Community detection in networks: a heuristic version of Girvan Newman algorithm

Identificazione delle comunità nei grafi: una versione euristica dell'algoritmo di Girvan Newman

Ilaria Bombelli and Lorenzo Di Rocco

Abstract Complex problems in different fields can be modeled by graph data structures. An important and challenging task is the community detection, i.e. the identification of highly connected components. The Girvan Newman algorithm identifies high-quality communities but is not computationally suitable for huge graph analysis. Therefore, we propose a heuristic version of this algorithm, by considering an approximated measure of the edge betweenness. We evaluate the performances of our proposal on benchmark networks.

Abstract *I grafi permettono di modellare problemi complessi di diversa natura. Una delle questioni più stimolanti e importanti è il rilevamento delle comunità, cioè l'identificazione di componenti altamente connesse. L'algoritmo di Girvan Newman è in grado di identificare tali componenti con successo, ma non è computazionalmente adatto ad un'analisi di grafi di grandi dimensioni. Pertanto, proponiamo una versione euristica di questo algoritmo, approssimando la misura di betweenness degli archi. Analizziamo le prestazioni della nostra proposta utilizzando reti benchmark.*

Key words: Network, Communities Detection, Edge Betweenness

Ilaria Bombelli
University of Rome La Sapienza, Statistical Sciences Department, Rome, Italy
e-mail: ilaria.bombelli@uniroma1.it

Lorenzo Di Rocco
University of Rome La Sapienza, Statistical Sciences Department, Rome, Italy
e-mail: lorenzo.dirocco@uniroma1.it

1 Introduction

Networks are interesting tools to represent complex problems that arise in different situations: for example, in a society, networks can represent how class students interact with one another; in technological field, networks can represent flights routes between cities.

In a network framework, a really challenging task is the detection of *communities* (or clusters of nodes). Plenty of literature deals with this task and, among the hierarchical algorithms, the most famous ones are Louvain [2] and Girvan Newman [5]. The former follows an *agglomerative* approach, while the latter is characterized by a *divisive* one.

The Girvan Newman (GN) algorithm suffers from an important drawback: it is unfeasible for community detection problems on large-scale networks, as also the authors highlight [5]. Indeed, the GN algorithm is based on the computation of the edge betweenness, which is practically an extremely expensive task.

For this reason, we propose a Heuristic Girvan Newman (HGN) algorithm, considering an approximation of the edge betweenness. In this way, the computational effort of each iteration of the GN algorithm is reduced.

In Section 2, we explore deeply the methodology underlying our proposal. In Section 3, we assess the performances of our proposal on benchmark networks. Finally, in Section 4 final remarks and future research directions are discussed.

2 Methods

Complex networks of highly connected data are commonly represented by graphs. A graph $G = (V,E)$ is a data structure built on top of a set of vertices (V) and a set of edges (E). The vertices are considered to be the entities of the network, while the edges describe the occurring interactions.

In network science, hubs identification is an important task and aims to identify pivotal vertices. Different measures have been proposed in the literature to quantify the importance of a vertex. Freeman's *betweenness* [4] is a widely used index that evaluates the centrality of a vertex considering the number of shortest paths passing through the vertex.

The definition of betweenness can be extended to the edges [5]. The so-called *edge betweenness* for an edge $e \in E$ is defined as follows :

$$ebw(e) := \sum_{\substack{j,k=1,\dots,|V| \\ j \neq k}} \frac{\sigma_{jk}(e)}{\sigma_{jk}} \tag{1}$$

where σ_{jk} is the number of shortest paths connecting node v_j and node v_k, and $\sigma_{jk}(e)$ is the number of shortest paths connecting node v_j and node v_k that run along the edge $e \in E$.

The greatest values of this index are usually associated with *bridge-like* edges, i.e. those connecting different communities of the network. The GN algorithm is one of the main methods for communities detection and it is based on a recursive procedure that at each iteration deletes the edge(s) with the greatest betweenness.

