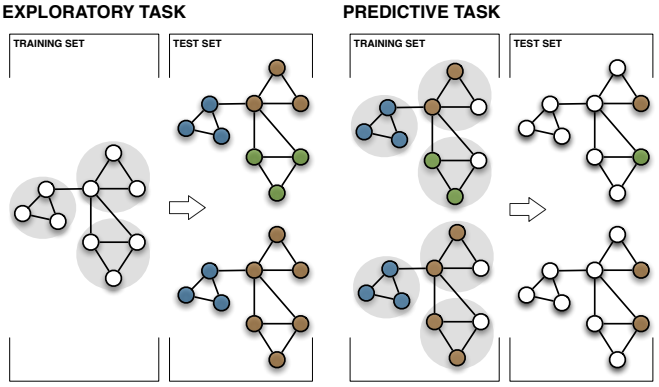
Exploratory and predictive tasks of network community detection

Lovro Šubelj
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, SI-1000 Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

Network community structure is a thoroughly investigated concept with various practical applications. However, due to the lack of data, past studies were mainly focused on networks of rather small or moderate size. Only recent research has thus shown that community structure revealed in large networks does not actually coincide with some ground truth clusters [2]. Despite this discouraging fact, we show that community information is still beneficial in practical scenarios [5].

Most of the past work focused on exploratory task of network community detection. Here, communities revealed by an algorithm are compared to some ground truth clusters using, e.g., normalized mutual information (NMI). Although exploratory analysis can provide a valuable insight, predictive analytics is far more common and useful in practice. In this case, revealed communities are utilized to predict the unknown node labels as, e.g., the most frequent labels in the concerned nodes' communities. We measure classification accuracy (CA), particularly the gain compared to a baseline approach that considers merely the neighborhoods of the concerned nodes.



As our first example, we consider a citation network of over 500 thousand papers published by the American Physical Society[1] (APS). We select journal and section information as ground truth clusters, while we withhold the information of all papers in 2013 for the predictive task. As our second example, we consider a reference network between over 100 thousand US diplomatic cables released by the WikiLeaks[2] (WL). We select privacy and embassy information as ground truth clusters, while we withhold the information of all cables in 2010 for the predictive task.

We apply 14 community detection algorithms to APS network and 26 algorithms to WL network. These include different spectral methods [3], modularity optimization [1], map equation algorithms [4] and approaches based on dynamical processes [6].

[1]http://www.aps.org/
[2]http://wikileaks.org/

CLUSTERS		EXPLORAT.	PREDICITVE	CORRELAT.
Data	#	NMI	CA (gain)	Spearman
APS	12	0.356	72.8% (6.2%)	−0.888
	301	0.365	41.4% (1.0%)	0.731
WL	3	0.131	51.3% (23.5%)	−0.724
	263	0.648	48.1% (13.7%)	0.911

The results can be summarized as follows:

(1) algorithms perform poorly on exploratory task with NMI below 0.5 in most cases. This is not surprising, since ground truth clusters were selected rather arbitrarily and likely do not even match the granularity of communities;

(2) algorithms perform surprisingly well on predictive task with up to 24% gain in CA. Thus, despite the lack of one-to-one correspondence with ground truth clusters, community structure is still beneficial for prediction;

(3) performance of algorithms on exploratory and predictive tasks reveals strong positive correlation in the case of smaller clusters. In other words, for smaller clusters, the same algorithms perform well on both tasks; and

(4) performance of algorithms on exploratory and predictive tasks reveals strong negative correlation in the case of larger clusters. In other words, for larger clusters, different algorithms perform well on different tasks.

Despite poor performance on exploratory task, state-of-the-art community detection algorithms quite efficiently solve predictive task. Note also that evaluating the algorithms on only one of the tasks, as most commonly done in the literature, will give misleading results in the case of larger clusters.

Authors thank American Physical Society and WikiLeaks for providing the data. The work has been supported by Slovenian Research Agency Program No. P2-0359, Slovenian Ministry of Education, Science and Sport Grant No. 430-168/2013/91, and European Union, European Social Fund.

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

[2] D. Hric, R. K. Darst, and S. Fortunato. Community detection in networks: Structural communities versus ground truth. *Phys. Rev. E*, 90(6):062805, 2014.

[3] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, 1998.

[4] 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.

[5] L. Šubelj. Large network community detection in practical scenarios. In *Proceedings of the International Workshop on Social Network Analysis*, Capri, Italy, 2015.

[6] 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, 2011.