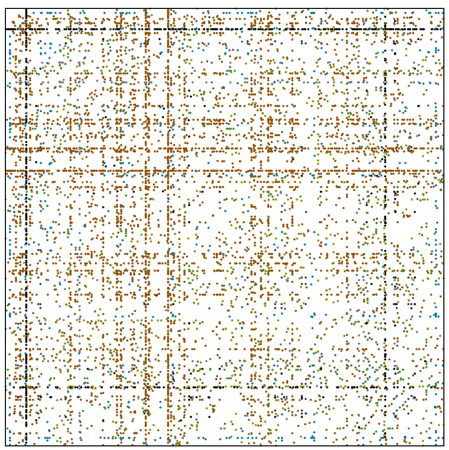
partitioning *blockmodeling*

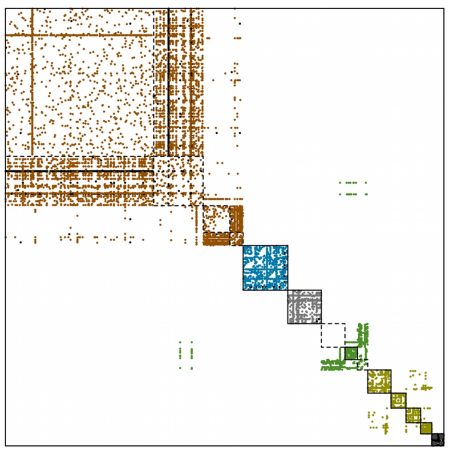
— standard *equivalence blockmodeling* [DBF05]

– define *node similarity* as *(structural) equivalence*

$$s_{ij} \sim |\Gamma_i \cap \Gamma_j|$$

1. *blockmodeling* by *(hierarchical) clustering* $\mathcal{O}(n^2)$
2. return *block model* at desired *clustering resolution*





see catrege implementation

‡ `javax.swing`, `javax.management`, `javax.naming`, `javax.print`, `javax.xml`, `javax.lang` etc.

community detection

advanced topics in *network science* (*ants*)

Lovro Šubelj & Jure Leskovec
University of Ljubljana
spring 2019/20

community *agglomerative*

— Ravasz *hierarchical clustering* [RSM+02]

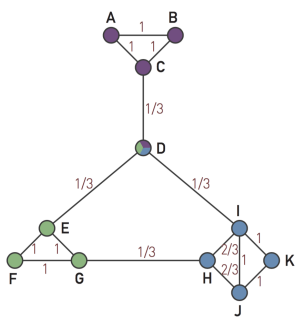
- define *node similarity* as *topological overlap*

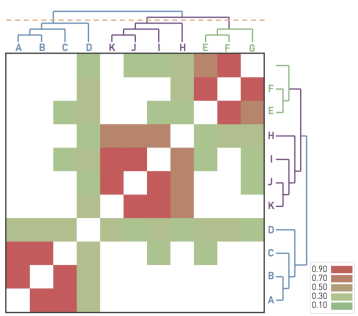
$$s_{ij} = \frac{|\Gamma_i \cap \Gamma_j| + A_{ij}}{\min(k_i, k_j)}$$

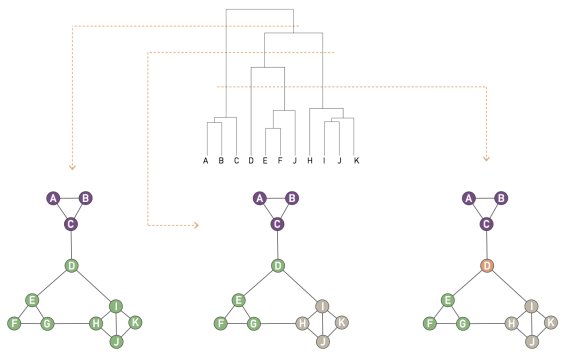
- define *cluster similarity* as *average linkage*

$$S_{ij} = \frac{1}{n_i n_j} \sum_{xy} s_{xy} \delta_{c_x c_i} \delta_{c_y c_j}$$