This divisive procedure provides high-quality clustering performances but is unpractical for large-scale graph processing. This drawback is due to the shortest paths search, which is particularly expensive from a computational viewpoint. A possible solution consists in parallelizing the elaborations on a High Performance Computing architecture, obtaining in this way a significant performance speed-up. However, it is worth considering also high-quality approximations of the edges betweenness to reduce the computational effort.

In this work, we propose a heuristic version of the GN algorithm that estimates the edges betweenness through an accurate sampling technique. More precisely, we take into account the approximation proposed by [8], which we call DIAM, as suggested by [1]. This approach evaluates an appropriate sample size r according to the graph structure, it randomly selects a sample of paths $S = \{p_i, i = 1, \ldots, r\}$ and it returns for each edge $e \in E$ the following estimator:

$$\widehat{ebw} = \frac{1}{r} \sum_{p \in S} \mathbb{I}_p(e) \tag{2}$$

where $\mathbb{I}_p(e)$ is 1 if e belongs to p, 0 otherwise. In the following, we provide more details about the procedure.

Firstly, the estimates of the edge betweenness $\{\widehat{ebw}_i, i = 1, \ldots, |E|\}$ are all set to 0. Then, r pairs of vertices are sampled uniformly from V. For each sampled pair (u, v): (a) all the shortest paths $S_{u,v}$ between u and v are computed, (b) a path p is selected uniformly at random from $S_{u,v}$, (c) the betweenness estimates of the edges belonging to p are increased by $1/r$.

The sample size formula is obtained considering an application of the Vapnik-Chervonenkis dimension [9] and it is defined as follows:

$$r := \frac{c}{\varepsilon^2} \left(\lfloor \log_2(\tilde{D}(G) - 1) \rfloor + 1 + \ln \frac{1}{\delta} \right) \tag{3}$$

where c is a positive constant (typically set to 0.5, as [7] suggested), $\varepsilon \in [0, 1]$ is the additive error, $1 - \delta$ (with $\delta \in [0, 1]$) is the probability of the error and $\tilde{D}(G)$ is the 2-approximation of the diameter of the graph $D(G)$ [8].

Summing up, the HGN algorithm is based on the following steps:

1. Calculate the approximated betweenness for all edges using DIAM.
2. Remove the edge(s) with the highest betweenness.
3. Recalculate the approximated betweenness for all edges affected by the removal.
4. Repeat from step 2 until no edges remain.

It is worth mentioning that in the worst-case scenario at each iteration GN computes all the shortest paths between $n(n-1)$ pairs of vertices, while HGN only between r

pairs. Moreover, our implementation allows to stop the algorithm when the required number of communities K is reached.

3 Experimental Analysis

We evaluate the performances of our proposal on benchmark networks in terms of *modularity* [3] and *Adjusted Rand Index* (ARI) [6]. We briefly describe the networks and then we show the main results.

3.1 Dataset

The benchmark networks that we consider are both characterized by an underlying known community structure. *Zachary karate club network* ([10]) represents social ties among the members of a university karate club collected by Wayne Zachary in 1977. *American college football network* ([5]) represents American football games between Division IA colleges during the regular season of Fall 2000. The former contains 78 edges, 34 vertices, and $K = 2$ known communities; the latter contains 616 edges, 115 vertices, and $K = 12$ known communities.

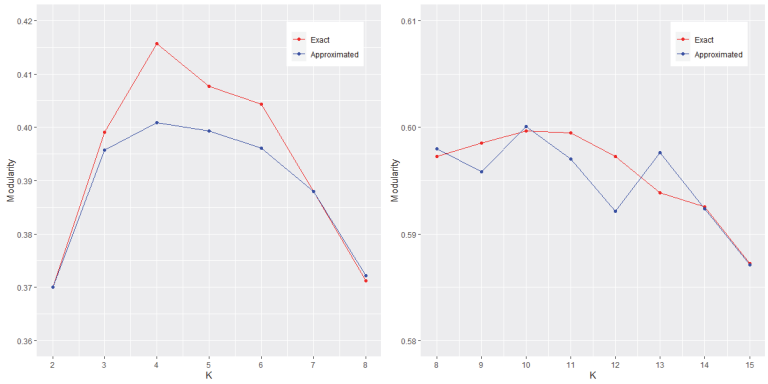
3.2 Results

As a preliminary analysis, we assess the accuracy of DIAM, since, to the best of our knowledge, there are no benchmarks in the literature. We notice that DIAM recognizes the highest-scoring edges, but, by sorting the edges in a decreasing order w.r.t the betweenness, mismatches between approximated ranking and exact one can occur. This implies that the edge deleted by HGN at i-th iteration does not always correspond to the one deleted at the same iteration by the GN algorithm. Anyway, HGN provides good performances in solving community detection problems.

