Unfolding network communities by combining defensive and offensive label propagation

Lovro Šubelj and Marko Bajec

Faculty of Computer and Information Science, University of Ljubljana

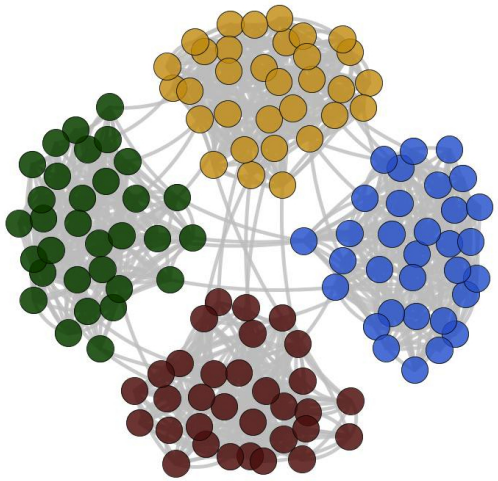
September 20, 2010[1]

[1]Workshop on the Analysis of Complex Networks (ACNE '10)

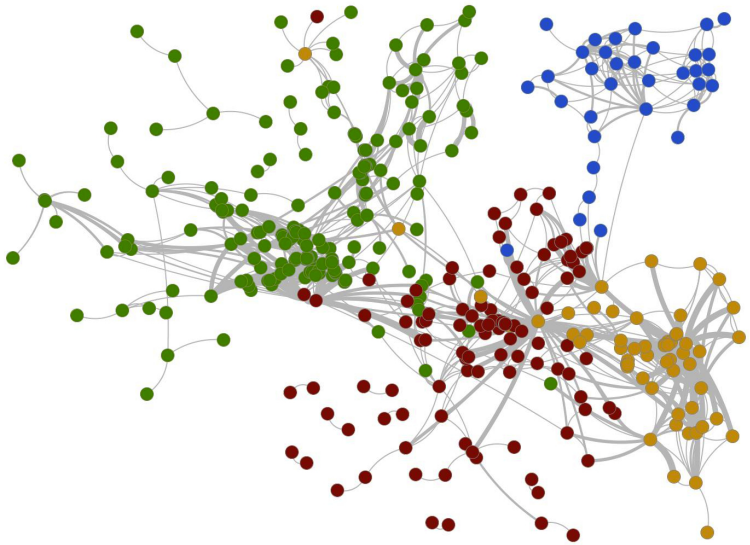
Outline

Network communities

- Intuitively, *communities* (or *modules*) are cohesive groups of nodes densely connected within, and only loosely connected between.
- Formally, e.g., notions of *weak* and *strong communities* [39], etc.



(a) Girvan-Newman [14] benchmark



(b) JUNG graph library

Play an important role in many real-world systems [15, 37].

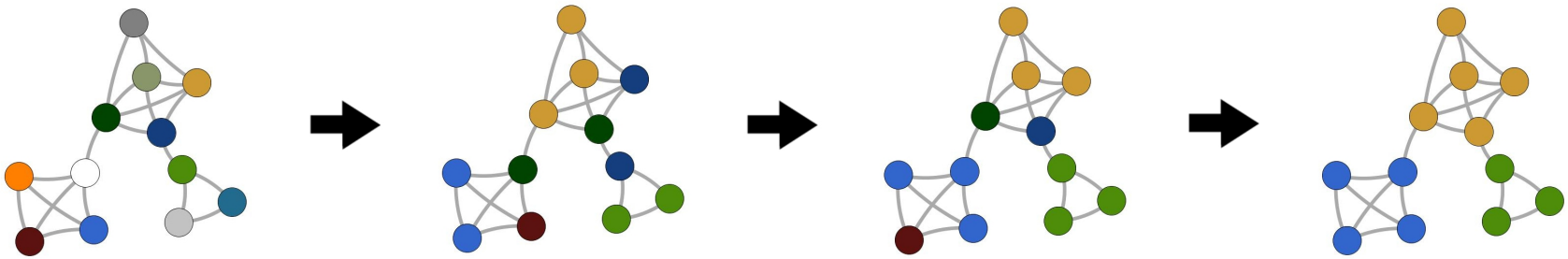
Outline

Label propagation algorithm

Undirected graph $G(N, E)$ with weights W (and communities C).

