Community detection

Leonid E. Zhukov

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

Network Science

Lecture outline

1. Overlapping commmunities
   - Clique percolation method
2. Multi-level optimization
   - Fast community unfolding
3. Random walk methods
   - Walktrap

Community detection

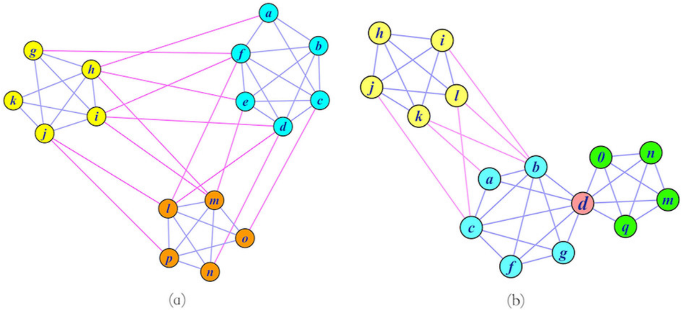
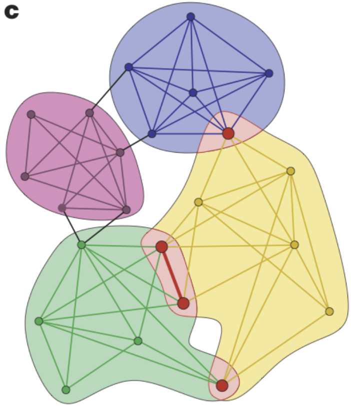


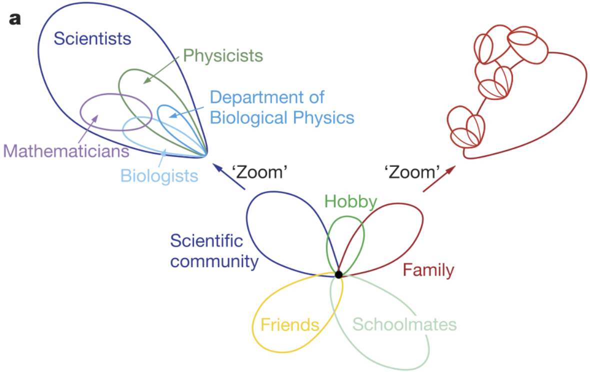
image from W. Liu , 2014

Overlapping communities



Palla, 2005

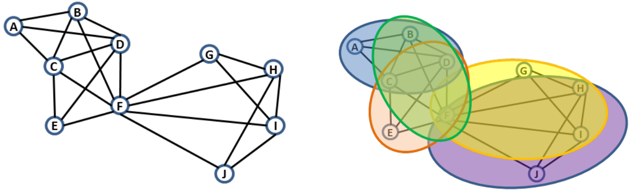
Overlapping communities



Palla, 2005

k-clique community

- k-clique is a clique (complete subgraph) with k nodes
- k-clique community a union of all k-cliques that can be reached from each other through a series of adjacent k-cliques
- two k-cliques are said to be adjacent if they share $k-1$ nodes.



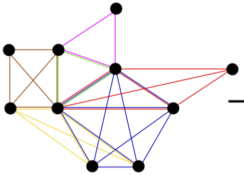
Adjacent 4-cliques

k-clique percolation

- Find all maximal cliques
- Create clique overlap matrix
- Threshold matrix at value $k-1$
- Communities = connected components

Palla, 2005

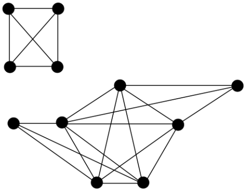
k-clique percolation



	5	3	2	1	3	1
	3	4	2	1	1	1
	2	2	3	2	1	2
	1	1	2	3	0	1
	3	1	1	0	4	2
	1	1	2	1	2	4

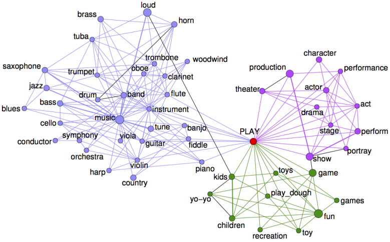
k=4

	1	1	0	0	1	0
	1	1	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	1	0	0	0	1	0
	0	0	0	0	0	1