1. bottom-up *agglomerative hierarchical clustering* $\mathcal{O}(n^2)$
2. cut *cluster dendrogram* at desired *clustering resolution*







community *divisive*

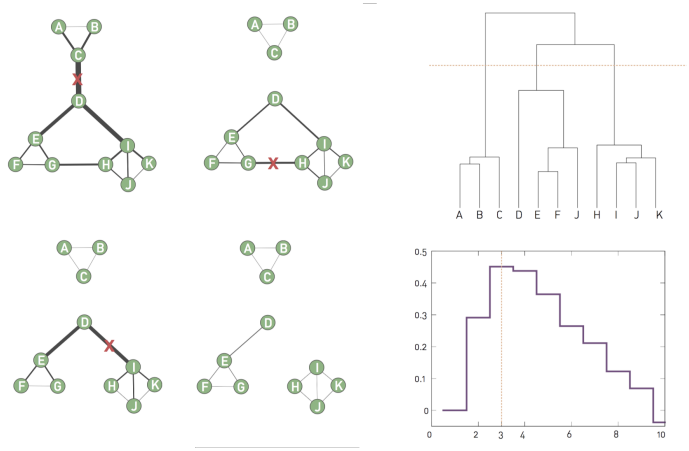
— Girvan-Newman *hierarchical clustering* [GN02]

– define *node dissimilarity* as *link betweenness*

$$\sigma_{ij} = \sum_{st \notin \{i,j\}} \frac{g_{st}^{ij}}{g_{st}}$$

1. top-down *divisive hierarchical clustering* $\mathcal{O}(nm^2)$
2. cut *cluster dendrogram* at *maximum modularity*

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta_{c_i c_j}$$

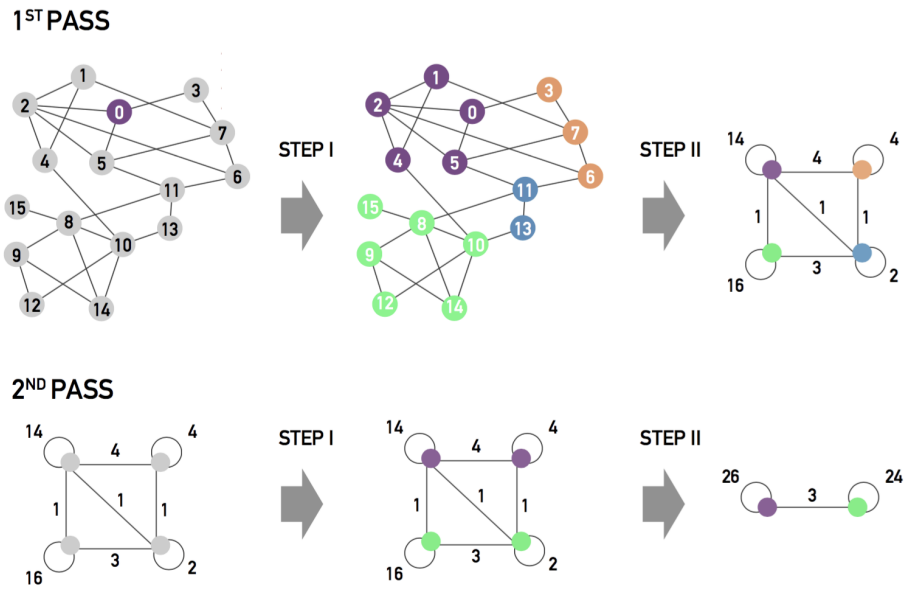


community *modularity*

— Louvain *modularity optimization* [BGLL08]

1. set *node community* by *modularity optimization* $\mathcal{O}(cm)$
2. *aggregate community nodes into supernodes* and repeat 1.
3. return *community structure maximizing modularity*