Label propagation algorithm [40] (*LPA*):

1. initialize nodes with unique labels, i.e., $\forall n \in N : c_n = l_n$,
2. set each node's label to the label shared by most of its neighbors[2], i.e., $\forall n \in N : c_n = \text{argmax}_l \sum_{m \in \mathcal{N}_n^l} w_{nm}$,
3. if not converged, continue to 2.

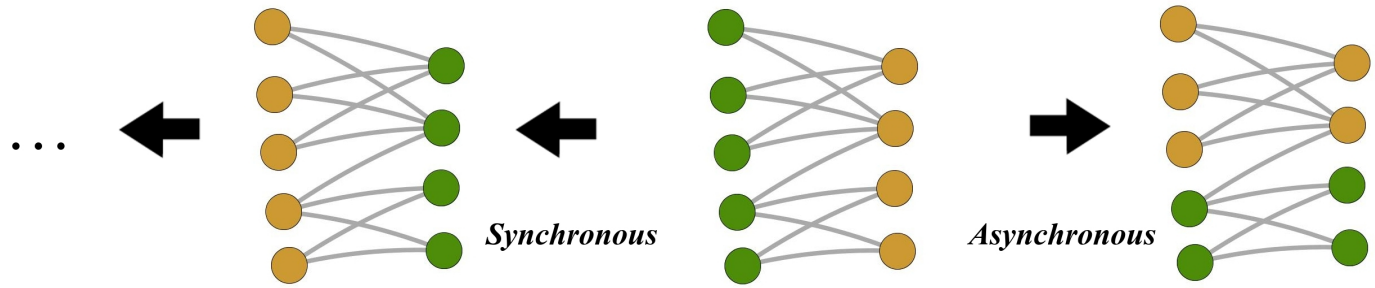


Near linear time complexity [40, 28, 46].

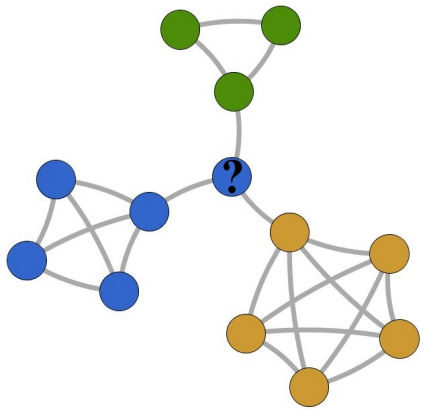
[2]Nodes are updated sequentially. Ties are broken uniformly at random.

Issues with label propagation

Oscillation of labels in, e.g., two-mode networks.
↪ Nodes are updated sequentially (*asynchronous*), in a random order [40].



Convergence issues for, e.g., overlapping communities.
↪ Node's label is retained, when among most frequent [40].

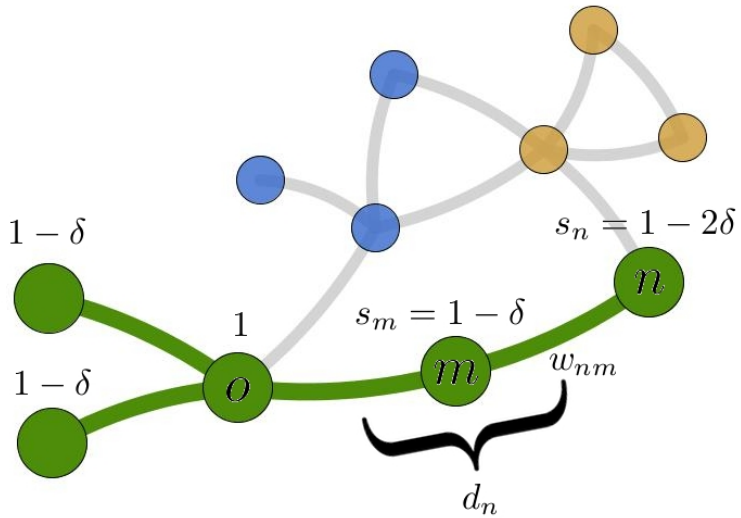


Label hop attenuation

Emergence of a *major community* (in large networks).
$\hookrightarrow$ Label *hop attenuation* [28]: each label l_n has associated a score s_n (initialized to 1) that decreases by $\delta \in [0, 1]$ after each step. Then,

$$\forall n \in N : c_n = \operatorname{argmax}_l \sum_{m \in \mathcal{N}_n^l} s_m w_{nm} \text{ and } s_n = \left(\max_{m \in \mathcal{N}_n^{c_n}} s_m \right) - \delta.$$

