CEIL: A Scalable, Resolution Limit Free Approach for Detecting Communities in Large Networks

Vishnu Sankar M
IIT Madras
Chennai, India
vishnusankar1512gmail.com

Balaraman Ravindran
IIT Madras
Chennai, India
ravicse.iitm.ac.in

Shivashankar S
Ericsson Reasearch
Chennai, India
s.shivashankarericsson.com

Abstract

Real world networks typically exhibit non uniform edge densities with there being a higher concentration of edges within modules or communities. Various scoring functions have been proposed to quantify the quality of such communities. In this paper, we argue that the popular scoring functions suffer from certain limitations. We identify the necessary features that a scoring function should incorporate in order to characterize good community structure and propose a new scoring function, CEIL (Community detection using External and Internal scores in Large networks), which conforms closely with our characterization. We also demonstrate experimentally the superiority of our scoring function over the existing scoring functions.

Modularity, a very popular scoring function, exhibits resolution limit, i.e., one cannot find communities that are much smaller in size compared to the size of the network. In many real world networks, community size does not grow in proportion to the network size. This implies that resolution limit is a serious problem in large networks. Modularity is still very popular since it offers many advantages such as fast algorithms for maximizing the score, and non-trivial community structures corresponding to the maxima. We show analytically that the CEIL score does not suffer from resolution limit. We also modify the Louvain method, one of the fastest greedy algorithms for maximizing modularity, to maximize the CEIL score. We show that our algorithm gives the expected communities in synthetic networks as opposed to maximizing modularity. We also show that the community labels given by our algorithm matches closely with the ground truth community labels in real world networks. Our algorithm is on par with Louvain method in computation time and hence scales well to large networks.

1 Introduction

Networks are the natural form of representation of the interactions between real world objects. Most of these real world networks exhibit the property of community structure [Newman and Girvan, 2002], i.e., nodes in the network can be partitioned into communities such that more edges are present between nodes belonging to the same community than between nodes belonging to different communities.

The quality of communities given by community detection algorithms is measured using scoring functions. Though several scoring functions have been proposed in the literature, Newman's modularity is the most widely used scoring function. The exact maximization of modularity is hard [Brandes *et al.*, 2006]. So, several heuristics were proposed to maximize modularity. Newman proposed a greedy maximization of modularity [Newman, 2004] to find communities. A randomized version of the greedy maximization of modularity and an ensemble method of modularity maximization was proposed by Ovelgonne and Geyer-Schulz [Ovelgnne and Geyer-Schulz, 2010]. Later, they improved the performance of this algorithm by introducing iterative ensemble method to maximize modularity [Ovelgonne and Geyer-Schultz, 2013]. A very fast approach to maximize modularity, widely known as Louvain method, was introduced by Blondel et al. [Blondel *et al.*, 2008]. Since this significantly outperformed other methods on time it remains very popular. Other than the modularity maximization algorithms, Raghavan et al. [Raghavan *et al.*, 2007] proposed a near linear label propagation algorithm to find communities in large networks. Pons and Latapy proposed a random walk based approach to find the similarity between nodes and hence find communities in the network [Pons and Latapy, 2006]. Palla et al. proposed the clique percolation technique [Palla *et al.*, 2005] which is mainly used to find overlapping communities.

Though several algorithms find better communities than Louvain method, it is arguably the fastest deterministic approach to find communities in a network. Its only limitation is the resolution limit of modularity. Conductance [Shi and Malik, 2000] is another popular scoring function used to measure the quality of communities. But, optimization of conductance will always give one community which is the graph itself. A comparison of several scoring functions against ground-truth communities revealed that none of them can be accepted as a single best scoring function since every scoring function captures only certain features of the community [Yang and Leskovec, 2012]. So, there is a need for a scoring function which captures the intuition behind communities, which can be used as an objective function to find communities and which can be optimized efficiently.

We propose a new scoring function, known as CEIL (Community detection using External and Internal score in Large networks) score, which is free from resolution limit and can be maximized using the same Louvain heuristics to find communities in a network.

The contributions of this paper are :

- Identification of the necessary features that a good scoring function should incorporate in order to find the relative ordering of communities in a network (Section 2.1).
- Introduction of a new scoring function (CEIL score) which captures all the features of communities (Section 2.3).
- A theoretical proof that CEIL score does not suffer from resolution limit (Section 3). A comparative study of several scoring functions indicate the versatility of our scoring function (Section 4).
- Adaptation of the Louvain algorithm to greedily optimize the CEIL score (Section 5) and empirical verification that the community labels given by this modified algorithm closely match the ground-truth labels (Section 6).

2 Introduction of CEIL score

The intuition behind a community is that it is well connected internally and well separated from the rest of the network, i.e., it should have more intra-community edges and fewer inter-community edges. A scoring function is the mathematical formulation of the quality of a community.

