Community structures evaluation in complex networks: A descriptive approach

Vinh-Loc Dao, Cécile Bothorel, Philippe Lenca

Community structures evaluation in complex networks: A descriptive approach

Vinh-Loc Dao, Cécile Bothorel, and Philippe Lenca

Abstract Evaluating a network partition just only via conventional quality metrics – such as modularity, conductance or normalized mutual of information – is usually insufficient. Indeed, global quality scores of a network partition or its clusters do not provide many ideas about their structural characteristics. Furthermore, quality metrics often fail to reach an agreement especially in networks whose modular structures are not very obvious. Evaluating the goodness of network partitions in function of desired structural properties is still a challenge.
Here, we propose a methodology that allows one to expose structural information of clusters in a network partition in a comprehensive way, thus eventually helps one to compare communities identified by different community detection methods. This descriptive approach also helps to clarify the composition of communities in real-world networks. The methodology hence bring us a step closer to the understanding of modular structures in complex networks.

1 Introduction

Modular structures have been noticed in a large range of real-world networks through many researches on social networks [6, 7, 11], computer networks such as the Internet [5, 13], biochemical networks [9, 12], etc. Nodes in networks have a tendency to connect preferably with the similar ones to establish functional groups, sometimes called clusters, modules or communities. Understanding modular structures of networks pays an essential role in the study of their functionalities.

Since the notion of community varies according to specific contexts, it seems not appropriate to use a global quality criteria in order to evaluate graph partitions. Depending on which kind of network is considered in which kind of application,

Vinh-Loc Dao
Institut Mines Telecom, Telecom Bretagne, UMR CNRS 6285 Lab-STICC, France.
e-mail: vinh.dao@telecom-bretagne.eu

one might need to decompose a network into clusters that possess specific features with desired structures. Once a network partition is available, the clusters need to be analyzed to verify the existence of features as well as their quality in the global image of the network.

In small networks, clusters can be evaluated manually by simple visualizations, however when the sizes grow, manual evaluation is not feasible. In these cases, expected concepts of community are mathematically translated into quality metrics such as conductance or modularity Q [3, 7, 11] in order to quantify the quality of clusters. Those quality functions score the goodness of clusters according to their associated concepts of community but can not identify or describe more specific structural patterns. In other words, many interested structure features in communities are invisible to quality functions.

In this work, we propose a methodology to describe communities through intra-cluster links and inter-cluster links in such a way that structural information is exposed comprehensively to evaluators. Such a description will help one to evaluate network partitions according to different concepts of community and to detect more sophisticated structures. Our results show that ground-truth communities composition in many real-world networks exposes a diversity in structural patterns, which are very different from the conventional notion of community.

2 Related works

Many researches have been conducted in order to understand the nature of ground-truth communities in real-world networks as well as ones identified by community detection algorithms over a broad range of networks. Although the notion of community is not straight forward, these researches provide essential information so that one can study several qualities of communities as well as their characteristics.

Leskovec *et al.* [14] compared the performance of 13 quality functions in term of their efficiencies to identify community goodness properties such as density, cohesiveness. Besides, the authors also analyzed the consistence of these quality functions' performances to many simulated perturbations.

Due to the fact that community structures may strongly differ from networks to networks. Creusefond *et al.* [4] proposed a methodology to identify groups of networks where quality functions perform consistently. The authors analyzed quality functions in three levels of granularity from node-level to community-level and network-level.

Guimerà *et al.* in [8] proposed a methodology that allows one to extract and display information about node roles in complex networks. Specifically, the role of a node in a network partition can be defined by its value of within-module connectivity and its participation into inter-cluster connections. Our work here is based on a similar method of illustration, but instead of analyzing roles of nodes in a network partition, we conduct a community-level analysis to expose the nature of communities that constitute the network.

3 Community anatomy via out degree fractions of nodes

The idea behind quality metrics is that given a partition, they indicate how the component subgraphs fit their concepts of community. In this section, we present a methodology to analyze communities in networks based on the analysis of **Out Degree Fraction (ODF)** of their nodes. We show that communities can be classified in several structural types based on the variation of their nodes' ODFs.