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta_{c_i c_j}$$



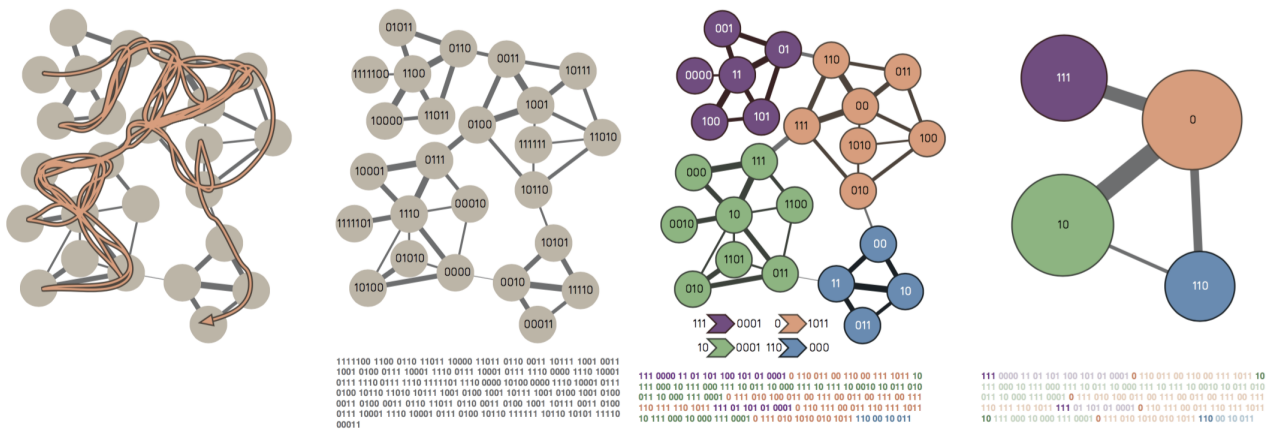
see findcommunities implementation

community *map equation*

— Infomap *map equation compression* [RB08]

1. set *node community* by *optimal coding* $\mathcal{O}(m \log m)$
2. *compress community nodes into supernodes* and repeat 1.
3. return *community structure maximizing map equation*

$$\mathcal{L} = \sum_i p_{i\rightsquigarrow} H(\widetilde{\mathcal{C}}) + \sum_i p_{i\hookleftarrow} H(\mathcal{C}_i)$$



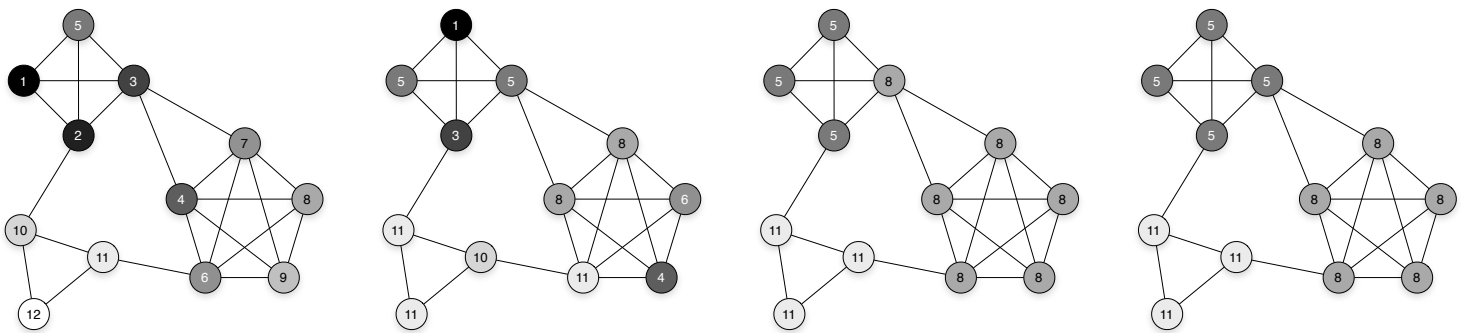
see mapequation implementation

community *propagation*

— Raghavan *label propagation* [RAK07, ŠB11]