Actually, $s_n = 1 - \delta d_n$, where $d_n = \left(\min_{m \in \mathcal{N}_n^{c_n}} d_m \right) + 1$.



Some issues not discussed (e.g., oscillation of labels [40], stability [47]).

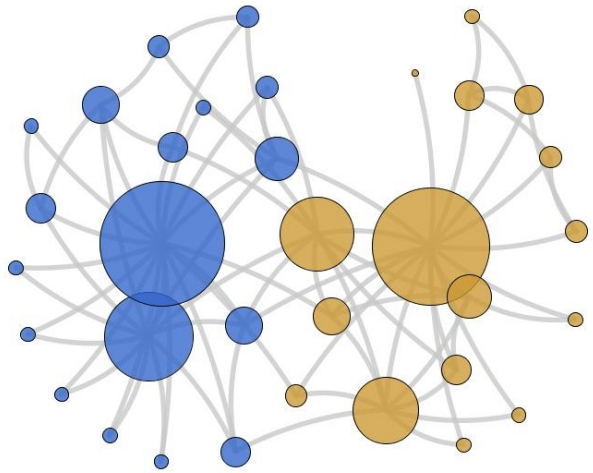
Outline

Node propagation preference

Applying *node preference* [28] (i.e., propagation strength) can improve the algorithm. Thus,

$$\forall n \in N : c_n = \underset{l}{\operatorname{argmax}} \sum_{m \in \mathcal{N}_n^l} f_m^{\alpha} s_m w_{nm},$$

for some preference f_n and parameter α.



(c) Zachary's karate club [50]

However, static measures for f_n do not work in general (see paper).

dDaLPA & oDaLPA algorithms

Estimate *diffusion* within (current) communities, i.e.,

$$p_n = \sum_{m \in \mathcal{N}_n^{c_n}} p_m / deg_m^{c_n},$$

using a random walker.

Apply preference to:

- the *core* of each (current) community, i.e.,

$$f_n^{\alpha} = p_n,$$

- the *border* of each (current) community, i.e.,

$$f_n^{\alpha} = 1 - p_n.$$

We get *defensive* and *offensive diffusion and label propagation algorithm* (*dDaLPA* and *oDaLPA* respectively.)

