NATIONAL RESEARCH UNIVERSITY

Network communities

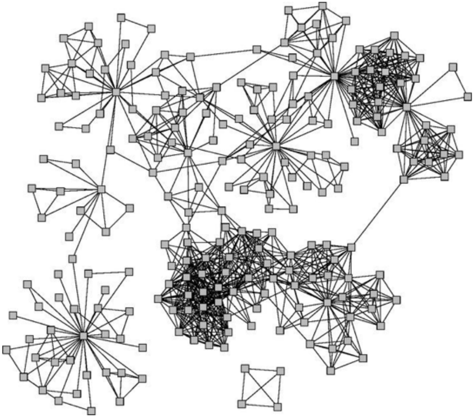
Social Network Analysis. MAGoLEGO course. Lecture 5

Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2019/sna

National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

Network communities



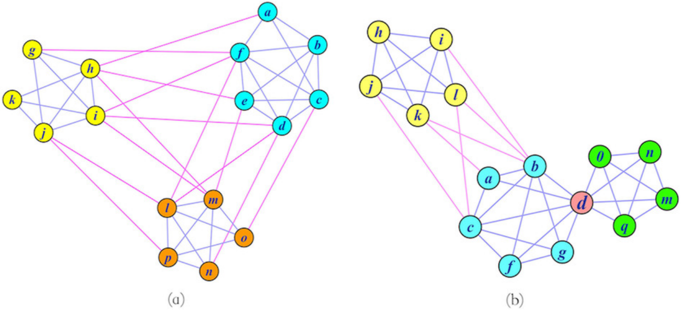
Connected and undirected graphs

Network communities

What makes a community (cohesive subgroup):

- Mutuality of ties. Everyone in the group has ties (edges) to one another
- Compactness. Closeness or reachability of group members in small number of steps, not necessarily adjacency
- Density of edges. High frequency of ties within the group
- Separation. Higher frequency of ties among group members compared to non-members

Wasserman and Faust



Community types:

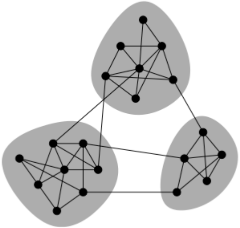
- Non-overlapping
- Overlapping

image from W. Liu , 2014

Network communities

Definition

Network communities are groups of vertices such that vertices inside the group connected with many more edges than between groups.



- Will consider non-overlapping communities, each node assigned only to one community

Community density

- Graph $G(V, E)$, $n = |V|$, $m = |E|$
- Community - set of nodes S
  n_s-number of nodes in S, m_s - number of edges in S
- Graph density

$$\rho = \frac{m}{n(n-1)/2}$$

- community internal density

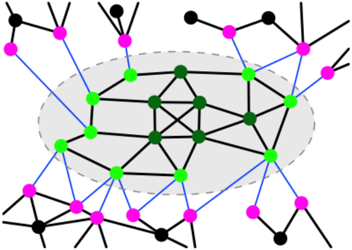
$$\delta_{int}(C) = \frac{m_s}{n_s(n_s-1)/2}$$

- external edges density

$$\delta_{ext}(C) = \frac{m_{ext}}{n_c(n-n_c)}$$

- community (cluster): $\delta_{int} > \rho$, $\delta_{ext} < \rho$

Graph cuts



Graph cut is a partition of the vertices of a graph $G(E, V)$ into two disjoint subsets: $V = V_1 + V_2$

$$Q = cut(V_1, V_2) = \sum_{i \in V_1, j \in V_2} e_{ij}$$

image from Fortunato 2016

Modularity

- Compare fraction of edges within the cluster to expected fraction in random graph with identical degree sequence

$$Q = \frac{1}{4}(m_s - E(m_s))$$

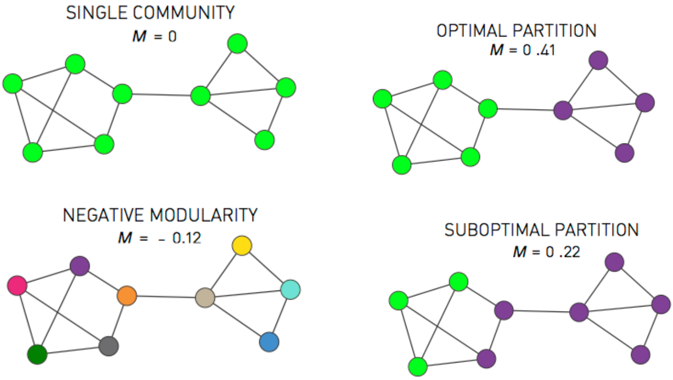
- Modularity score

$$Q = \sum_u \left(\frac{m_u}{m} - \left(\frac{k_u}{2m} \right)^2 \right)$$

m_u - number of internal edges in a community u,
k_u - sum of node degrees within a community