1. set *node community* by *neighbors frequency* $\mathcal{O}(cm)$
2. *randomly shuffle nodes* and repeat 1. *until convergence*
3. return *community structure connected components*

$$\forall i : c_i = \arg\max_c \sum_j A_{ij}\delta_{c_j c}$$



see balanced implementation

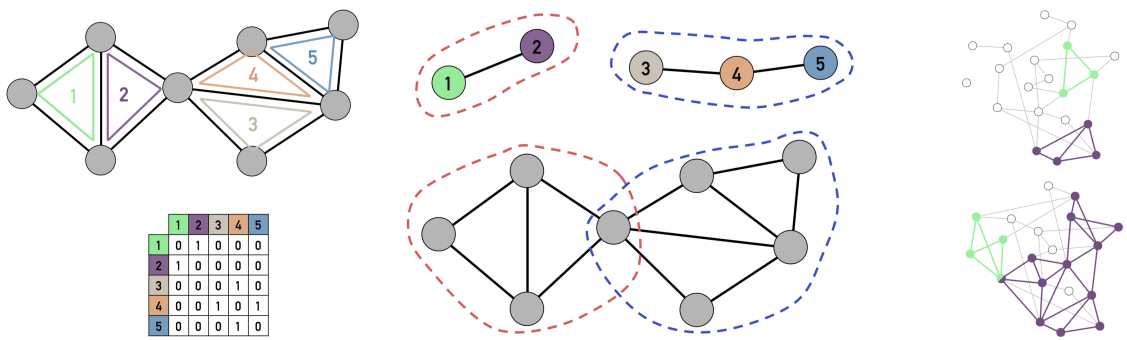
community *percolation*

— Palla *clique percolation* [PDFV05, KKKS08]

1. find *k-node cliques* by *sequential enumeration* $\mathcal{O}(n_k)$
2. *merge clique nodes into supernodes* and *link adjacent*

   adjacent k-node cliques share $k-1$ nodes
3. return *clique structure connected components*

   clique percolation at $(kn-n)^{\frac{1}{1-k}}$



see kclique implementation

community *links*

— Ahn *link clustering* [EL09, ABL10]

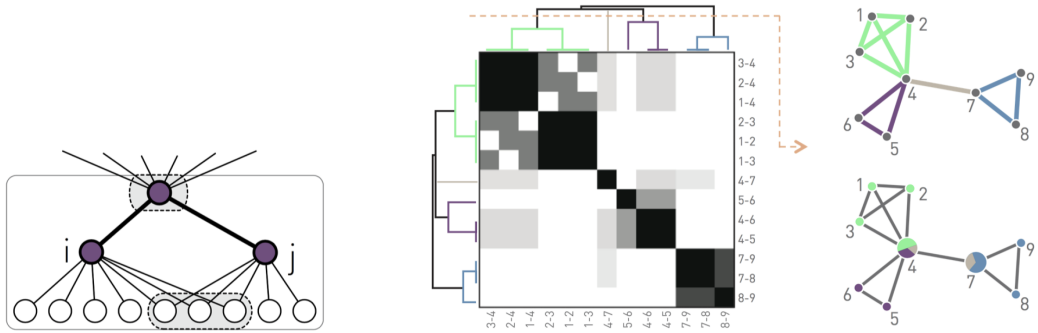
- define *link similarity* as *neighbors index*

$$\forall ij \in \Gamma_x : s_{ij}^x = \frac{|\Gamma_i^+ \cap \Gamma_j^+|}{|\Gamma_i^+ \cup \Gamma_j^+|}$$

- define *cluster similarity* as *single linkage*

$$S_{ij} = \max_{xy \in \Gamma_z} (s_{xy}^z \delta_{c_{xz} c_i} \delta_{c_{yz} c_j})$$