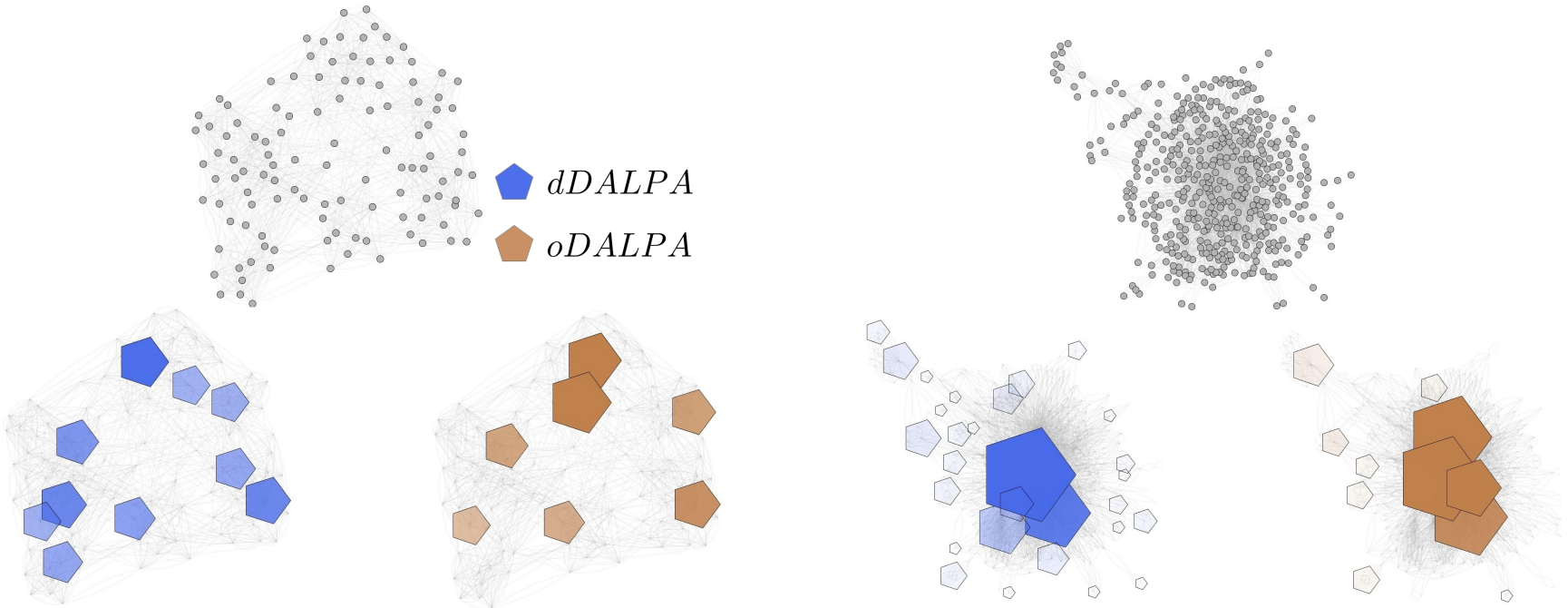
dDaLPA & *oDaLPA* algorithms, cont.

Algorithm (*dDaLPA*)

$\{$*Initialization.*$\}$
while not *converged* **do**
  shuffle(N)
  for $n \in N$ **do**
    $c_n \leftarrow \text{argmax}_l \sum_{m \in \mathcal{N}_n^l} p_m(1 - \delta d_m) w_{nm}$ $\{1 - p_m$ *for oDaLPA.*$\}$
    $p_n \leftarrow \sum_{m \in \mathcal{N}_n^{c_n}} p_m / deg_m^{c_n}$ $\{deg_m$ *for oDaLPA.*$\}$
    if c_n *has changed* **then**
      $d_n \leftarrow (\min_{m \in \mathcal{N}_n^{c_n}} d_m) + 1$
    end if
  end for
  $\{$*Re-estimation of* δ *(see paper).*$\}$
end while

Defensive preservation & offensive expansion of comm.

- *dDaLPA defensively preserves* the communities – high "recall".
- *oDaLPA offensively expands* the communities – high "precision".

