A novel approach to evaluate community detection algorithms on ground truth

Giulio Rossetti, Luca Pappalardo, and Salvatore Rinzivillo

Abstract Evaluating a community detection algorithm is a complex task due to the lack of a shared and universally accepted definition of community. In literature, one of the most common way to assess the performances of a community detection algorithm is to compare its output with given ground truth communities by using computationally expensive metrics (i.e., Normalized Mutual Information). In this paper we propose a novel approach aimed at evaluating the adherence of a community partition to the ground truth: our methodology provides more information than the state-of-the-art ones and is fast to compute on large-scale networks. We evaluate its correctness by applying it to six popular community detection algorithms on four large-scale network datasets. Experimental results show how our approach allows to easily evaluate the obtained communities on the ground truth and to characterize the quality of community detection algorithms.

1 Introduction

Evaluating the results provided by a community detection algorithm is one of the most difficult tasks of complex network analysis, since there is no a shared and universally accepted definition of what a community is [1, 2]. Each approach hence defines its own idea of community and maximizes a specific quality function (e.g. modularity, density, conductance, etc.). Even though the communities identified by a given algorithm on a network are consistent with its community definition, it is not guaranteed that they are able to capture the real sub-topology of the network. For this reason, the common way to state the quality of a community detection algorithm is to evaluate the similarity between the communities it produces and the ground truth communities of the network. Generally, the communities produced by the algorithm are compared to the ground truth communities specified in the network dataset using

Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR) e-mail: name.surname@isti.cnr.it

metrics like the Normalized Mutual Information score (NMI) [3]. Unfortunately the computational complexity of this metric is quadratic in the number of communities of the network, which makes it unsuitable on large-scale complex networks where a large number of communities emerge.

In this paper we propose a novel community evaluation approach that leverages ground truth communities and copes with the computational issues that arise when calculating NMI on large community sets. To do that we define two measures, namely *community precision* and *community recall*, which provide information about how much the nodes of a given community tend to be in the same ground truth community. In particular, community precision quantifies the level of label homophily between a community and a ground truth community, while the community recall quantifies the ratio of nodes in the ground truth community covered by a given algorithm community. To validate our methods we apply six popular community detection algorithms on four large-scale networks with ground truth communities. We then compute the proposed community precision and community recall metrics on the produced community sets in order to compare them on the ground truth. We show how the evaluation can be easily performed through density scatter plots, where the presence and position of visual clusters well identify the properties of the community sets in terms of precision and recall. The evaluation can be also summarized into a single number using the $F1$-measure (the harmonic mean of community precision and community recall), which provides a clear and concise evaluation of the quality of a community set.

The paper is organized as follows. Section 2 revises the main works in community detection and community evaluation. Section 3 introduces the community precision and the community recall metrics and Section 4 describes our experiments, the community detection algorithms used (Section 4.1), the network datasets (Section 4.2) and the results obtained (Section 4.3). Finally, Section 5 concludes the paper illustrating some possibile improvements of the proposed metrics.

2 Related Works

Community detection has become during the last decade one of the most challenging and studied problems in complex network analysis, due to its relevance for a wide range of applications such as the study of information and disease spreading [4, 5], the prediction of future interactions and activities of individuals [6, 7], and even the analysis of the patterns of human mobility [8, 9]. Two surveys by Fortunato [1] and Coscia et al. [2] explore all the most popular techniques to find communities in complex networks, highlighting that several algorithms have been proposed in literature to detect different definitions of network community. The plethora of many community definitions makes the evaluation of a community detection algorithm a difficult task. In literature, the most used evaluation method is to compare the community set produced by an algorithm on a network with ground truth communities of the same network. Due to the scarse availability of real networks with

ground truth communities, the evaluation of an algorithm is often performed using synthetic network generators that also provide ground truth communities (such as the LFR benchmark [10]). In such scenario, the comparison is generally done by the Normalized Mutual Information score (NMI) a measure of similarity borrowed from information theory [3, 11, 12], defined as:

$$NMI(X,Y) = \frac{H(X)+H(Y)-H(X,Y)}{(H(X)+H(Y))/2} \tag{1}$$

where $H(X)$ is the entropy of the random variable X associated to an algorithm community, $H(Y)$ is the entropy of the random variable Y associated to a ground truth community, and $H(X,Y)$ is the joint entropy. NMI ranges in the interval [0,1] and is maximized when the algorithm community and ground truth community are identical. One drawback of NMI is that, assuming that the algorithm community set and the ground truth community set have approximately the same size n, the overall NMI computation requires $O(n^2)$ comparisons, making it unsuitable for large-scale networks.