- Modularity score range $Q \in [-1/2, 1)$, single community $Q = 0$

Modularity



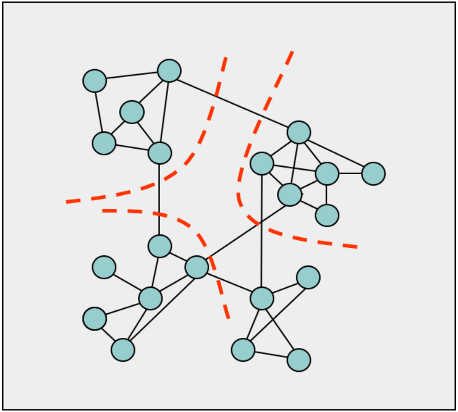
- The higher the modularity score - the better are communities

image from A.L. Barabasi 2016

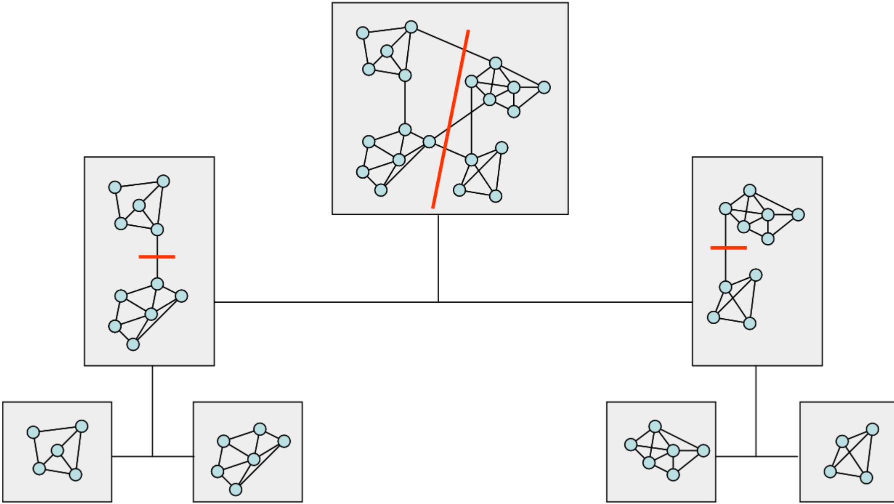
Community detection

- Combinatorial optimization problem:
  - optimization criterion (density, graph cut, modularity score)
  - optimization method
- Exact solution NP-hard
  (bi-partition: $n = n_1 + n_2,\ n!/(n_1!n_2!)$ combinations)
- Solved by greedy, approximate algorithms or heuristics
- Recursive top-down 2-way partition, multiway partition

Multiway partitioning



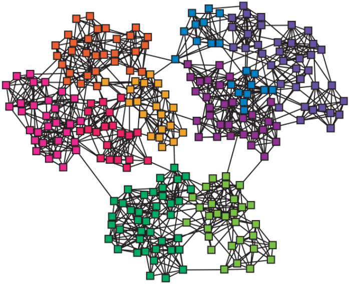
Recursive partitioning



Edge betweenness

Focus on edges that connect communities.
Edge betweenness -number of shortest paths $\sigma_{st}(e)$ going through edge e

$$C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$$



Edge betweenness algorithm

Newman-Girvan, 2004

Algorithm: Edge Betweenness

Input: graph G(V,E)

Output: Dendrogram/communities

repeat

For all $e \in E$ compute edge betweenness $C_B(e)$;

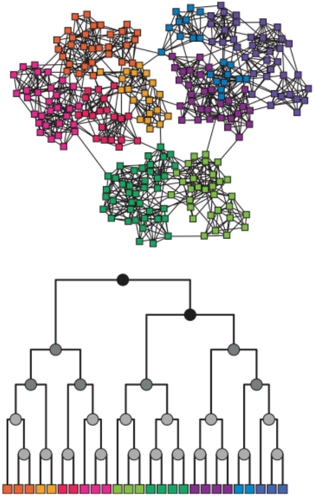
remove edge e_i with largest $C_B(e_i)$;

until *edges left*;

If bi-partition, then stop when graph splits in two components (check for connectedness)