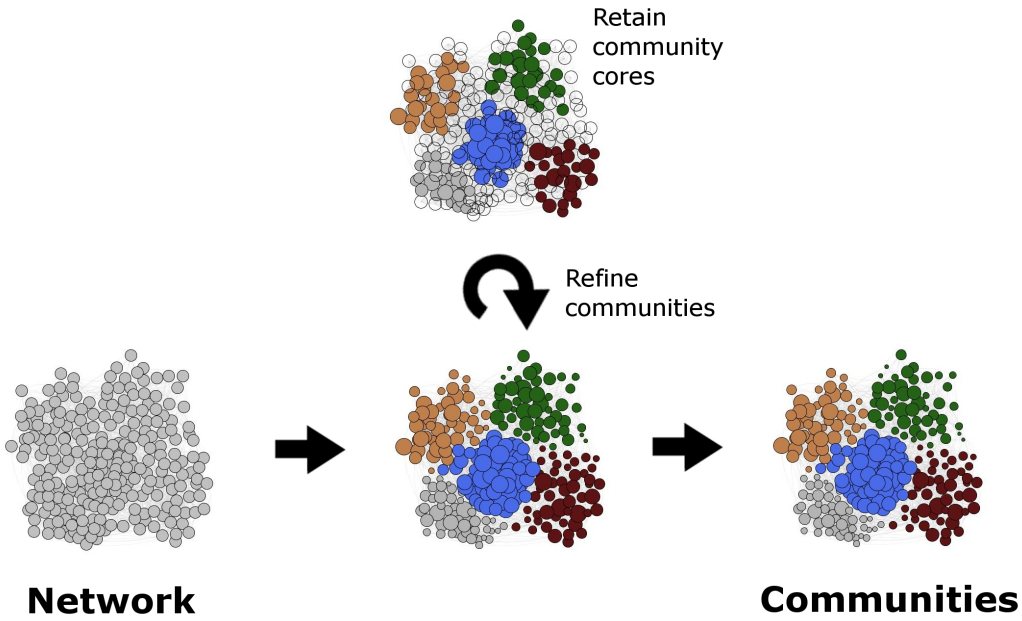


(d) American college football league [14]. (e) Nematode Caenorhabditis elegans [21].

Combining the two strategies

Find initial communities with *dDaLPA*, and refine them with *oDaLPA* – high "recall" and "precision".

However, simply running the algorithms successively does not work. Thus, relabel some of the nodes, e.g., a half.



We get *K-Cores* algorithm.

K-Cores algorithm

Algorithm (*K-Cores*)

$C \leftarrow dDaLPA(G, W)$ {*Defensive propagation.*}
while $|C|$ *decreases* **do**
  for $c \in C$ **do**
    $m_c \leftarrow median(\{p_n |\ n \in N \wedge c_n = c\})$
    {*Relabel nodes with* $c_n = c$ *and* $p_n \leq m_c$ *(i.e. retain cores).*}
  end for
  $C \leftarrow oDaLPA(G, W)$ {*Offensive propagation.*}
end while

Outline

Experimental testbed

Experimental testbed:

- Lancichinetti et al. [22] benchmark networks (see paper),
- random graph à la Erdös-Rényi [10] (see paper),
- 22 real-world networks (moderate size),
- 9 large real-world networks (over 10^6 edges).

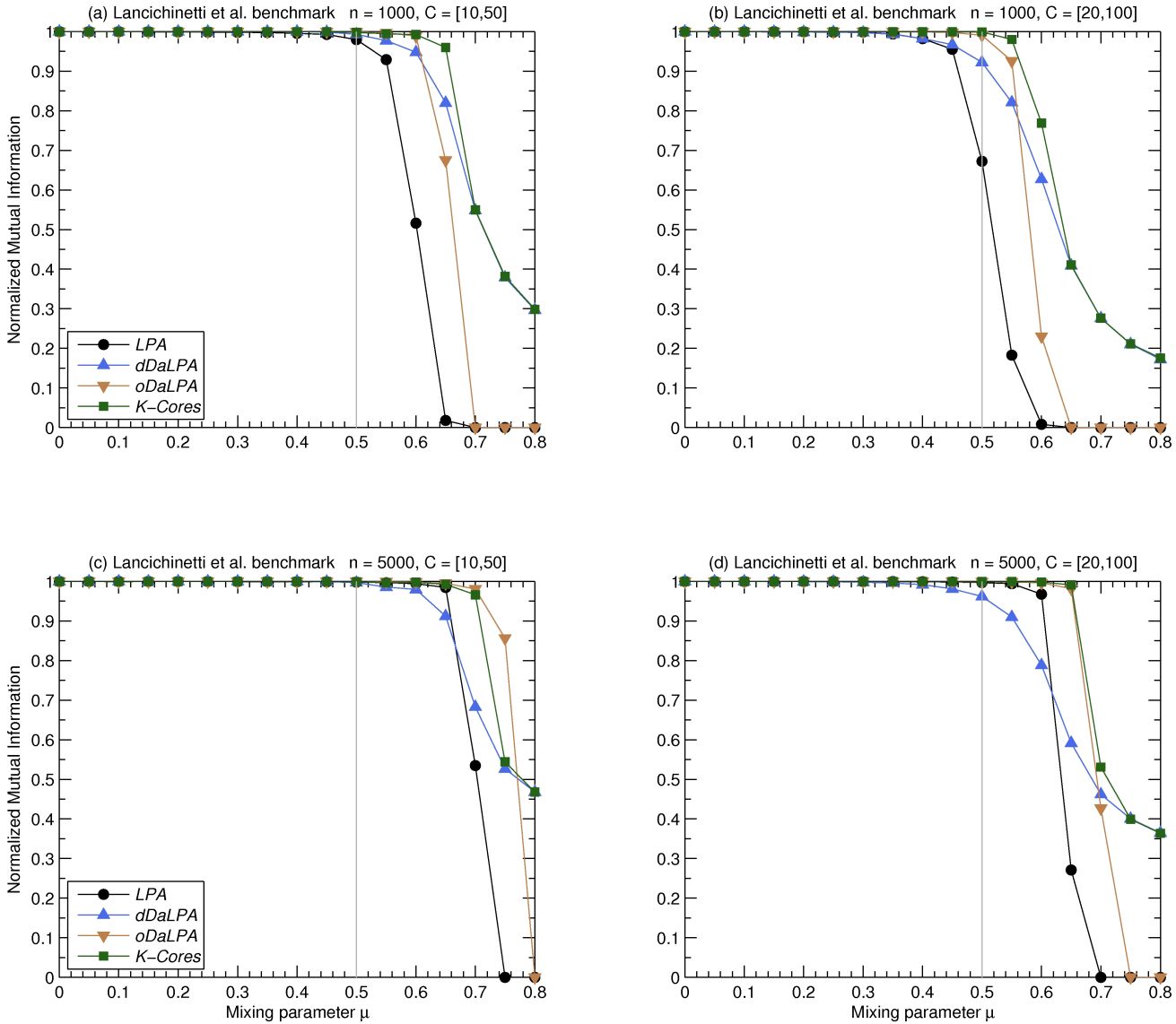
Results are assessed in terms of *modularity* Q, i.e.,