3 Approach definitions

The computation of NMI on large community sets is often prohibitive: following equation (1) given the algorithm community set X of size m and ground truth community set Y of size n, to compute NMI we need to identify the communities best matches with cost $O(mn)$. Assuming $m \simeq n$ the NMI computation requires $O(n^2)$ comparisons thus making it often unsuitable for large-scale networks.

To overcome this drawback, we propose a novel approach that provides valuable insights on the quality of the community sets produced by a community detection algorithm. Given a community set X produced by an algorithm and the ground truth community set Y, for each community $x \in X$ we label its nodes with the ground truth community $y \in Y$ they belong to. We then match community x with the ground truth community with the highest number of labels in the algorithm community. This procedure produces (x,y) pairs having the highest homophily between the node labels in x and all the ground truth communities. We then measure the quality of the mappings by the two following measures:

- *Precision*: the percentage of nodes in x labeled as y, computed as

$$P = \frac{|x \cap y|}{|x|} \in [0,1] \tag{2}$$

- *Recall*: the percentage of nodes in y covered by x, computed as

$$R = \frac{|x \cap y|}{|y|} \in [0,1]. \tag{3}$$

Given a pair (x, y) the two measures describe the overlap of their members: a perfect match is obtained when both *precision* and *recall* are 1. We thus have a many-to-one mapping: multiple communities in X can be connected to a single ground truth community in Y. This policy enables the adoption of the proposed methodology both in case of algorithms producing crisp partitions or algorithm producing overlapping communities. Moreover, analyzing the *precision* and *recall* of each pair we are able to detect both underestimations and overestimations made by the adopted algorithm.

We can combine *precision* and *recall* into their harmonic mean obtaining the $F1$-measure, a concise quality score for the individual pairing:

$$F1 = 2\frac{precision * recall}{precision + recall}. \tag{4}$$

Given a network, the $F1$ score can be averaged among all the identified pairs in order to summarize the overall correspondence between the algorithm community set and ground truth community set. The mean $F1$, along with its standard deviation, makes possible to compare the performances of different algorithms on the same network with ground truth communities. The proposed approach as complexity $O(|V| + |C|) \simeq O(|V|)$ since it is composed by two steps: (i) node labeling (linear in the number of nodes $|V|$) and (ii) communities F1-computation (linear in the number of identified communities $|C|$). The averaging of community F1s has constant cost.

4 Experiments

In this section we evaluate the proposed methodology on the community sets produced by popular community detection algorithms on large-scale real-world networks with ground truth communities. In Section 4.1 we introduce the algorithms used and in Section 4.3 we evaluate the quality of the algorithms by using the proposed approach[1].

4.1 Community detection algorithms

We use six different community detection algorithms designed to maximize different functions: LOUVAIN, INFOHIERMAP, CFINDER, DEMON, ILCD and EGO-NETWORK.

LOUVAIN is an heuristic method based on modularity optimization [13] and it is proven to be fast and scalable on large-scale networks. The modularity optimization is performed in two steps. First, the method searches for "small" communities by

[1] A Python implementation of our approach is available at: http://goo.gl/kWIH2I

optimizing modularity locally. Second, it aggregates nodes belonging to the same community and builds a new network whose nodes are communities. These steps are repeated iteratively until a maximum modularity is obtained, producing a complete non-overlapping partitioning of the graph. As most of the approaches based on modularity optimization, it suffers from a "scale" problem that causes the extraction of few huge communities and an high number of tiny ones.