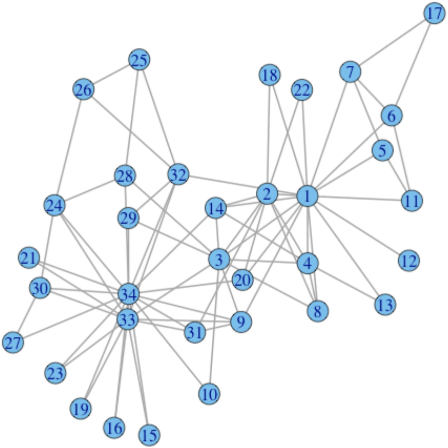
Edge betweenness

Hierarchical algorithm, dendrogram



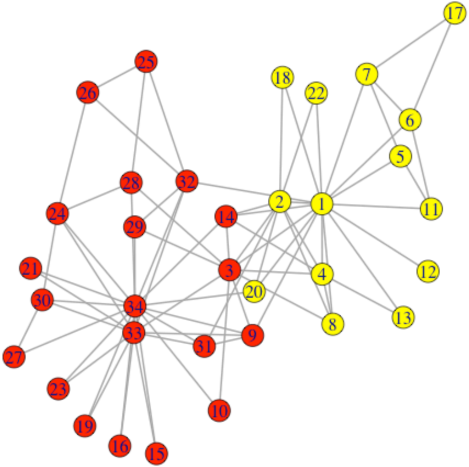
Edge betweenness

Zachary karate club



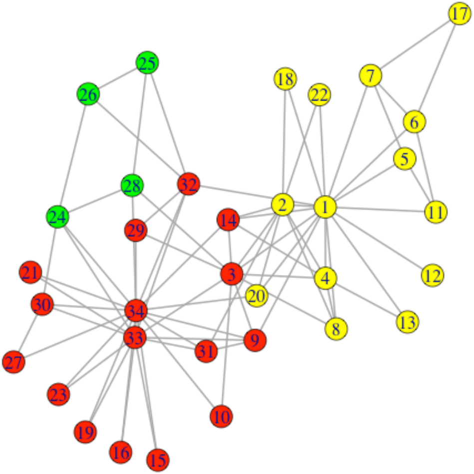
Edge betweenness

Zachary karate club

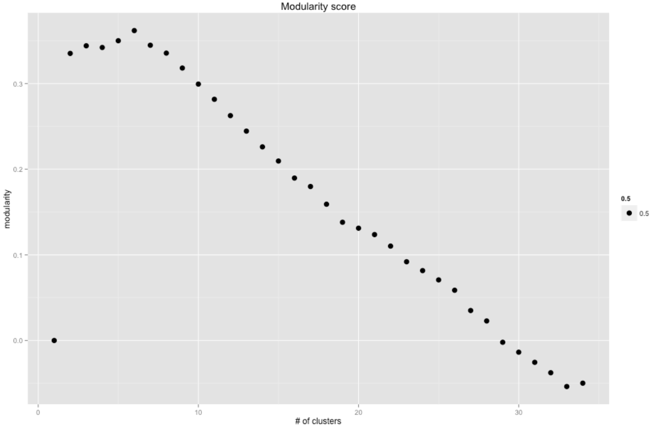


Edge betweenness

Zachary karate club



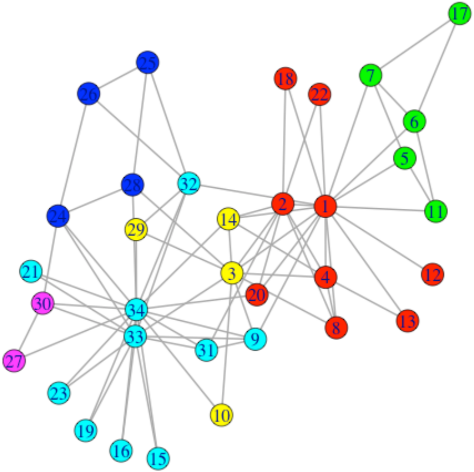
Edge betweenness



igraph:modularity()

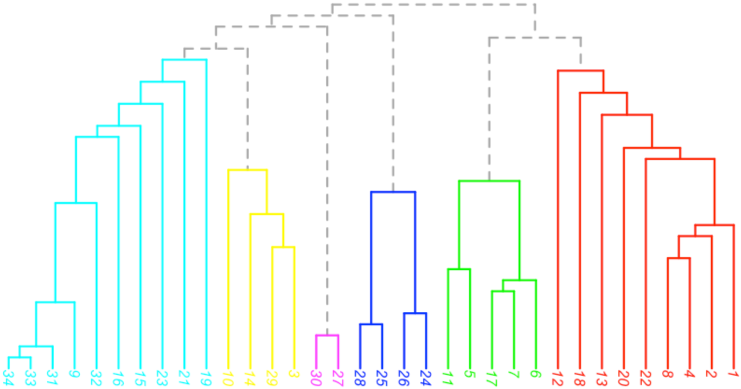
Edge betweenness

best: clusters = 6, modularity = 0.345



Edge betweenness

Zachary karate club



igraph:dendPlot()