1. bottom-up *agglomerative hierarchical clustering* $\mathcal{O}(m^2)$
2. cut *cluster dendrogram* at desired *clustering resolution*



see linkcomm implementation

community *measures*

— *degree* K, *expansion* E and *Flake* F [FLG00, RCC+04] of $\{C\}$

$$K = \frac{1}{n}\sum_{ij} A_{ij}\delta_{c_i c_j} = \langle k \rangle - E \qquad F = \frac{|\{i{:}\sum_j A_{ij}\delta_{c_i c_j} < k_i/2\}|}{n}$$

— *normalized mutual information* NMI [DDGDA05] of $\{C\}$,$\{D\}$

- p_c & p_{cd} are *standard* & *joint distributions* of $\{C\}$,$\{D\}$
- $H(\mathcal{C})$ & $H(\mathcal{C}|\mathcal{D})$ are *standard* & *conditional entropies*
- MI & VI are *mutual* & *variation* of *information*

$$NMI = \frac{2MI(\mathcal{C},\mathcal{D})}{H(\mathcal{C})+H(\mathcal{D})} = \frac{2H(\mathcal{C})-2H(\mathcal{C}|\mathcal{D})}{H(\mathcal{C})+H(\mathcal{D})} = \frac{2H(\mathcal{C})+2\sum_{CD} p_{cd}\log\frac{p_{cd}}{p_d}}{-\sum_C p_c \log p_c + H(\mathcal{D})}$$

— *normalized variation of information* NVI [Mei07, KLN08]

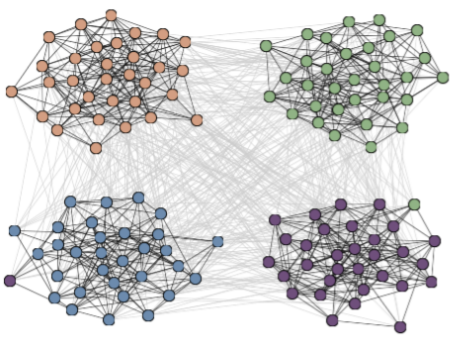
$$NVI = \frac{VI(\mathcal{C},\mathcal{D})}{\log n} = \frac{H(\mathcal{C}|\mathcal{D})+H(\mathcal{D}|\mathcal{C})}{\log n}$$

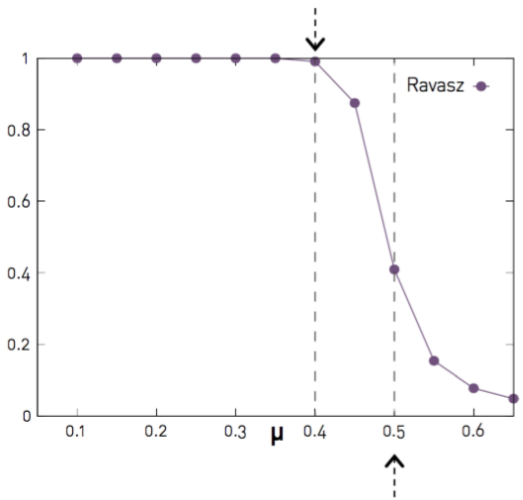
community *benchmarks*

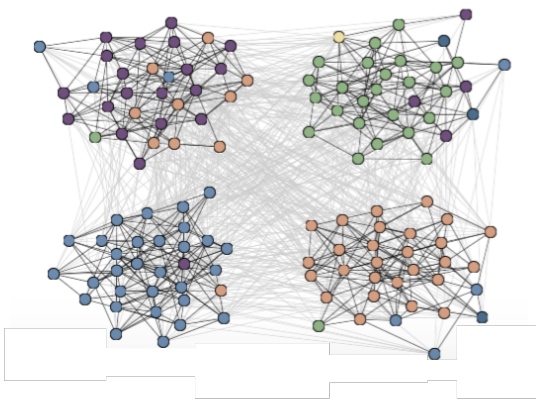
— Girvan-Newman *synthetic graphs* [GN02]

— *planted partition* controlled by *mixing parameter* μ

$n = 128$ $\quad \langle k \rangle = \langle k^{int} \rangle + \langle k^{ext} \rangle = 16$ $\quad \mu = \frac{\langle k^{ext} \rangle}{\langle k \rangle}$