INFOHIERMAP is one of best performing hierarchical non-overlapping clustering algorithms for community detection [14] studied to optimize community conductance. The graph structure is explored with a number of random walks of a given length and with a given probability of jumping into a random node. The underlying intuition is that random walkers are trapped in a community and exit from it very rarely. Each walk is described as a sequence of steps inside a community followed by a jump. By using unique names for communities and reusing a short code for nodes inside the community, this description can be highly compressed, in the same way as re-using street names (nodes) inside different cities (communities). The renaming is done by assigning a Huffman coding to the nodes of the network. The best network partition will result in the shortest description for all the walks.

CFINDER is an algorithm for finding dense overlapping groups of nodes in networks, based on the Clique Percolation Method (CPM) [15]. Its community definition is based on the observation that a typical member in a community is linked to many other members, but not necessarily to all other nodes in the community. In other words, a community can be interpreted as a union of smaller complete subgraphs that share nodes. These complete subgraphs are called k-cliques, where k is the number of nodes in the subgraph, and a k-clique-community is defined as the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques. Two k-cliques are said to be adjacent if they share $k-1$ nodes.

DEMON is an incremental algorithm that uses an approach based on the extraction of ego networks, that is, the set of nodes connected with a certain ego node u [16]. The communities are extracted by using a bottom-up approach: each node gives the perspective of the communities surrounding it and then all the different perspectives are merged together in an overlapping structure. In practice, the ego network of each node is extracted and a label propagation is performed on this structure ignoring the presence of the ego itself, since it will be judged by its peer neighbors. Then, with equity, the vote of everyone in the network is combined. The result of this combination is a set of overlapping modules, the guess of the real communities in the global system, made not by an external observer, but by the actors of the network itself.

ILCD is an algorithm for the detection of overlapping communities in dynamic networks. It can also be used on static networks and works on large-scale networks. It is not based on the modularity, but, on the contrary, on the idea that communities are defined locally (intrinsic communities) [17].

EGO-NETWORKS is a naive algorithm that models the communities as the set of induced subgraphs obtained considering each node with its neighbors. This approach provides the highest overlap among the considered approaches: each node u belongs exactly to $|\Gamma(u)|+1$ communities, where $\Gamma(u)$ identifies its neighbors set.

4.2 Network data

We use four large-scale network datasets in our experiments: DBLP, Youtube, Amazon and LiveJournal [18], filtering them on the nodes covered by the ground truth partition (network statistics shown in Table 1).[2].

The DBLP network is a co-authorship network where two authors of computer science papers are connected if they publish at least one paper together. The ground truth communities are defined by the publication venue, e.g. journal or conference, hence authors who published to a certain journal or conference form a community.

Youtube is a popular video-sharing website where the users form friendships each other and can create groups which other users can join. The user-defined groups are the ground truth communities of the network.

The Amazon network has been collected by crawling Amazon website. It is based on Customers-Who-Bought-This-Item-Also-Bought feature of the Amazon website. If a given product i is frequently co-purchased with product j, the graph contains an undirected edge from i to j. Each product category provided by Amazon defines each ground truth community.

LiveJournal is a free online blogging community where users can declare friendships to each other. It also allows users to form a group which other members can then join. Each of these user-defined groups is a ground truth community.