Fast community unfolding

V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, 2008 ”The Louvain method”

- Heuristic method for greedy modularity optimization
- Find partitions with high modularity
- Multi-level (multi-resolution) hierarchical scheme
- Scalable

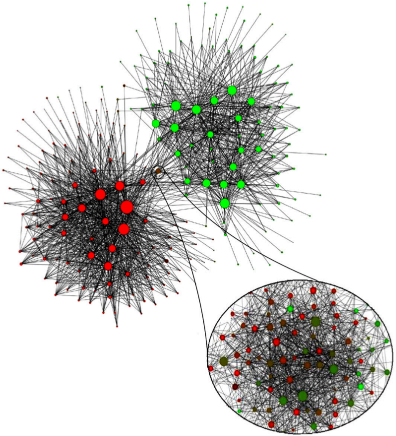
Modularity:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j) = \sum_u\left(\frac{m_u}{m} - \left(\frac{k_u}{2m}\right)^2\right)$$

V. Blondel et.al., 2008

Fast community unfolding

Multi-resolution scalable method



2 mln mobile phone network

V. Blondel et.al., 2008

Fast community unfolding algorithm

Input: Graph G(V,E)

Output: Communities

Assign every node to its own community;

repeat

  repeat

    For every node evaluate modularity gain from removing node from its community and placing it in the community of its neighbor;

    Place node in the community maximizing modularity gain;

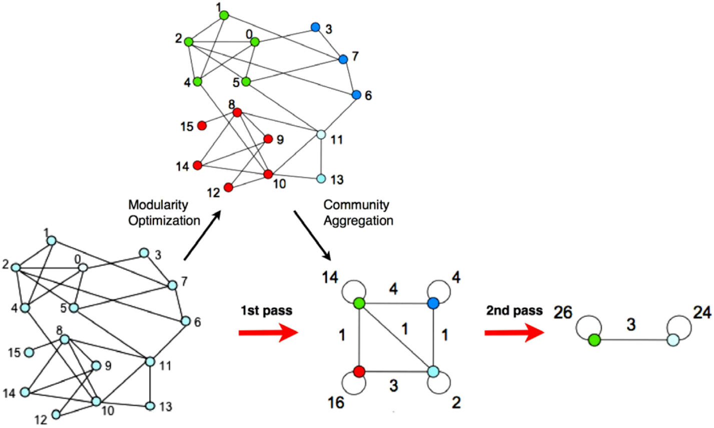
  until *no more improvement (local max of modularity)*;

  Nodes from communities merged into "super nodes" ;

  Weight on the links added up

until *no more changes (max modularity)*;