$$Q = \frac{1}{2|E|} \sum_{n,m \in N} \left(A_{nm} - \frac{deg_n deg_m}{2|E|} \right) \delta(c_n, c_m).$$

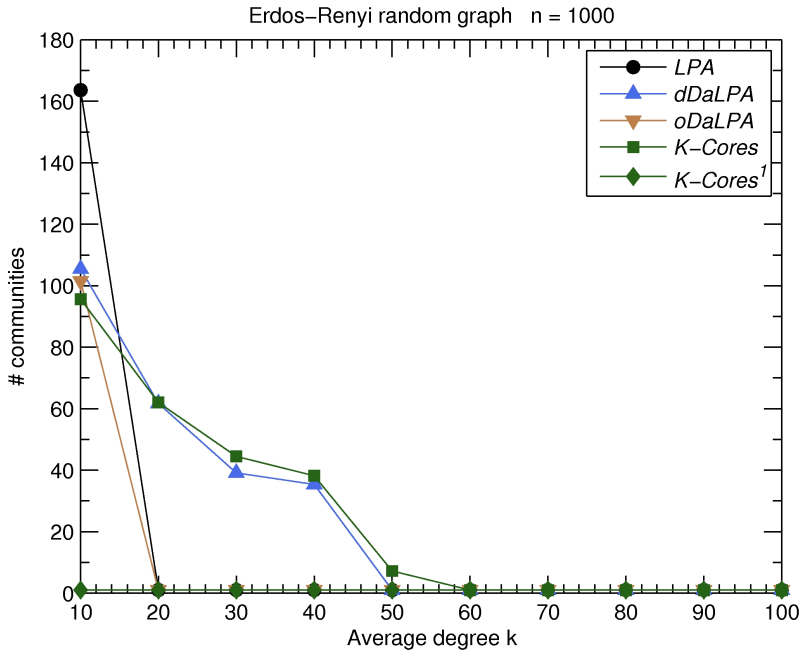
and *Normalized Mutual Information*, i.e.,