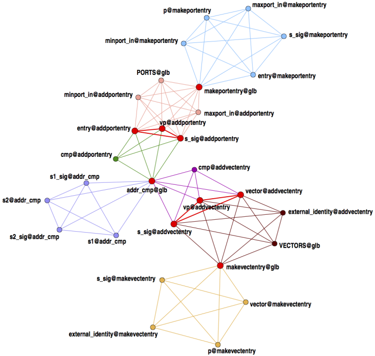


Palla, 2005

k-clique percolation



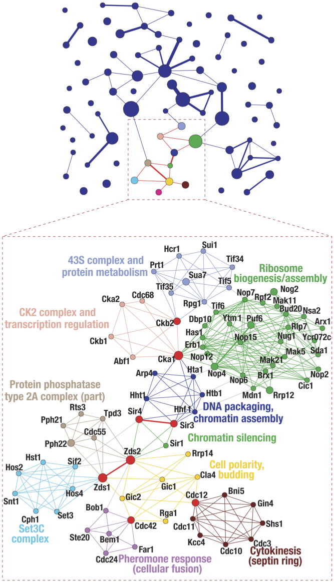
$k = 4$



$k = 5$

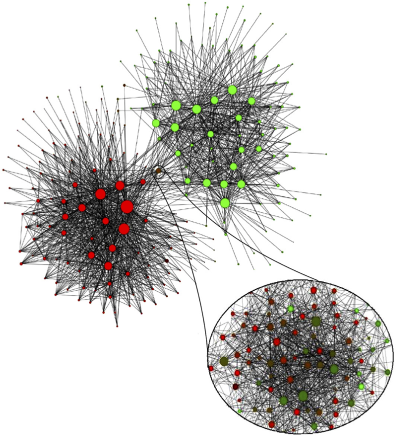
Palla, 2005

k-clique percolation



Fast community unfolding

Multi-resolution scalable method



2 mln mobile phone network
V. Blondel et.al., 2008

Fast community unfolding

"The Louvain method"

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

Modularity:

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

V. Blondel et.al., 2008

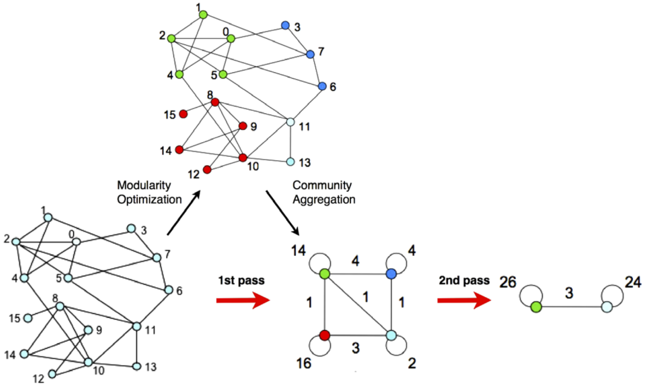
Fast community unfolding

Algorithm

- Assign every node to its own community
- Phase I
  - 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
  - repeat until no more improvement (local max of modularity)
- Phase II
  - Nodes from communities merged into "super nodes"
  - Weight on the links added up
- Repeat until no more changes (max modularity)

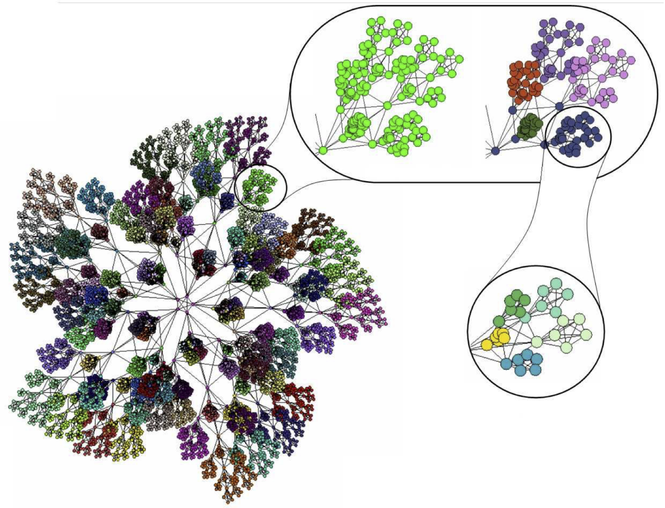
V. Blondel et.al., 2008

Fast community unfolding



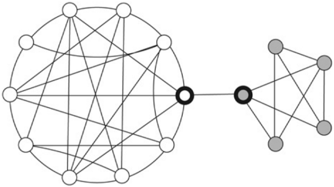
V. Blondel et.al., 2008

Fast community unfolding



V. Blondel et.al., 2008

Communities and random walks



- Random walks on a graph tend to get trapped into densely connected parts corresponding to communities.

Walktrap community

Walktrap

- Consider random walk on graph
- At each time step walk moves to NN uniformly at random $P_{ij} = \frac{A_{ij}}{d(i)}$, $P = D^{-1}A$, $D_{ii} = diag(d(i))$
- P_{ij}^t - probability to get from i to j in t steps, $t \ll t_{mixing}$
- Assumptions: for two i and j in the same community P_{ij}^t is high
- if i and j are in the same community, then $\forall k$, $P_{ik}^t \approx P_{jk}^t$
- Distance between nodes:

$$r_{ij}(t) = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik}^t - P_{jk}^t)^2}{d(k)}} = ||D^{-1/2}P_i^t - D^{-1/2}P_j^t||$$