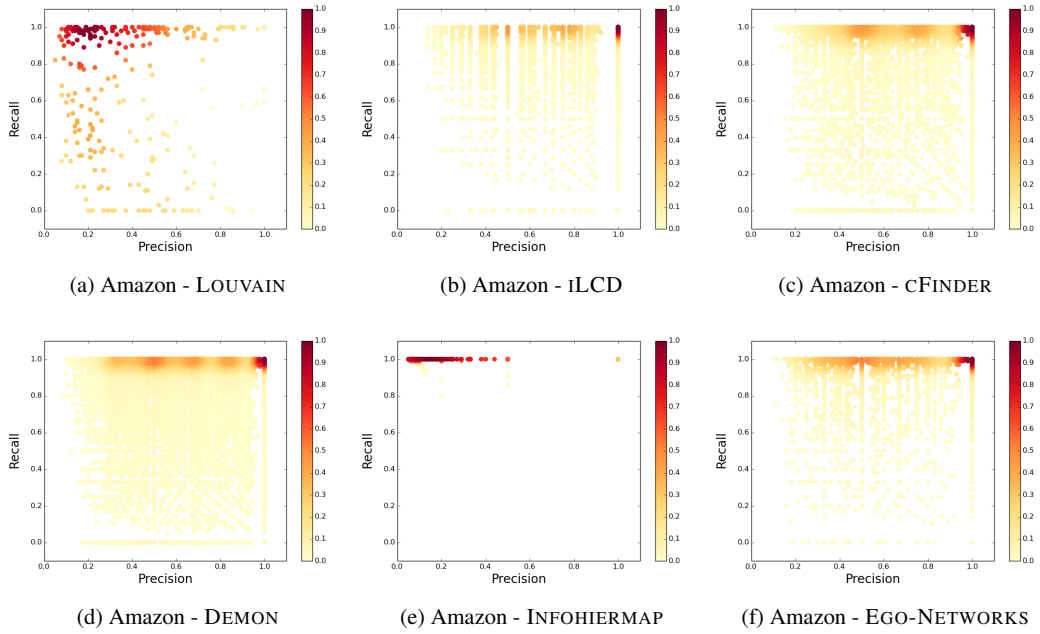
Network	**Nodes**	**Edges**	**Clustering**	**Diameter**	**ground truth com.**
Amazon	334,863	925,872	0.3967	44	75,149
DBLP	317,080	1,049,866	0.6324	21	13,477
Youtube	1,134,890	2,987,624	0.0808	20	8,385
LiveJournal	3,997,962	34,681,189	0.2843	17	287,512

Table 1: Networks Statistics of the four large-scale real-world networks analyzed.

4.3 Results

We apply the six algorithms introduced in Section 4.1 to extract communities from the four large-scale network datasets described in the Section 4.2. We then use the proposed evaluation approach to compare the obtained community sets and rank the tested algorithms. Figures 1, 2, 3 and 4 show the density scatter plots describing community precision and community recall computed on the six community sets produced on the Amazon, DBLP, Youtube and LiveJournal networks respectively. In this representation, we report the community precision on the x-axis and the community recall on the y-axis: the color of a point (x,y) in a scale from yellow to

[2] The network datasets are available at: https://snap.stanford.edu/data/



(a) Amazon - LOUVAIN (b) Amazon - ILCD (c) Amazon - CFINDER

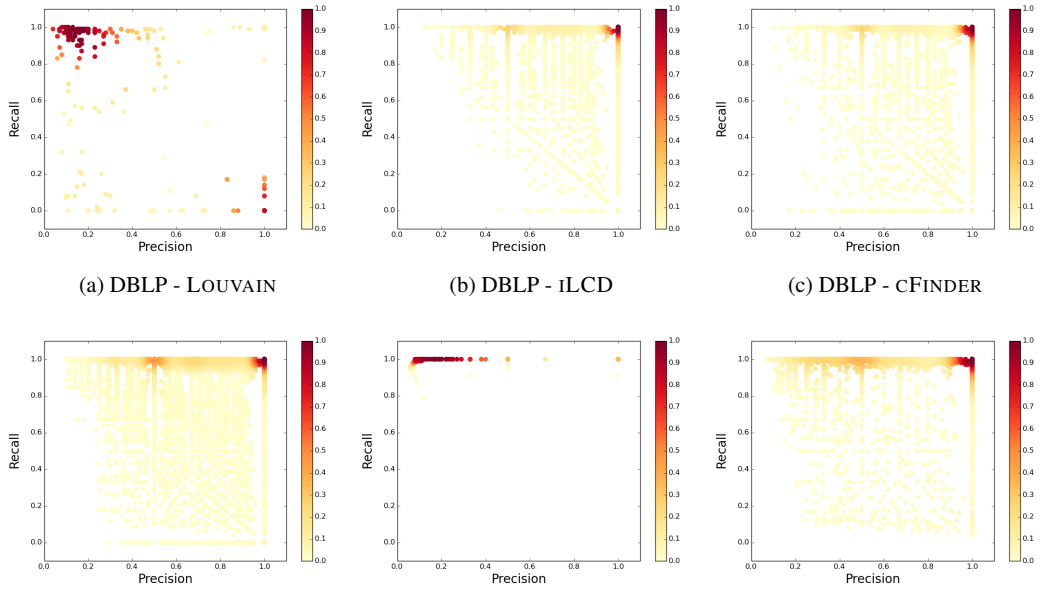
(d) Amazon - DEMON (e) Amazon - INFOHIERMAP (f) Amazon - EGO-NETWORKS