3.1 Community structures in term of ODF

A graph $G = (V, E)$ is composed of a set of $n = |V|$ nodes and $m = |E|$ edges where $E = (u, v) : u, v \in V$. Given a cluster S of n_S nodes, which is a subgraph of G, a function $f(S)$ quantifies a quality metric of S according to a particular notion of community. Let $d(u)$ be the degree of node u. The out degree fraction of node u in community S is measured by:

$$ODF_S(u) = \frac{|(u, v) \in E : v \notin S|}{d(u)}$$

When evaluating a community, one would normally not only want to know the average fraction of out degrees in that community, but also be curious about how are they distributed over nodes. By observing the average and the standard deviation of *ODF* values of nodes in a community, one could deduce the composition of its population. From now on, for given a community S, *meanODF* and *sdODF* denote the average and the standard deviation of *ODF* values of nodes in S respectively. They are calculated as following:

- $meanODF(S) = \frac{\sum_{u \in S} ODF_S(u)}{n_S}$
- $sdODF(S) = \left(\frac{\sum_{u \in S} [ODF_S(u) - meanODF(S)]^2}{n_S - 1} \right)^{1/2}$

As a *meanODF* value indicates the average out degree fraction of nodes in a community, a low *meanODF* implies that nodes in the community connect mostly with other nodes inside the community while a high *meanODF* means that nodes connect preferably to nodes in other communities rather than to the ones in its own. We could refer low *meanODF* and high *meanODF* characteristics to assortative structure and disassortative structure respectively. A medium value of *meanODF* in this case signifies a hybrid structure of the community as shown in Fig. 1.

We know that a standard deviation of a variable help us to understand the fluctuation of its values. Thus, to understand the composition of a community, we inspect its *sdODF* value. A low *sdODF* value implies that community's out degrees are proportionally distributed among nodes in a way to limit the variation of *ODF* values. Meanwhile, a high *sdODF* argues a diverse connection patterns of nodes in the community. In other words, based on *sdODF* value of a community, one can deter-

mine whether is there a clear division of roles [8] among nodes in the community (high *sdODF*) or nodes are just basically regular ones (low *sdODF*).

One might wonder why we chose the average and the standard deviation of *ODF* values of nodes in order to describe a community. In fact, each quality metric has its own meaning and reveals a different aspect of community structure [14]. Because the notion of community also changes according to domains of application and analysis purposes, there is actually no universal metric that can generalize the goodness of communities. Generally, one would expect a clustering where the majority of edges reside between nodes in a same cluster while there are few edges that cross to other clusters. The *meanODF* and *sdODF* are used since together, they can describe the distribution edges among nodes in an informative way. However, quality metrics could be chosen differently to match with specific concepts of community.

3.2 Community structures classification via nodes' ODFs analysis

Follow this line of argumentation, we classify communities into different structural groups based on their node orientations and their structure homogeneities. Community structures in real networks are undeniably much more complex and can not just only be described by *meanODF* and *sdODF* values. However, this simplification helps one to have a general view of networks by qualifying community anatomy. Here, we suggest to classify communities into 6 following groups, which are illustrated in Fig. 1:

- *Conventional communities* (S1 - low *meanODF* and low *sdODF*): This structure corresponds to the traditional definition of community where the majority of edges locate inside communities. Most of actual community detection methods are based on this notion. In addition, community's out degrees are homogeneously spread over its nodes.
- *Casual communities* (S2 - medium *meanODF* and low *sdODF*): Modular structure is not very clear in this type of community since there is not a clear propensity in node connections inside and outside of communities.
- *Extrovert communities* (S3 - high *meanODF* and low *sdODF*): This structure exposes an explicit disassortative structure where members in a same community are not joined together generally, but rather connect with members of other communities.
- *Full-core communities* (S4 - low *meanODF* and high *sdODF*): This group of communities shows a striking similarity with ones of $S1$ structure since both possess relatively dense inner connections. The only distinction between $S1$ and $S4$ structure is that $S4$ contains a few numbers of *active connector* nodes, which attract most out links. These connectors form a peripheral zone, whereas the other nodes constitute a core as illustrated in Fig. 1.
- *Half-core communities* (S5 - medium *meanODF* and high *sdODF*): These communities also display core-periphery structure, but there is not anymore a quantity dominance of core nodes over periphery nodes like that of in structure $S4$.