P. Pons and M. Latapy, 2006

Walktrap

Computing node distance r_{ij}

- Direct (exact) computation: $P^t_{ij} = (P^t)_{ij}$ or $P^t_i = P^t p^0_i$, $p^0_i(k) = \delta_{ik}$
- Approximate computation (simulation):
  - Compute K random walks of length t starting form node i
  - Approximate $P^t_{ik} \approx \frac{N_{ik}}{K}$, number of walks end up on k

Distance between communities:

$$P^t_{Cj} = \frac{1}{|C|} \sum_{i \in C} P^t_{ij}$$

$$r_{C_1 C_2}(t) = \sqrt{\sum_{k=1}^{n} \frac{(P^t_{C_1 k} - P^t_{C_2 k})^2}{d(k)}} = ||D^{-1/2} P^t_{C_1} - D^{-1/2} P^t_{C_2}||$$

P. Pons and M. Latapy, 2006

Walktrap

Algorithm (hierarchical clustering)

- Assign each vertex to its own community $S_1 = \{\{v\}, v \in V\}$
- Compute distance between all adjacent communities $r_{C_i C_j}$
- Choose two "closest" communities that minimizes (Ward's methods):

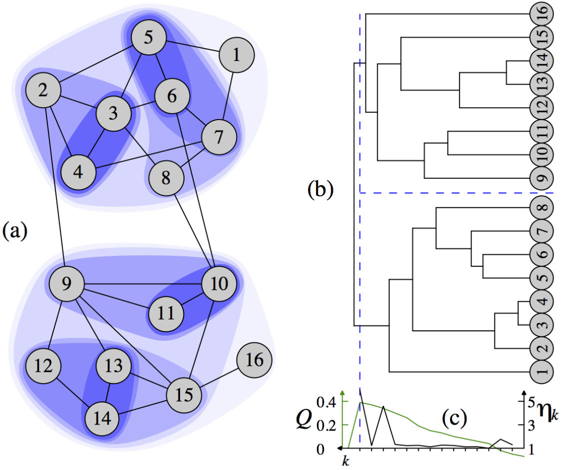
$$\Delta\sigma(C_1, C_2) = \frac{1}{n}\left(\sum_{i \in C_3} r_{iC_3}^2 - \sum_{i \in C_1} r_{iC_1}^2 - \sum_{i \in C_2} r_{iC_2}^2\right)$$

  and merge them $S_{k+1} = (S_k \backslash \{C_1, C_2\}) \cup C_3,\ C_3 = C_1 \cup C_2$
- update distance between communities

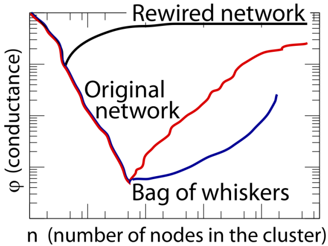
After $n-1$ steps finish with one community $S_n = \{V\}$

P. Pons and M. Latapy, 2006

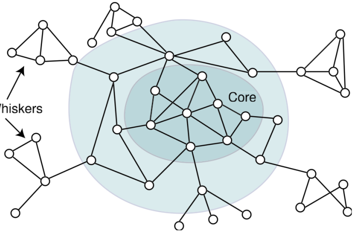
Walktrap



P. Pons and M. Latapy, 2006



(a) Typical NCP plot



(b) Caricature of network structure

Best conductance of a vertex set S of size k:

$$\Phi(k) = \min_{S \in V, |S|=k} \phi(S), \quad \phi(S) = \frac{cut(S, V \backslash S)}{\min(vol(S), vol(S \backslash V))}$$

where $vol(S) = \sum_{i \in S} k_i$ - sum of all node degrees in the set

J. Leskovec, K. Lang, 2010

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	$O(n^2 m)$
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	$O(m^3 n)$
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	$O(m^\beta \log n)$, $\beta \sim 1.3$

Fortunato, 2010

References

- 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).
- J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. In WWW 08: Procs. of the 17th Int. Conf. on World Wide Web, pages 695-704, 2008.

References

- M.A Porter, J-P Onella, P.J. Mucha. Communities in Networks, Notices of the American Mathematical Society, Vol. 56, No. 9, 2009
- S. E. Schaeffer. Graph clustering. Computer Science Review, 1(1), pp 27-64, 2007.
- S. Fortunato. Community detection in graphs, Physics Reports, Vol. 486, Iss. 3-5, pp 75-174, 2010

Summary

Lectures 1-10

- Network characteristics:
  – Power law node degree distribution
  – Small diameter
  – High clustering coefficient (transitivity)
- Network models:
  – Random graphs
  – Preferential attachement
  – Small world
- Centrality measures:
  – Degree centrality
  – Closeness centrality
  – Betweenness centrality
- Link analysis:
  – Page rank
  – HITS

Summary

Lectures 1-10

- Structural equivalence
  - Vertex equivalence
  - Vertex similarity
- Assortative mixing
  - Assortative and disassortative networks
  - Mixing by node degree
  - Modularity
- Network structures:
  - Cliques
  - k-cores
- Network communities:
  - Graph partitioning
  - Overlapping communities
  - Heuristic methods
  - Random walk based methods