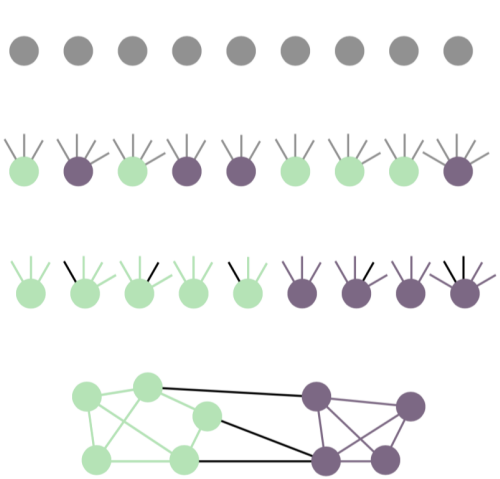
community *benchmarks*

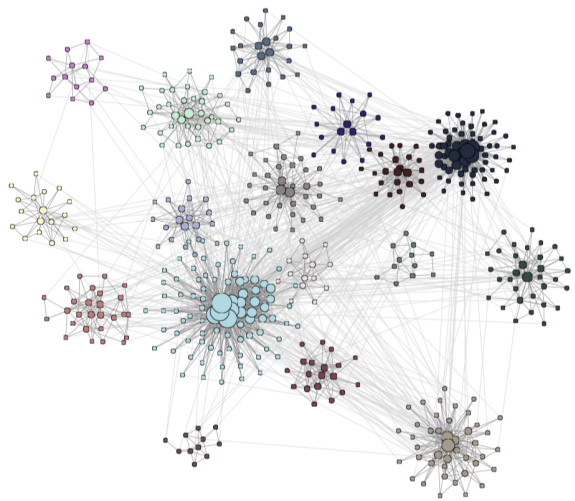
— Lanchichinetti *synthetic graphs* [LFR08]

— *power-law distributions* $p_k \sim k^{-\gamma_k}$ & $p_s \sim s^{-\gamma_s}$

— *planted communities* controlled by *mixing parameter* μ

$n = 1000,\ n_c \in [10, 50]$ $\gamma_k \in [2, 3],\ \gamma_s \in [1, 2]$ $\mu = \frac{\langle k^{ext} \rangle}{\langle k \rangle}$





clustering *references*

Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann.
Link communities reveal multiscale complexity in networks.
Nature, 466(7307):761–764, 2010.

A.-L. Barabási.
Network Science.
Cambridge University Press, Cambridge, 2016.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre.
Fast unfolding of communities in large networks.
J. Stat. Mech., P10008, 2008.

Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj.
Generalized Blockmodeling.
Cambridge University Press, Cambridge, 2005.

Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas.
Comparing community structure identification.
J. Stat. Mech., page P09008, 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.

T. S. Evans and R. Lambiotte.
Line graphs, link partitions and overlapping communities.
Phys. Rev. E, 80(1):016105, 2009.

clustering *references*

M. Fiedler.

Algebraic connectivity of graphs.
Czech. Math. J., 23:298–305, 1973.

Gary William Flake, Steve Lawrence, and C. Lee Giles.

Efficient identification of web communities.
In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, USA, 2000.

M. Girvan and M. E. J Newman.

Community structure in social and biological networks.
P. Natl. Acad. Sci. USA, 99(12):7821–7826, 2002.

Jussi M. Kumpula, Mikko Kivelä, Kimmo Kaski, and Jari Saramäki.

Sequential algorithm for fast clique percolation.
Phys. Rev. E, 78(2):026109, 2008.

Brian W. Kernighan and S. Lin.

An efficient heuristic procedure for partitioning graphs.
Bell Sys. Tech. J., 49(2):291–308, 1970.

Brian Karrer, Elizaveta Levina, and M. E. J. Newman.

Robustness of community structure in networks.
Phys. Rev. E, 77(4):046119, 2008.

Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi.

Benchmark graphs for testing community detection algorithms.
Phys. Rev. E, 78(4):046110, 2008.

F. Lorrain and H. C. White.

Structural equivalence of individuals in social networks.
J. Math. Sociol., 1(1):49–80, 1971.

clustering *references*

Marina Meila.

Comparing clusterings: An information based distance.
J. Multivariate Anal., 98(5):873–895, 2007.

Mark E. J. Newman.

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

Gergely Palla, Imre Derényi, Illes Farkas, and Tamas Vicsek.

Uncovering the overlapping community structure of complex networks in nature and society.
Nature, 435(7043):814–818, 2005.

Tiago P. Peixoto.

Model selection and hypothesis testing for large-scale network models with overlapping groups.
Phys. Rev. X, 5(1):011033, 2015.

Usha Nandini Raghavan, Reka Albert, and Soundar Kumara.

Near linear time algorithm to detect community structures in large-scale networks.
Phys. Rev. E, 76(3):036106, 2007.

M. Rosvall and C. T. Bergstrom.

Maps of random walks on complex networks reveal community structure.
P. Natl. Acad. Sci. USA, 105(4):1118–1123, 2008.

Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi.

Defining and identifying communities in networks.
P. Natl. Acad. Sci. USA, 101(9):2658–2663, 2004.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and Albert László Barabási.

Hierarchical organization of modularity in metabolic networks.
Science, 297(5586):1551–1555, 2002.

clustering *references*

Lovro Šubelj and Marko Bajec.

Robust network community detection using balanced propagation.
Eur. Phys. J. B, 81(3):353–362, 2011.

D. R. White and K. P. Reitz.

Graph and semigroup homomorphisms on networks of relations.
Soc. Networks, 5(2):193–234, 1983.

course *project*

advanced topics in *network science* (*ants*)

Lovro Šubelj & Jure Leskovec
University of Ljubljana
spring 2019/20

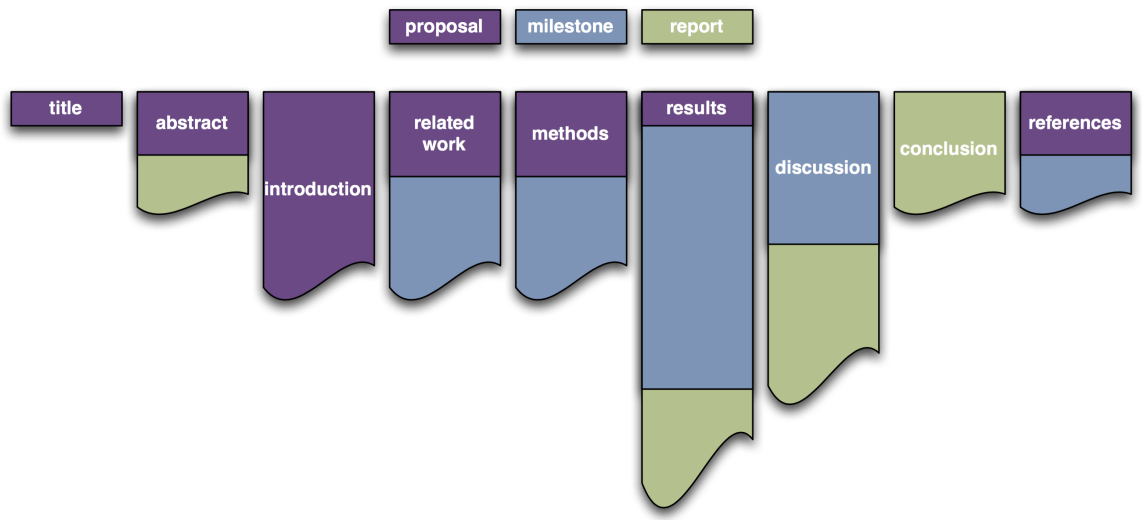
project *overview*

- substantial *network science project*
- networks *combined with PhD research*
- *scientific paper submitted* (to arXiv.org)
- project should go *beyond this course*