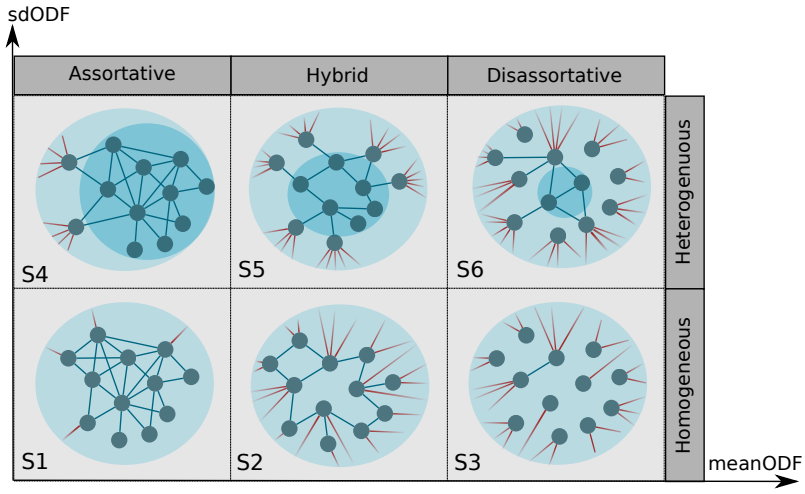

Fig. 1 Six representative community structures that can be measured by community's nodes out degree fractions (*meanODF* and *sdODF*). Blue edges represent intra-community connections and red edges (stubs) represent inter-community connections. Dark background zones in $S4, S5, S6$ structures illustrate a core-periphery arrangement.

- *Seed-core communities* (S6 - high *meanODF* and high *sdODF*): Core-periphery structure in this class of communities is degenerated or even disappeared since out-bound connectors predominate in the whole community. Most nodes connect mainly outside their community with a few exceptions. This structure have many similarities with $S3$ structure and $S5$ structure and can be considered as a transition state of community evolution between $S3$ and $S5$.

Here, a node is considered as a core node if it connects mostly inside its community whether a periphery node is the one that attaches communities together.

3.3 Network partition evaluation methodology

We propose a methodology to decompose network partitions into classes of structurally similar communities. For a given network partition:

1. Compute *meanODF* and *sdODF* values over all communities (cf Sect. 3.1)
2. Present each community by its couple of values (*meanODF*, *sdODF*) to observe the distribution of these quality metrics.
3. Choose thresholds for each quality metric in order to describe desired structure qualities for communities.
4. Identify structure profiles of all communities based on a representative map (cf Fig. 1) defined from step 3

As previously mentioned in section 3.1, quality metrics reveal different aspects of community structures. Thus, replacing *meanODF* and *sdODF* in step 1 by other quality metrics could also provide further structural information on community structures of networks under consideration. A list of quality metrics and their performances in detecting ground-truth communities in several networks can be found in [14].

Based on requirements of each specific context, thresholds to be chosen in step 3 could be varied and must not cover the whole ranges of *meanODF* and *sdODF*. In this case, the methodology also serves as a filter to eliminate unqualified communities. The choice of thresholds is, in fact, relative and can be a reference for analysis purposes.

4 Community description experiment on real-world networks

We analyze undirected, unweighted and scale-free networks [2] with ground-truth communities on SNAP dataset [10]. These communities are overlapped and may not cover the whole network, which means one node can belong to no community or can be members of many communities at a same time. The community sizes, the overlap sizes and the community memberships per node in these networks follow a power-law distribution [14].

Table 1 Network summary: N number of nodes, E number of edges, C number of communities, S average community size, O community memberships per node, $\bar{\mu}$ average conductance [14] of communities.