Fig. 1: Density scatter plots describing community precision and community recall on the six community sets extracted from the Amazon network.

red indicates the number of community matchings having precision x and recall y: the more red is the color the higher is the volume. We have a perfect match when both precision and recall are 1 (top-right corner of the plot): in this scenario, the algorithm community is identical to the corresponding ground truth community. The proposed visualization also allows an intuitive identification of the community scale:

- pairings having maximal recall and low precision (i.e. points that clusters close to the upper left corner of the plot) identifies network substructures that overestimate the ground truth;
- pairings having low recall and maximal precision (i.e. points that clusters close to the lower right corner) identifies network substructures that underestimate the ground truth.

The former scale tells us that the algorithm produces communities that group together more nodes than it should, while in the latter case the ground truth communities are fragmented in smaller communities. From the plots, for the Amazon and DBLP networks a difference among the algorithms clearly emerges: while DEMON, ILCD, CFINDER and EGO-NETWORKS produce community sets with high precision and high recall denoting a high correspondence to the ground truth communities, LOUVAIN and INFOHIERMAP produce community sets with low precision



(a) DBLP - LOUVAIN (b) DBLP - ILCD (c) DBLP - CFINDER

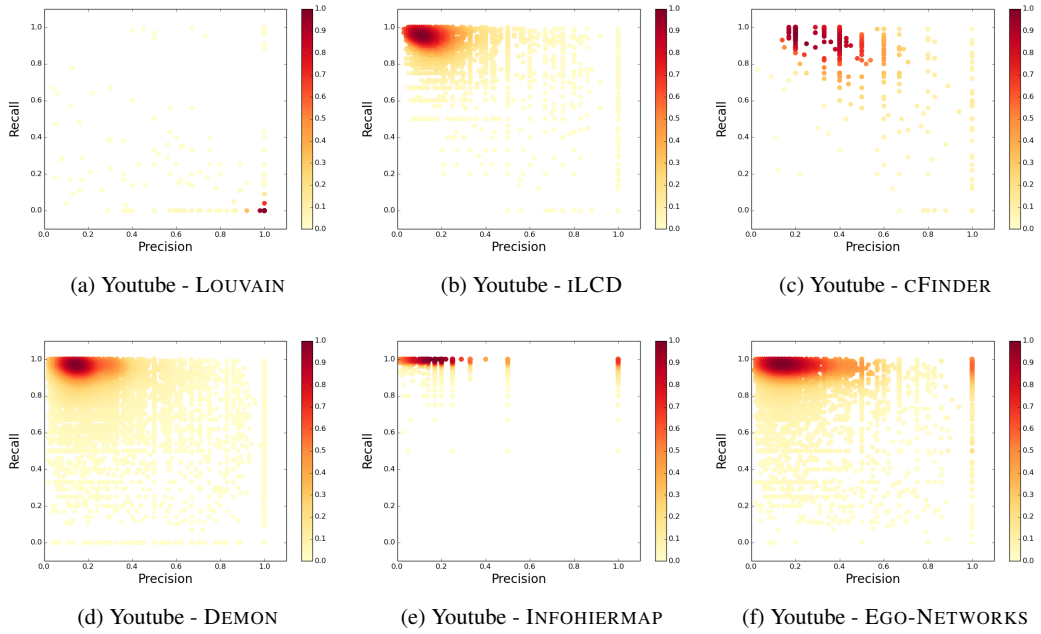
(d) DBLP - DEMON (e) DBLP - INFOHIERMAP (f) DBLP - EGO-NETWORKS

Fig. 2: Density scatter plots describing community precision and community recall on the six community sets extracted from the DBLP network.