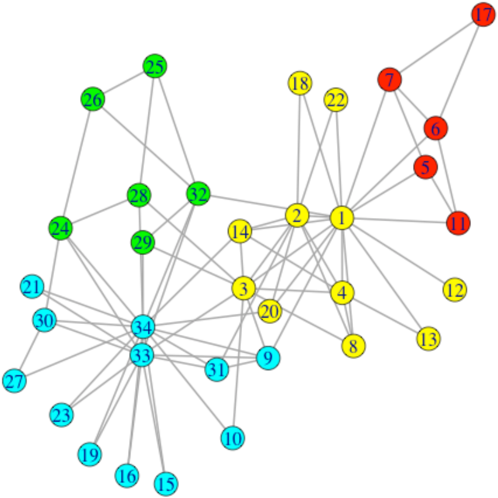
Fast community unfolding



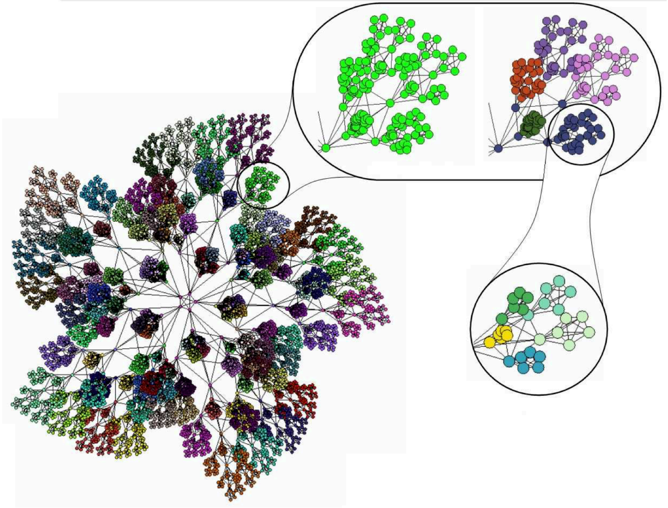
V. Blondel et.al., 2008

Fast community unfolding

clusters = 4, modularity = 0.445



Fast community unfolding



V. Blondel et.al., 2008

Community detection algorithms

Author	Ref.	Label	Order
Eckmann & Moses	(Eckmann and Moses, 2002)	EM	$O(m\langle k^2\rangle)$
Zhou & Lipowsky	(Zhou and Lipowsky, 2004)	ZL	$O(n^3)$
Latapy & Pons	(Latapy and Pons, 2005)	LP	$O(n^3)$
Clauset et al.	(Clauset *et al.*, 2004)	NF	$O(n\log^2 n)$
Newman & Girvan	(Newman and Girvan, 2004)	NG	$O(nm^2)$
Girvan & Newman	(Girvan and Newman, 2002)	GN	$O(n^2m)$
Guimerà et al.	(Guimerà and Amaral, 2005; Guimerà *et al.*, 2004)	SA	parameter dependent
Duch & Arenas	(Duch and Arenas, 2005)	DA	$O(n^2\log n)$
Fortunato et al.	(Fortunato *et al.*, 2004)	FLM	$O(m^3n)$
Radicchi et al.	(Radicchi *et al.*, 2004)	RCCLP	$O(m^4/n^2)$
Donetti & Muñoz	(Donetti and Muñoz, 2004, 2005)	DM/DMN	$O(n^3)$
Bagrow & Bollt	(Bagrow and Bollt, 2005)	BB	$O(n^3)$
Capocci et al.	(Capocci *et al.*, 2005)	CSCC	$O(n^2)$
Wu & Huberman	(Wu and Huberman, 2004)	WH	$O(n+m)$
Palla et al.	(Palla *et al.*, 2005)	PK	$O(\exp(n))$
Reichardt & Bornholdt	(Reichardt and Bornholdt, 2004)	RB	parameter dependent

Author	Ref.	Label	Order
Girvan & Newman	(Girvan and Newman, 2002; Newman and Girvan, 2004)	GN	$O(nm^2)$
Clauset et al.	(Clauset *et al.*, 2004)	Clauset et al.	$O(n\log^2 n)$
Blondel et al.	(Blondel *et al.*, 2008)	Blondel et al.	$O(m)$
Guimerà et al.	(Guimerà and Amaral, 2005; Guimerà *et al.*, 2004)	Sim. Ann.	parameter dependent
Radicchi et al.	(Radicchi *et al.*, 2004)	Radicchi et al.	$O(m^4/n^2)$
Palla et al.	(Palla *et al.*, 2005)	Cfinder	$O(\exp(n))$
Van Dongen	(Dongen, 2000a)	MCL	$O(nk^2)$, $k<n$ parameter
Rosvall & Bergstrom	(Rosvall and Bergstrom, 2007)	Infomod	parameter dependent
Rosvall & Bergstrom	(Rosvall and Bergstrom, 2008)	Infomap	$O(m)$
Donetti & Muñoz	(Donetti and Muñoz, 2004, 2005)	DM	$O(n^3)$
Newman & Leicht	(Newman and Leicht, 2007)	EM	parameter dependent
Ronhovde & Nussinov	(Ronhovde and Nussinov, 2009)	RN	$O(m^\beta\log n)$, $\beta\sim 1.3$

References

- S. Fortunato. Community detection in graphs, Physics Reports, Vol. 486, Iss. 3–5, pp 75-174, 2010
- S. E. Schaeffer. Graph clustering. Computer Science Review, 1(1):27–64, 2007.
- Modularity and community structure in networks, M.E.J. Newman, PNAS, vol 103, no 26, pp 8577-8582, 2006
- Finding and evaluating community structure in networks, M.E.J. Newman, M. Girvan, Phys. Rev E, 69, 2004
- U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (3) (2007) 036106.
- G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814?818.
- P. Pons and M. Latapy, Computing communities in large networks using random walks, Journal of Graph Algorithms and Applications, 10 (2006), 191-218.
- V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. P10008 (2008).