Network	N	E	C	S	O	$\bar{\mu}$	Community nature
Livejournal[a]	4.0M	34.7M	664414	10.79	6.24	**0.95**	User-defined communities
Youtube[a]	1.13M	3.0M	16386	7.89	2.45	**0.91**	User-defined groups
DBLP[a]	0.32M	1.05M	13477	53.41	2.76	**0.62**	Publication venues
Amazon[a]	0.33M	0.93M	75149	30.22	7.16	**0.58**	Product categories

[a] http://snap.stanford.edu/data/

Livejournal network is an online blogging community where users declare their friendships. *Youtube* network represents a social network on Youtube video sharing website. *DBLP* computer science bibliography co-authorship network is constructed in a way that two authors are connected if they published at least one paper together. *Amazon* co-purchased network represents products which are frequently bought together on Amazon website. A description of these networks and measures on their ground-truth communities can be found in Table 1.

Here, we take the conductance $\bar{\mu}$ as an example to demonstrate the weaknesses of conventional quality metrics [1]. The latter represents average portion of boundary edges in ground-truth communities of a network. This metric could tell us a global score of community quality, but they can not distinguish many different struc-

tures that exist simultaneously in networks. For instance, the average conductance $\bar{\mu}$ shows that there are above 90% of edges in *Live journal* and *Youtube* that cross communities, meanwhile these numbers are about 60% in *DBLP* and *Amazon*. However, one could not gain more insight into the differences of community structure between *Livenetwork* and *Youtube*, or between *DBLP* and *Amazon*. Thus, we describe the ground-truth communities of these networks in the next part by applying the methodology presented in section 3.3.

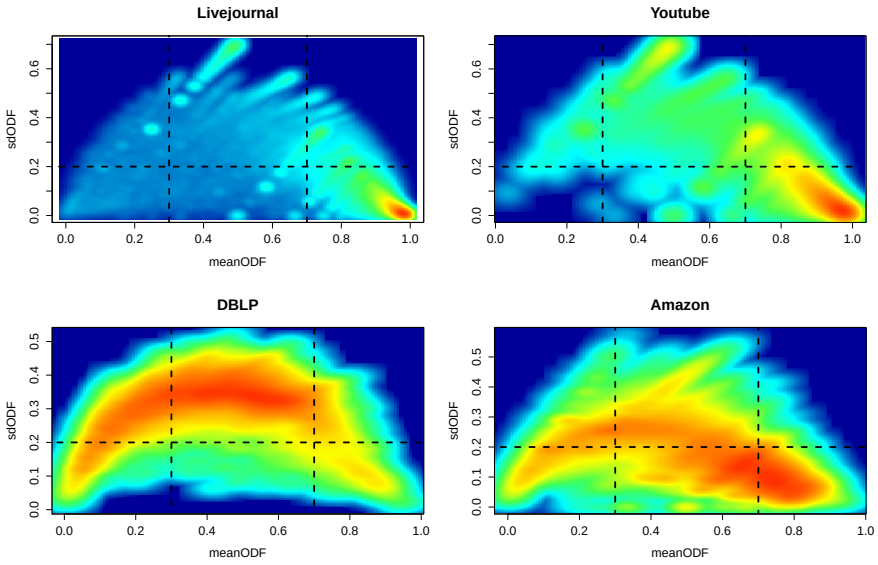

Fig. 2 The density of ground-truth communities on a *meanODF*, *sdODF* space. The dashed lines represent thresholds between the 6 presented structures $S1$ to $S6$ (cf Sect. 3.2)

Fig. 2 presents the landscape of *meanODF*, *sdODF* values of all ground-truth communities in the 4 networks (cf Sect. 3.3, step 1 and 2). We classify these communities into the 6 groups as presented in section 3.2 by choosing thresholds for *meanODF* at 0.3, 0.7 and for *sdODF* at 0.2 (cf Sect. 3.3, step 3). The landscape helps us to analyze the composition of ground-truth community structures in each network. We remind that the density landscapes in Fig. 2 do not represent the networks themselves, but the community structures in these networks.