- *analytical derivation* of *theoretical results*
- *empirical evaluation* of *methods* or *models*
- *design of novel methods*, *models* or *algorithms*
- *scalable implementation* of existing *algorithms*

project *delivery*

- course *project delivery* breakdown
  - *reaction paper* and *project proposal*
  - *midterm project milestone report*
  - final *eight-page scientific paper*
- informal *project presentations* to get *feedback*
  - *proposal/milestone presentations* in front of *class*
  - *final presentation* in front of *INA students* (*seminar*)



project *proposal*

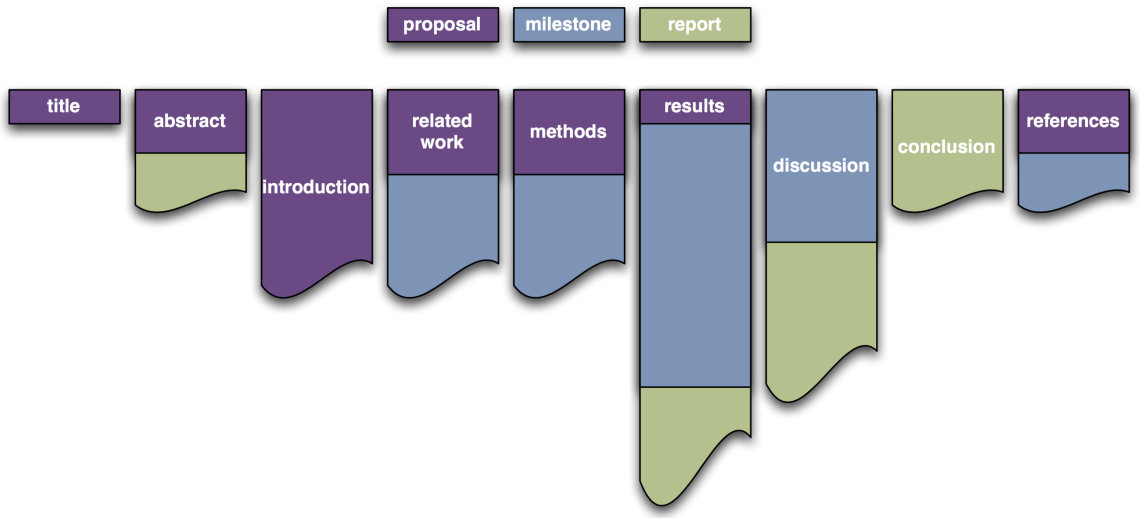
- *cover* component ($\leq$ *paragraph*)
  - project *title* and short *abstract*
- *introductory* component ($\leq$ *page*)
  - project *motivation* and *background*
  - *detailed* project *problem description*
- *reaction paper* component ($>$ *page*)
  - survey of existing *related work*
  - *critical evaluation* of existing work
  - highlight *strengths* and *weaknesses*
- *project proposal* component ($\leq$ *page*)
  - *detailed* project *proposal description*
  - proposed *methods* and expected *results*
  - highlight *value*, *novelty* and *feasibility*

project *milestone*

— *midterm project milestone*

— *at least* 50% of *research done*

— *at least* 50% of final *paper written*



project *deadlines*

- *hard-copy* in *submission box* (grading)
- *electronic version* to eUcilnica (archive)
- *cover sheet* with signed *honor code*

week	lectures	presentations	assignments
. . .	. . .	. . .	. . .
7		*proposal* (Apr 2nd)	
8			*proposal* (Apr 9th)
9			
10			
11	*advanced topics in network science*	*milestone* (Apr 30th)	
12			*milestone* (May 7th)
13			
14			
15	*applications of network science*	*project* (May 28th)	
16			*paper* (Jun 4th)
. . .			

— *assignments* due on *Thursdays at 3:00pm*

— *late days* expire on *Tuesdays at 3:00pm*

— *presentations* on *Thursdays at 6:00pm*