Network	LOUVAIN	INFOHIERMAP	CFINDER	DEMON	ILCD	EGO-NETWORKS
Amazon	.40 (.26)	.46 (.29)	.72 (.27)	.70 (.24)	**.74 (.23)**	.72 (.22)
DBLP	.26 (.24)	.45 (.31)	**.82 (.24)**	.75 (.24)	.81 (.23)	.81 (.22)
Youtube	.16 (.05)	**.59 (.32)**	.50 (.20)	.36 (.10)	.35 (.20)	**.58 (.28)**
LiveJournal	.01 (.06)	.66 (.30)	.21 (.30)	.56 (.29)	**.71 (.04)**	.52 (.30)

Table 2: The average $F1$-measure for the four networks. Each row shows the average $F1$-measure (standard deviation within brackets) achieved when matching the communities identified by the algorithms and the ground truth communities of a specific network. In bold the best score for each network.

(low label homophily) and high recall (they cover a large fragment of the corresponding ground truth community). On the Youtube network LOUVAIN shows very high precision and very low recall, while the other algorithms behave the opposite producing communities with low precision and high recall. On the LiveJournal network all the algorithms produce communities with high precision, while the recall varies a lot across the communities. Table 2 summarizes all these observation reporting, for each algorithm and dataset, the average $F1$-measure computed on the identified pairings. We observe how the average $F1$-measure is useful to understand two main aspects of community evaluation:



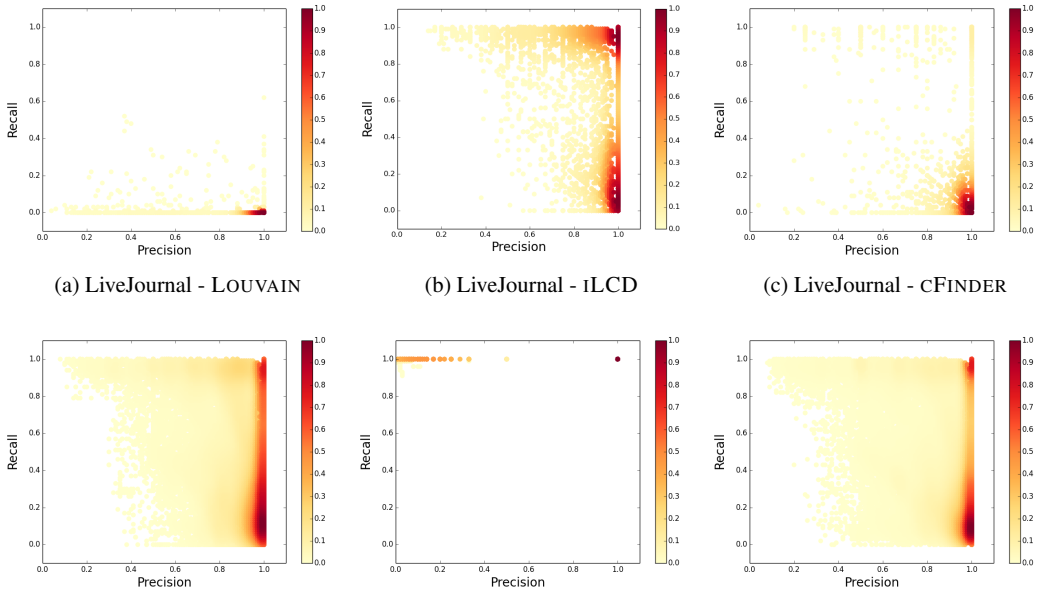
(a) Youtube - LOUVAIN (b) Youtube - ILCD (c) Youtube - CFINDER

(d) Youtube - DEMON (e) Youtube - INFOHIERMAP (f) Youtube - EGO-NETWORKS

Fig. 3: Density scatter plots describing community precision and community recall on the six community sets extracted from the Youtube network.