We can see that the structural patterns of ground-truth communities within 4 networks are totally distinct. Normally, one would expect that ground-truth communities in a network have a quite similar structure, but the density landscapes in Fig. 2 illustrate a more complex community composition. While in *Livejournal* and *Youtube* networks, the majority of communities have a similar structure, those in *DBLP* and *Amazon* networks vary in a large range. Table 2 describes a global composition of the 4 networks in terms of the 6 basic structural groups ($S1$ to $S6$). We

Table 2 The composition of ground-truth communities in the 4 networks (in percentage)

Network	S1	S2	S3	S4	S5	S6
Livejournal	0.29	0.74	**90.17**	0.31	3.88	4.61
Youtube	0.08	2.36	**65.36**	1.37	**17.55**	**13.28**
DBLP	6.28	2.07	4.87	**23.44**	**57.86**	5.48
Amazon	8.33	**31.13**	**23.57**	9.13	**26.63**	1.21

find that $S3$ structure occupies around 90% and 65% of communities in *Live journal* and *Youtube* networks respectively. This implies the fact that most users in these networks usually have friendships outside their communities rather than inside. In addition, there are many closely-knit members in *Youtube* network, who are not very active outside their communities ($S5$ and $S6$).

In *DBLP* and *Amazon* network, although there is aways a dominance of some structures, we notice a more equilibrate repartition of communities over the landscapes. In the case of *DBLP*, nearly 60% of publication venues ($S5$) attract a variety type of authors in term of cooperation profile. These communities could represent traditional publication venues which gather at the same time high influence authors and newcomers. Meanwhile, there is about 23.44% publication venues where presented just a few active *eminent* authors. In *Amazon* network, the high presence of $S2$ and $S3$ structures explains that products are more often co-purchased with ones of other categories. Besides, there are also many miscellaneous product categories ($S1$, $S4$, $S5$) which consist of a high portion of products that are mostly complemented by ones in the same categories. Further analysis in natures and functionalities of products need to be conducted in order to understand this commercial network.

5 Conclusion

We know that optimizing a particular quality function could discard many interesting community structures. The methodology proposed in this paper presented a new approach to community analysis, where specialists can evaluate network partitions themselves according to contextual concepts with more insight into community structures. We also extended the notion of community, which is actually generalized for most community detection methods and then described communities in real networks in an informative way that many quality metrics fail to do. The extended notion could be applied in order to identify more complex structures in networks. Furthermore, we continue to enrich this notion by employing other pairs of metrics to describe more sophisticated characteristics that exist in real-world communities.

References

[1] Almeida, H., Guedes, D., Meira Jr., W., Zaki, M.J.: Is there a best quality metric for graph clusters? In Gunopulos, D. *et al.* (eds.) ECML PKDD 2011, Part I. LNCS, vol. 6911, 44–59. Springer, Heidelberg (2011)

[2] Barabási, A.L., Albert, R.: Emergence of Scaling in Random Networks. Science **286** (5439), 509–512 (1999)

[3] Clauset, A., Newman, M. E. J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**, 066111 (2004)

[4] Creusefond, J., Largillier, T., Peyronnet, S.: On the Evaluation Potential of Quality Functions in Community Detection for Different Contexts. Advances in Network Science: 12th International Conference and School, NetSci-X (2016)

[5] Faloutsos, M., Faloutsos, P., Faloutsos, C.,: On Power-Law Relationships of the Internet Topology. In: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, 251–262 (1999)

[6] Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., Elovici, Y.,: Link Prediction in Social Networks Using Computationally Efficient Topological Features. In: 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 73–80 (2011)

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

[8] Guimer, R., Amaral, L.A.N.: Cartography of complex networks: modules and universal roles. Journal of Statistical Mechanics: Theory and Experiment, P02001 (2005)

[9] Guimer, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Letters to Nature **7028**, 895-900 (2005)

[10] Leskovec, J., Krevl, A.: SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data (2014). Reference date: 13/12/2016

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

[12] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.-L.: Hierarchical Organization of Modularity in Metabolic Networks. Science **297** (5586), 1551–1555 (2002)

[13] Yook, S.H., Jeong, H., Barabási, A.-L.: Modeling the Internet's large-scale topology. In: Proceedings of the National Academy of Science **99**, 13382-13386 (2002)

[14] Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. Knowledge and Information System **42**, 181-213 (2015) http://dx.doi.org/10.1007/s10115-013-0693-z