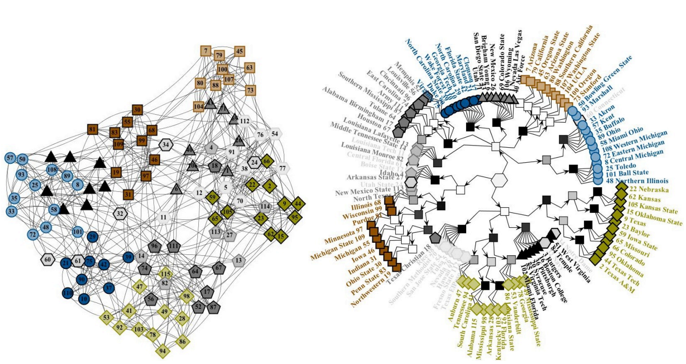
Network group discovery by hierarchical label propagation

Lovro Šubelj, Marko Bajec
University of Ljubljana, Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
{lovro.subelj,marko.bajec}@fri.uni-lj.si

Complex real-world networks often reveal characteristic groups of nodes with common linking pattern. Besides densely linked groups known as communities, networks also consist of groups of structurally equivalent nodes denoted modules, and different mixtures of these, with core/periphery and hub & spokes structures as special cases. These are of value in various applications, especially in large social and information networks. However, despite an outburst of community detection algorithms in the last decade, approaches for other groups of nodes are relatively rare and often limited. We here present a hierarchical label propagation algorithm (*HPA*) proposed in [1] for general group detection. The algorithm requires no apriori knowledge and has near ideal complexity, while the main novelty is that groups are revealed through an adequate hierarchical refinement that enables straightforward discovery of different types of groups.

Fig. 1 demonstrates the benefits of a general group detection approach like *HPA*. We apply the algorithm to a famous social network that represents matches played among US college football teams in the 2000 season [2]. The network has been of considerable interest in the past literature, since it reveals clear communities that coincide with the division into conferences. Revealed group hierarchy in Fig. 1b indeed contains communities on higher levels, however, several of these are further refined into well defined configurations of modules on lower levels. For instance, group of nodes at the top of Fig. 1a is in fact a complete multi-partite graph on ten nodes. Whether the particular group hierarchy would be of any interest in practical applications remains unclear, still, most of the groups present would remain overlooked under the standard community framework.

HPA is validated on various synthetic and real-world networks, and rigorously compared against a larger number of state-of-the-art approaches on group detection, hierarchy discovery and link prediction tasks (that may be of independent interest). Analysis shows that *HPA* is comparable to the state-of-the-art in community detection, while superior in general group detection and link prediction (Fig. 2). Moreover, while different approaches can accurately solve the community detection problem, there is an absence of other reliable approaches for general groups.



(a) Social network (b) Revealed with *HPA*

Figure 1: Group hierarchy of a social network, where node shapes correspond to a known sociological division (see text for details). Shades of inner nodes of the hierarchy are proportional to linking probabilities between groups represented by the sub-hierarchies.

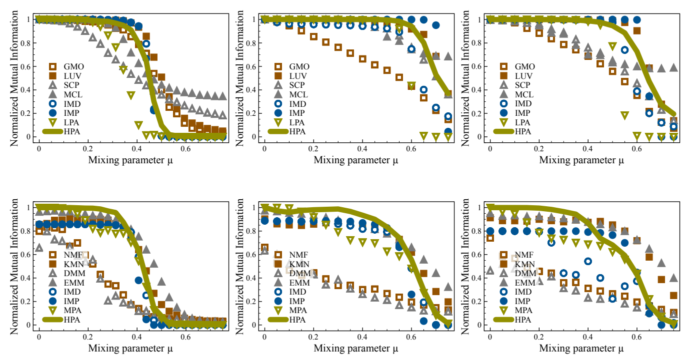


Figure 2: Comparison of different approaches on (top) community and (bottom) general group detection tasks (see [1] for details). The distribution of links is controlled by a mixing parameter μ, where lower values correspond to a clearer group structure. Notice that merely *HPA* can accurately reveal general groups of nodes planted in these networks (bottom).

HPA algorithm is compared against twelve state-of-the-art approaches. For community detection task, we consider greedy optimization of modularity (*GMO*), multi-stage modularity optimization or Louvain method (*LUV*), sequential clique percolation (*SCP*), Markov clustering approach (*MCL*), structural compression or Infomod (*IMD*), random walk-based compression or Infomap (*IMP*) and label propagation algorithm (*LPA*).

For general group detection task, we adopt symmetric nonnegative matrix factorization (*NMF*), k-means data clustering (*KMN*), mixture models (*EMM*), degree-corrected mixture models (*DMM*), model-based propagation algorithm (*MPA*), structural compression (*IMD*) and the best approach above (*IMP*). Note that *NMF*, *KMN*, *EMM* and *DMM* demand the number of groups apriori, which is a big disadvantage in practice.

The work has been supported 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] Šubelj, L. and Bajec, M. (2014) *Physica A* **397**, 144–156.
[2] Girvan, M. and Newman, M. E. J. (2002) *P. Natl. Acad. Sci. USA* **99(12)**, 7821–7826.