$$NMI = \frac{2I(C,P)}{H(C) + H(P)}, \text{ where } I(C,P) = H(C) - H(C|P).$$

Lancichinetti et al. benchmark



Erdös-Rényi random graph



Real-world networks

Type	Network	Nodes	Edges	*LPA*	*dDaLPA*	*oDaLPA*	*K-Cores*
Communication	*uni*	1133	5451	0.364	0.481	0.389	**0.518**
	enron	36692	367662	0.355	0.514	0.380	**0.516**
Social	*football*	115	616	0.592	0.593	0.595	**0.600**
	jazz	198	2742	0.346	**0.418**	0.377	**0.418**
	wiki	7115	103689	0.056	0.195	0.046	**0.202**
	epinions	75879	508837	0.106	0.288	0.111	**0.291**
Protein	*yeast*	2114	4480	0.665	0.733	0.720	**0.793**
Metabolic	*elegans*	453	2025	0.122	0.172	0.131	**0.173**
Peer-to-peer	*gnutella*	62586	147892	0.338	0.412	0.387	**0.447**
Web	*blogs*	1490	16718	0.400	0.424	0.424	**0.426**
Collaboration	*genrelat*	5242	28980	0.737	0.769	0.779	**0.820**
	codmat^3	27519	116181	0.596	0.611	0.627	**0.687**
	codmat^5	36458	171736	0.548	0.575	0.590	**0.648**
	hep	12008	237010	0.484	**0.585**	0.518	**0.585**
	astro	18772	396160	0.326	**0.538**	0.337	**0.538**
Software	*engine*	139	243	0.689	0.724	0.726	**0.747**
	jung	436	1303	0.611	0.587	0.623	**0.631**
	javax	2089	7934	0.723	0.687	0.725	**0.768**
Power	*power*	4941	6594	0.595	0.690	0.698	**0.820**
Internet	*oregon*^{3}	767	3591	0.302	0.210	**0.354**	0.210
	oregon^6	22963	48436	0.498	0.347	**0.541**	0.347
	nec	75885	357317	0.683	0.628	0.688	**0.736**

Tabela: Mean modularities Q (100 to 100000 runs).

Large real-world networks

DPA – faster alternative for *K-Cores*.

DPA^+ – *DPA* with simple hierarchical investigation.

DPA^* – *DPA* with hierarchical *core extraction* technique.

For more see [46].

Network	Nodes	Edges	*LPA*	*K-Cores*	*DPA*	DPA^+	DPA^*
amazon	0.3M	1.2M	0.681/15	0.783/273	0.700/34	**0.883/65**	0.856/78
ndedu	0.3M	1.5M	0.838/53	0.891/471	0.860/50	0.897/37	**0.901/58**
road	1.1M	3.1M	0.552/10	0.847/895	0.626/82	**0.985/136**	0.883/142
google	0.9M	4.3M	0.801/15	0.889/444	0.820/59	0.962/45	**0.967/48**
skitter	1.7M	11.1M	0.746/25	-	0.755/126	0.680/52	**0.801/76**
movie	0.4M	15.0M	0.524/21	-	0.533/147	0.474/39	**0.606/71**
nber	3.8M	16.5M	0.576/109	-	0.582/336	0.707/112	**0.739/308**
live	4.8M	69.0M	0.673/100	-	0.548/206	0.683/73	**0.688/125**
webbase	14.5M	101.0M	0.894/38	-	0.923/114	0.942/43	**0.954/39**

Tabela: Peak modularities Q and # iterations (1 to 10 runs).

Conclusion

- Different advanced label propagation algorithms.
- Two unique strategies of community formation –
  different types of networks favor different formation strategies.
- Extensions and improvements for large networks.

For more see [46].

For material see `http://wwwlovre.appspot.com/?navigation=research_main`.

Thank you.

- Web graph from the Stanford WebBase Project (crawl from January 2010). http://diglib.stanford.edu:8091/~testbed/doc2/WebBase/ (2010)
- Adamic, L.A., Glance, N.: The political blogosphere and the 2004 U.S. election. In: Proceedings of the International Workshop on Link Discovery. pp. 36–43 (2005)
- Albert, R., Jeong, H., Barabási, A.: The diameter of the world wide web. Nature 401, 130–131 (1999)
- Barabási, A., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
- Barber, M.J., Clark, J.W.: Detecting network communities by propagating labels under constraints. Phys. Rev. E 80(2), 026129 (2009)

- Blondel, V.D., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. P10008 (2008)
- Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70(6), 066111 (2004)
- Danon, L., Díaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. J. Stat. Mech. P09008 (2005)
- Donetti, L., Munoz, M.A.: Detecting network communities. J. Stat. Mech. P10012 (2004)
- Erdős, P., Rényi, A.: On random graphs 1. Publ. Math. Debrecen 6, 290–297 (1959)
- Fortunato, S.: Community detection in graphs. Phys. Rep. 486(3-5), 75–174 (2010)
- Freeman, L.: A set of measures of centrality based on betweenness. Sociometry 40(1), 35–41 (1977)


