NETWORK GROUP DISCOVERY BY LABEL PROPAGATION

Lovro Šubelj

University of Ljubljana,
Faculty of Computer and Information Science

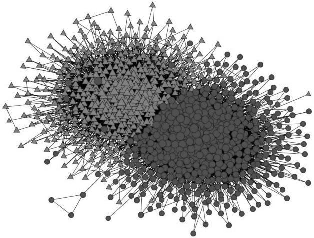
9.7.2014

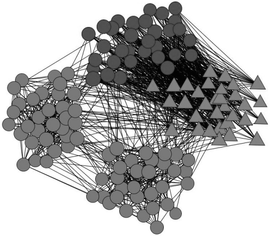
NETWORK GROUPS

community densely linked nodes sparsely linked between (Girvan and Newman, 2002)

module nodes linked to similar other nodes (Newman and Leicht, 2007)

other mixtures of these



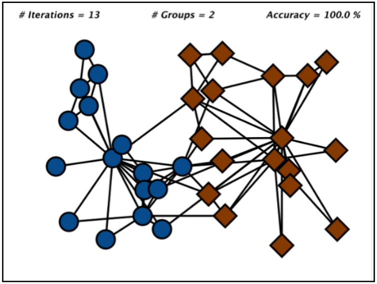


LABEL PROPAGATION

Label propagation algorithm: (Raghavan et al., 2007)

$$g_i = \operatorname*{argmax}_{g} \sum_{j \in \Gamma_i} \delta(g_j, g)$$

g_i is group label of node i and Γ_i are its neighbors.



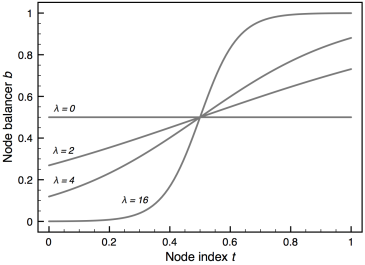
Algorithm has near linear complexity $\mathcal{O}(m)$, where m is number of links.

BALANCED PROPAGATION

Balanced propagation algorithm: (Šubelj and Bajec, 2011a)

$$g_i = \operatorname*{argmax}_g \sum_{j \in \Gamma_i} b_j \cdot \delta(g_j, g) \qquad b_i = \frac{1}{1 + e^{-\lambda(t_i - \frac{1}{2})}}$$

b_i is balancer of node i and $t_i \in (0, 1]$ is its normalized index.



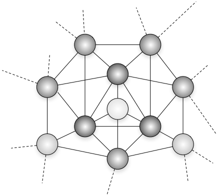
 # Partitions found in Zachary network in 1000 *runs drops from* 184 *to* 19.

ADVANCED PROPAGATION

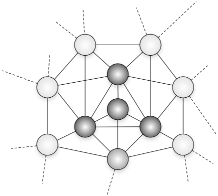
Defensive propagation algorithm: (Šubelj and Bajec, 2011b)

$$g_i = \operatorname*{argmax}_g \sum_{j \in \Gamma_i} p_j b_j \cdot \delta(g_j, g)$$

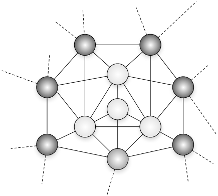
p_i is probability that random walker on group g_i visits node i.



By degrees



Defensive



Offensive

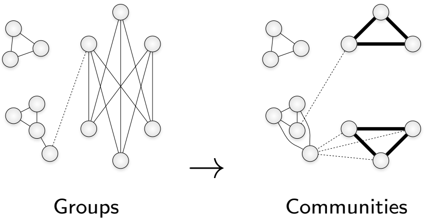
Defensive algorithm has high recall, offensive algorithm has high precision.

GENERAL PROPAGATION

General propagation algorithm: (Šubelj and Bajec, 2012)

$$g_i = \underset{g}{\operatorname{argmax}} \left(\tau_g \cdot \overbrace{\sum_{j \in \Gamma_i} p_j b_j \cdot \delta(g_j, g)}^{\text{Community detection}} + (1 - \tau_g) \cdot \overbrace{\sum_{\substack{j \in \Gamma_i \\ k \in \Gamma_j \setminus \Gamma_i}} \frac{p'_j b_k}{k_j} \cdot \delta(g_k, g)}^{\text{Module detection}} \right)$$

k_i is degree of node i and $\tau_g \in [0, 1]$ is parameter of group g.