- First, it summarizes how well the communities produced by an algorithm corresponds to the ground truth communities. For instance, from our experiments is clear how LOUVAIN shows lower correspondence with the ground truth than all the other algorithms: this result is clearly due to the so called *scale* problem of modularity based approaches. Indeed, as shown from all the density scatter plots, LOUVAIN produces either huge or tiny communities thus providing respectively an overestimation (high recall, low precision – i.e. Amazon and DBLP) or a underestimation (low recall, high precision – i.e. LiveJournal and Youtube) of the ground truth communities;
- Second, the $F1$-measure helps also in evaluating the quality of the ground truth itself: on the Youtube dataset for example no algorithm produces communities with high correspondence with the ground truth ones, denoting either a low quality of the ground truth communities or that the community definition underlying the ground truth radically differs from the community definition of the tested algorithms.

However the $F1$-measure indicator is computed as an average of community-pairs $F1$s and it can show a high standard deviation (Table 2). For this reason we report in Figure 5, for each network and algorithm, the complete distribution of $F1$ across



(a) LiveJournal - LOUVAIN (b) LiveJournal - ILCD (c) LiveJournal - CFINDER

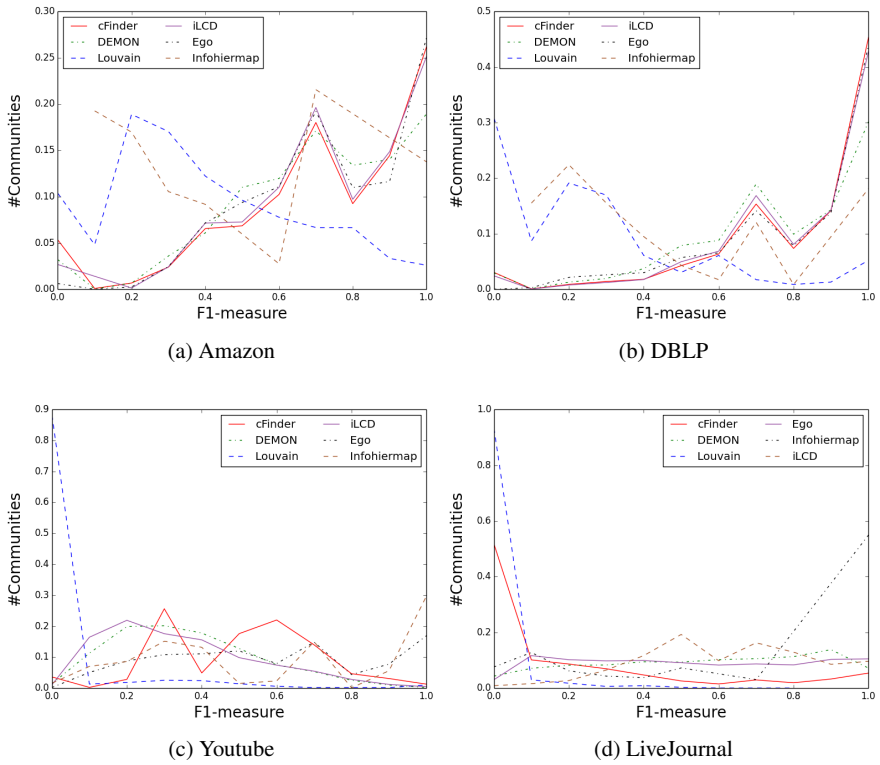
(d) LiveJournal - DEMON (e) LiveJournal - INFOHIERMAP (f) LiveJournal - EGO-NETWORKS

Fig. 4: Density scatter plots describing community precision and community recall on the six community sets extracted from the LiveJournal network.

the community pairs. We observe how the distributions endorse the validity of the proposed indicator even in presence of high standard deviation.