Freeman, L.C.: Centrality in social networks: Conceptual clarification. Soc. Networks 1(3), 215–239 (1979)


Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. In: Proceedings of the National Academy of Sciences of United States of America. pp. 7821–7826 (2002)


Gleiser, P., Danon, L.: Community structure in jazz. Adv. Complex Syst. 6(4), 565 (2003)


Gregory, S.: Finding overlapping communities in networks by label propagation (2009)


Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F., Arenas, A.: Self-similar community structure in a network of human interactions. Phys. Rev. E 68(6), 065103 (2003)


Hall, B.H., Jaffe, A.B., Tratjenberg, M.: The NBER patent citation data file: Lessons, insights and methodological tools. Tech. rep., National Bureau of Economic Research (2001)


Hoerdt, M., Jaeger, M., James, A., Magoni, D., Maillard, J., Malka, D., Merindol, P.: Internet IPv4 overlay map produced by network cartographer (nec). http://www.labri.fr/perso/magoni/nec/ (2003)


Jeong, H., Mason, S.P., Barabási, A., Oltvai, Z.N.: Lethality and centrality of protein networks. Nature 411, 41–42 (2001)


Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.: The large-scale organization of metabolic networks. Nature 407, 651–654 (2000)


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


Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web 1(1) (2007)

- Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: Proceedings of the ACM International Conference on World Wide Web (2010)
- Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: Densification laws, shrinking diameters and possible explanations. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2005)
- Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data 1(1) (2007)
- Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. e-print arXiv:0810.1355 (1) (2008)
- Leung, I.X.Y., Hui, P., Liò, P., Crowcroft, J.: Towards real-time community detection in large networks. Phys. Rev. E 79(6), 066107 (2009)


Liu, X., Murata, T.: Advanced modularity-specialized label propagation algorithm for detecting communities in networks. Physica A 389(7), 1493 (2009)


Liu, X., Murata, T.: Community detection in large-scale bipartite networks. In: Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology. vol. 1, pp. 50–57 (2009)


MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297 (1967)


Newman, M.E.J.: The structure of scientific collaboration networks. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 98, pp. 404–409 (2001)


Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74(3), 036104 (2006)


Newman, M.E.J.: Symmetrized snapshot of the structure of the internet at the level of autonomous systems. http://www-personal.umich.edu/~mejn/netdata/ (2006)


Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69(2), 026113 (2004)


Newman, M.E.J., Leicht, E.A.: Mixture models and exploratory analysis in networks. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 104, pp. 9564–9569 (2007)


Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814 (2005)


Pang, C., Shao, F., Sun, R., Li, S.: Detecting community structure in networks by propagating labels of nodes. In: Proceedings of the International Symposium on Neural Networks. pp. 839–846 (2009)


Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 101, pp. 2658–2663 (2004)


Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76(3), 036106 (2007)


Richardson, M., Agrawal, R., Domingos, P.: Trust management for the semantic web. In: Proceedings of the International Semantic Web Conference. vol. 2, pp. 351–368 (2003)


Ronhovde, P., Nussinov, Z.: Local resolution-limit-free potts model for community detection. Phys. Rev. E 81(4), 046114 (2010)


Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. In: Proceedings of the National Academy of Sciences of United States of America. vol. 105, pp. 1118–1123 (2008)

- Seidman, S.B.: Network structure and minimum degree. Soc. Networks 5(3), 269–287 (1983)
- Strogatz, S.H.: Exploring complex networks. Nature 410, 268 (2001)
- Subelj, L., Bajec, M.: Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. Submitted to Phys. Rev. E
- Tibély, G., Kertész, J.: On the equivalence of the label propagation method of community detection and a potts model approach. Physica A 387(19-20), 4982–4984 (2008)
- Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
- Wen, L., Kirk, D., Dromey, R.G.: Software systems as complex networks. In: Proceedings of the IEEE International Conference on Cognitive Informatics. pp. 106–115 (2007)


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