Figure 1 (a) shows that the modularity trends of HGN and GN, when applied to Zachary karate club dataset, are similar: therefore, $K = 4$ is the number of communities that maximizes the modularity in both cases. Moreover, by fixing $K = 2$ as the true partition suggests, both methods induce the same partition of the network, misclassifying only two units. Indeed, when comparing the two partitions against the true one, we got the same value of the ARI.

When applied to Football network, HGN leads to an oscillatory behaviour of the modularity (see Figure 1 (b)). For $K \in \{8, 10, 14, 15\}$ the values of the modularities corresponding to GN and HGN are almost the same, while greater differences are observed for $K \in \{9, 11, 12, 13\}$. However, the greatest difference, observed for $K = 12$, is equal to 0.00513. Therefore we can conclude that all the values are close to

the exact ones. Fixing $K = 12$, HGN leads to an ARI equal to 0.883, while the ARI of GN is equal to 0.885. Hence, HGN provides good results in terms of the ARI index. Indeed, comparing Figure 2 (a) and (b), we observe that the number of misclassifications in both the partitions is almost the same and the communities are made up of almost the same units.



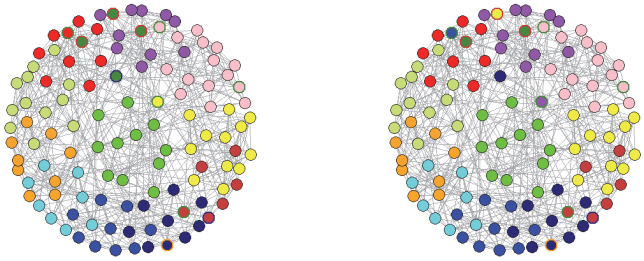
(a) Karate (b) Football

Fig. 1: Modularity values, as a function of the number of communities K, obtained by applying the exact (GN) and the approximated (HGN) algorithm with $\varepsilon = 0.05$ and $\delta = 0.1$.

4 Conclusion

In this work, we proposed a heuristic reformulation of the GN algorithm to face the problem of community detection. Since nowadays many complex problems can be represented by large-scale graphs, this approach can be a useful tool to avoid computationally heavy procedures.

Considering the good results obtained on small benchmark networks, future direction may involve the application of our method on a larger graph to accurately assess the computation time gain. Moreover, we are considering also other approaches to improve the edge betweenness approximation.



(a) GN (b) HGN

Fig. 2: Partitions obtained by applying GN and HGN on Football network with $K = 12$, $\varepsilon = 0.05$ and $\delta = 0.1$. Nodes belonging to the same community are identified by the same color. Nodes border colors compare the true partition with the obtained one: if the border is black, the units are correctly classified; otherwise, the color border identifies the belonging community in the true partition.

References

1. AlGhamdi, Z., Jamour, F., Skiadopoulos, S., Kalnis, P.: A benchmark for betweenness centrality approximation algorithms on large graphs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pp. 1–12 (2017)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
3. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE transactions on knowledge and data engineering **20**(2), 172–188 (2007)
4. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
5. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the national academy of sciences **99**(12), 7821–7826 (2002)
6. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1), 193–218 (1985)
7. Löffler, M., Phillips, J.M.: Shape fitting on point sets with probability distributions. In: European symposium on algorithms, pp. 313–324. Springer (2009)
8. Riondato, M., Kornaropoulos, E.M.: Fast approximation of betweenness centrality through sampling. Data Mining and Knowledge Discovery **30**(2), 438–475 (2016)
9. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. In: Measures of complexity, pp. 11–30. Springer (2015)
10. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of anthropological research **33**(4), 452–473 (1977)