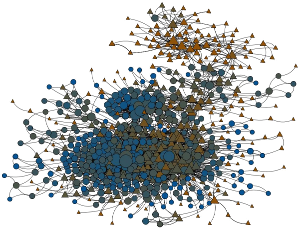
Group parameters τ have to be set accordingly (conductance, clustering).

HIERARCHICAL PROPAGATION

Hierarchical propagation algorithm: (Šubelj and Bajec, 2014)

$$\tau_{g_i} = \begin{cases} 1 & \text{if } d_i \geq p \text{ and } \langle d \rangle \geq p \\ 0 & \text{if } d_i < p \text{ and } \langle d \rangle < p \\ 0.5 & \text{else} \end{cases}$$

d_i is corrected clustering of node i and p is clustering of configuration model.

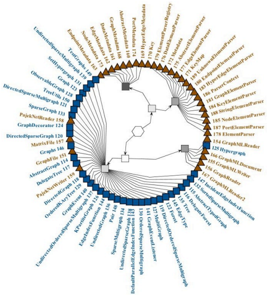


Communities are in dense parts ($d \gg 0$), modules are in sparse parts ($d \approx 0$).

HIERARCHICAL PROPAGATION (II)

Hierarchical propagation algorithm: (Šubelj and Bajec, 2014)

- group detection by propagation → communities
- bottom-up group agglomeration → hierarchy
- top-down group refinement → modules

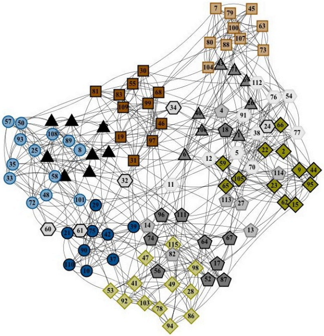


Alternative group hierarchies are compared by maximum likelihood.

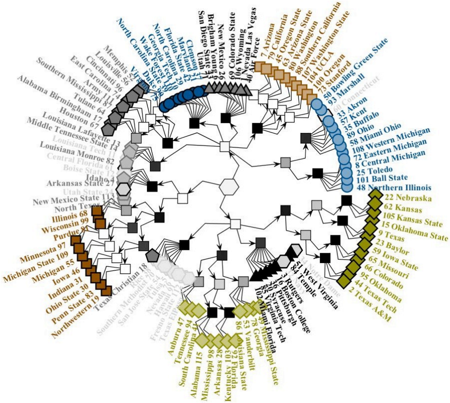
SOCIAL NETWORKS

Node shapes show sociological division into groups, (Girvan and Newman, 2002)

shades of inner nodes of hierarchy are proportional to link density.



American football network

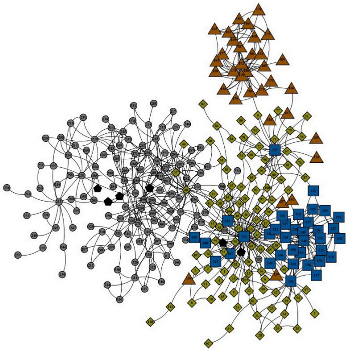


Group hierarchy

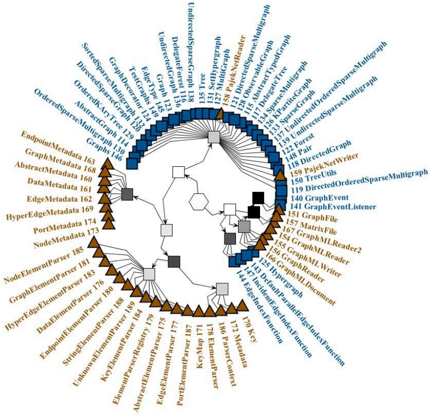
SOFTWARE NETWORKS

Node shapes show developer division into packages, (O'Madadhain et al., 2005)

shades of inner nodes of hierarchy are proportional to link density.



JUNG dependency network



Group hierarchy

REAL-WORLD NETWORKS

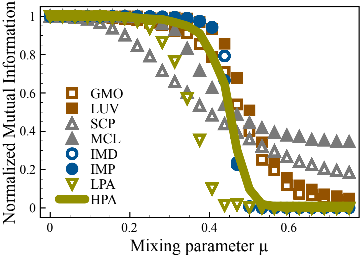
Label propagation algorithm (LPA), multi-stage modularity optimization or Louvain method (LUV), random walk compression or Infomap (IMP), *k*-means data clustering (KMN), mixture model with expectation-maximization (EMM) and hierarchical propagation algorithm (HPA).