5 Conclusion

Evaluating the quality of community detection algorithms is a hard task, especially because the problem itself is ill-posed: each algorithm optimizes a different quality metric introducing its own community definition. In this paper we tackled the problem of estimating the correspondence between algorithm communities and ground truth communities. When available, the information about ground truth communities of a network can be used to compare the results provided by a set of algorithms: so far the NMI has been the common way to perform this task. However, NMI has a major drawback: its computational complexity is quadratic in the number of communities. For this reason we introduced a novel and fast approach to estimate the quality of the communities produced by an algorithm that can be applicable to large-scale networks. With the support of visual tools, our methodology provides a reliable index that captures the quality of a community set and describes if the adopted



(a) Amazon (b) DBLP

(c) Youtube (d) LiveJournal

Fig. 5: Distribution of F1-measure on the community pairings generated by the six algorithms the Amazon (a), DBLP (b), Youtube (c) and LiveJournal (d) networks.

algorithm underestimates or overestimates the ground truth community structure. As future works, we plan to use the proposed approach to identify and characterize sub-profiles among the communities extracted: by applying clustering techniques using precision and recall as features, we can group communities according to their degree of correspondence to the ground truth and then study their network features.

Acknowledgements This work was partially funded by the European Community's H2020 Program under the funding scheme "FETPROACT-1-2014: Global Systems Science (GSS)", grant agreement #641191 CIMPLEX "Bringing CItizens, Models and Data together in Participatory, Interactive SociaL EXploratories", https://www.cimplex-project.eu.

Our research is also supported by the European Community's H2020 Program under the scheme "INFRAIA-1-2014-2015: Research Infrastructures", grant agreement #654024 *"SoBigData: Social Mining & Big Data Ecosystem"*, http://www.sobigdata.eu.

References

1. S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75 – 174, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6TVP-4XPYXF1-1/2/99061fac6435db4343b2374d26e64ac1
2. M. Coscia, F. Giannotti, and D. Pedreschi, "A classification for community discovery methods in complex networks," *Stat. Anal. Data Min.*, vol. 4, no. 5, pp. 512–546, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1002/sam.10133
3. A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E*, vol. 78, no. 4, p. 046110, Oct. 2008. [Online]. Available: http://pre.aps.org/abstract/PRE/v78/i4/e046110
4. S. Bhat and M. Abulaish, "Overlapping social network communities and viral marketing," in *International Symposium on Computational and Business Intelligence*, Aug 2013, pp. 243–246.
5. X. Wu and Z. Liu, "How community structure influences epidemic spread in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, pp. 623–630, 2008.
6. G. Rossetti, L. Pappalardo, R. Kikas, D. Pedreschi, F. Giannotti, and M. Dumas, "Community-centric analysis of user engagement in skype social network," in *Proceedings of the 2015 ACM/IEEE International Conference on Advances in Social Network Analysis and Mining*, 2015.
7. G. Rossetti, R. Guidotti, D. Pennacchioli, D. Pedreschi, and F. Giannotti, "Interaction prediction in dynamic networks exploiting community discovery," in *Proceedings of the 2015 ACM/IEEE International Conference on Advances in Social Network Analysis and Mining*, 2015.
8. S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, F. Giannotti, and D. Pedreschi, "Discovering the geographical borders of human mobility," *KI - Künstliche Intelligenz*, 2012.
9. J. P. Bagrow and Y.-R. Lin, "Mesoscopic structure and social aspects of human mobility," *PLoS ONE*, vol. 7, no. 5, p. e37676, May 2012. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0037676
10. A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Phys. Rev. E*, vol. 80, no. 1, p. 016118, Jul. 2009.
11. A. F. McDaid, D. Greene, and N. J. Hurley, "Normalized mutual information to evaluate overlapping community finding algorithms," *CoRR*, vol. abs/1110.2515, 2011.
12. "Detecting the overlapping and hierarchical community structure in complex networks," *New J. Phys. p*, 2009.
13. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.
14. M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.
15. G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, June 2005.
16. M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: a local-first discovery method for overlapping communities." in *KDD*, Q. Y. 0001, D. Agarwal, and J. Pei, Eds. ACM, 2012, pp. 615–623.
17. R. Cazabet, F. Amblard, and C. Hanachi, "Detection of overlapping communities in dynamical social networks," in *SocialCom*, 2010, pp. 309–314.
18. Y. Jaewon and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, 2015.