Large network community detection in practical scenarios

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, many of the past studies were focused on networks of rather small or moderate size. Thus, only recent research has shown that community structure revealed in large networks does not actually coincide with some ground truth clusters [2]. Despite this disturbing fact, we here show that community information can still be beneficial in practical scenarios.

As our first example, we consider a citation network of over 500 thousand papers published by American Physical Society[1] until 2013. We predict the journal information of all papers in 2013 based on the journal information of the papers published until 2012 and complete citation information. Constructing a simple majority classifier for each paper based merely on its cited papers, and thus merely on the neighbors of the corresponding node, gives 67% prediction accuracy. However, extending nodes' neighborhoods to the entire communities in which they reside, boosts the accuracy for additional 5%.

As our second example, we consider a collaboration network between over 300 thousand authors compiled from DBLP[2] computer science repository [6]. We predict the entire list of publication venues of an author (e.g., journal or conference) based on the publication venues of other authors and complete collaboration information. The accuracy of a classifier based merely on nodes' neighborhoods, and thus merely on immediate collaborations, is 31%, whereas the classifier based on community information has 35% prediction accuracy.

As our final example, we consider a reference network between over 100 thousand US diplomatic cables until 2010 released by WikiLeaks[3]. We predict the classifications of all cables in 2010 (e.g., secret or confidential) based on the classifications of the cables until 2009 and complete reference information. The classifier based merely on nodes' neighborhoods, and thus merely on referenced cables, has 28% prediction accuracy, whereas the community information improves the accuracy by almost 20%.

We stress that above superior performance is obtained by the algorithms based on label propagation [5], in contrast to more standard community detection approaches like spectral methods [3], modularity optimization [1] and map equation algorithm [4].

Authors thank American Physical Society, University of Trier and WikiLeaks for providing the data. The work has been supported in part by the Slovenian Research Agency Program No. P2-0359, by the Slovenian Ministry of Education, Science and Sport Grant No. 430-168/2013/91, and by the 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 and M. Bajec. Group detection in complex networks: An algorithm and comparison of the state of the art. *Physica A*, 397:144–156, 2014.

[6] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.*, 42(1):181–213, 2015.

[1] http://www.aps.org/
[2] http://dblp.uni-trier.de/
[3] http://wikileaks.org/