	Community detection			Group detection		
	LPA	LUV	IMP	KMN	EMM	HPA
American football network	0.892	0.876	**0.922**	0.845	0.823	0.909
	0.796	0.771	**0.890**	0.698	0.683	0.850
Southern women network	0.184	0.309	0.417	0.677	0.827	**0.932**
	0.093	0.174	0.273	0.560	0.720	**0.936**

Normalized Mutual Information and Adjusted Rand Index

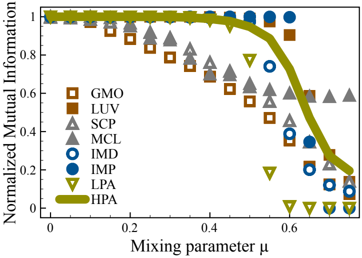
SYNTHETIC NETWORKS

Greedy optimization of modularity (GMO), multi-stage modularity optimization or Louvain (LUV), sequential clique percolation (SCP), Markov clustering (MCL), structural compression or Infomod (IMD), random walk compression or Infomap (IMP), label propagation algorithm (LPA) and hierarchical propagation algorithm (HPA).



4 communities

(Girvan and Newman, 2002)

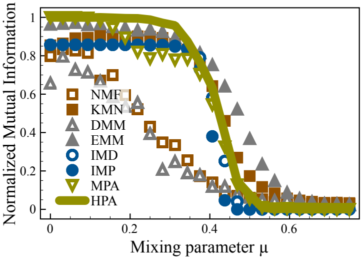


$\geq$ 10 communities

(Lancichinetti et al., 2008)

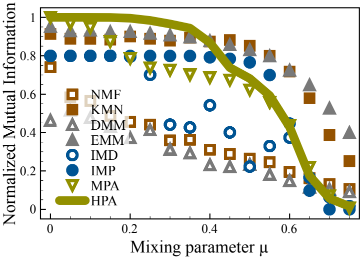
SYNTHETIC NETWORKS (II)

Symmetric nonnegative matrix factorization (NMF), *k*-means data clustering (KMN), (degree-corrected) mixture model (EMM & DMM), structural compression or Infomod (IMD) and random walk compression or Infomap (IMP), model-based propagation algorithm (MPA) and hierarchical propagation algorithm (HPA).



2 communities & bipartite modules

(Šubelj and Bajec, 2012)



3 communities & tripartite modules

(Šubelj and Bajec, 2014)

CONCLUSIONS

Hierarchical propagation algorithm: (Šubelj and Bajec, 2014)

- non-overlapping community and module detection
- easy to implement or extend with domain knowledge
- benefits in group detection, hierarchy discovery, link prediction

Community detection	$\longrightarrow$	CHECK COMMUNITIES	$\longrightarrow$	Module detection
Infomap		corrected clustering		data clustering
(Rosvall and Bergstrom, 2008)		(Soffer and Vázquez, 2005)		(Lin et al., 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.

A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78(4):046110, 2008.

C.-Y. Lin, J.-L. Koh, and A. L. P. Chen. A better strategy of discovering link-pattern based communities by classical clustering methods. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 56–67, Hyderabad, India, 2010.

M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *P. Natl. Acad. Sci. USA*, 104(23):9564, 2007.

J. O'Madadhain, D. Fisher, S. White, P. Smyth, and Y.-B. Boey. Analysis and visualization of network data using JUNG. *J. Stat. Softw.*, 10(2):1–35, 2005.

U. N. Raghavan, R. Albert, and S. 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.

S. N. Soffer and A. Vázquez. Network clustering coefficient without degree-correlation biases. *Phys. Rev. E*, 71(5):057101, 2005.

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

L. Šubelj and M. Bajec. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. *Phys. Rev. E*, 83(3): 036103, 2011b.

L. Šubelj and M. Bajec. Ubiquitousness of link-density and link-pattern communities in real-world networks. *Eur. Phys. J. B*, 85(1):32, 2012.

L. Šubelj and M. Bajec. Group detection in complex networks: An algorithm and comparison of the state of the art. *Physica A*, 397:144–156, 2014.