2.1 Characterization of a good scoring function

A community is characterized by the following three features.

- Number of nodes forming the community.
- Number of intra-community edges present in the community.
- Number of inter-community edges incident on the community.

A scoring function should incorporate all these three features of the communities because numerous examples can be constructed such that there are two communities which differ in only one of the above three features. In the such cases, the "third" feature becomes necessary to distinguish between the two communities. For example, consider a scoring function which does not include the feature 'number of nodes forming the community'. It will give same score to a community A in a network having 10 intra-community edges, 2 inter-community edges and 5 nodes and to a community B in the same network having 10 intra-community edges, 2 inter-community edges and 10 nodes. But community A has higher internal density than community B and should be given higher score than community B. Similarly, the other two features are also essential in assigning scores to communities in order to find the relative order of communities in a network.

We find that none of the existing scoring functions take into account all the three necessary features to characterize the communities. For e.g., Modularity [Newman, 2004] and conductance [Shi and Malik, 2000] do not consider the parameter 'number of nodes forming the community' while Triad Participation Ratio does not consider the parameter 'number of inter-community edges incident on the community'. So, we propose a new scoring function (CEIL score) characterizing the communities by taking into account all the necessary features.

2.2 Notations

Let G be a simple unweighted network with N nodes and E edges. Let S be the set of all communities in G and C be the size of S. We denote one of the communities in G by s and a_s is the number of intra-community edges in s, b_s is the number of inter-community edges incident on s and n_s is the number of nodes in s.

2.3 Existing scoring functions

Few of the widely used scoring functions are,

Definition 2.1. Conductance is given by the ratio of number of inter-community edges to the total degree of all nodes in the community [Shi and Malik, 2000].

Definition 2.2. Triad Participation Ratio is given by the fraction of nodes in the community that are participating in at least one triad.

Definition 2.3. Modularity is given by the difference between the actual number of intra-community edges present in the community and the expected number of intra-community edges that would be present in the community if the edges in the network were distributed randomly with identical degree distribution [Newman, 2004].

2.4 CEIL score

A community is said to be well connected internally if a large fraction of the pairs of nodes belonging to the community are connected, i.e., the internal well connectedness of a community increases as the number of intra-community edges increases.

Definition 2.4. The internal score of a community s is the internal edge density of that community.

$$\text{Internal Score}(s) = \begin{cases} \frac{a_s}{\binom{n_s}{2}} & \text{if } n_s \geq 2 \\ 0 & \text{if } n_s = 1 \end{cases}$$

Internal score ranges from 0 to 1. It takes the value of 0 when there are no intra-community edges in the community and takes the value of 1 when every node in the community is connected to every other node in the community.

A community is said to be well separated from rest of the network, if the number of inter-community edges is less, i.e., the external well separability increases as the number of inter-community edges decreases.

Definition 2.5. The external score of a community s is the fraction of total number of edges incident on that community that are intra-community edges.

$$\text{External Score}(s) = \frac{a_s}{a_s + b_s}$$

External score ranges from 0 to 1. It takes the value of 0 when every edge incident on the community is an inter-community edge and takes the value of 1 when every edge which is incident on the community is an intra-community edge.

A community structure is said to be good, if it is well connected within itself and is well separated from rest of the network, i.e., if it has high internal and external scores.

Definition 2.6. The CEIL score of a community s is the product of internal and external score of that community.

$$\text{CEIL Score}(s) = \text{Internal Score}(s) \times \text{External Score}(s)$$

Since we are interested in the relative order of communities and not on the absolute score a community gets, we chose product over geometric mean. CEIL score of a community ranges from 0 to 1. It takes the value of 0 when the community has no intra-community edge, and takes the value of 1 when the community is a clique that is isolated from the rest of the network. A high score represents a good community as it can be obtained only when both the internal and external scores are high.

A network typically has many communities.

Definition 2.7. The CEIL score of a network G is the weighted sum of scores of all the communities in that network.

$$\text{CEIL Score}(G) = \sum_{s \in S} \frac{n_s}{N} \times \text{Community Score}(s)$$

CEIL score of a network ranges from 0 to 1 in a simple, unweighted network. Computationally, CEIL score has the same complexity as that of modularity or conductance.

3 Resolution Limit

Resolution limit is the inability of an algorithm to detect communities which are much smaller in size when compared to the size of the network.

Theorem 3.1. *CEIL score does not suffer from resolution limit.*

Proof. Without loss of generality, let us consider the same network which was used to prove the resolution limit of modularity maximization. The network has two communities, 1 and 2, connected to each other as well as to the rest of the network as in figure 1. Let n_1 and a_1 be the number of nodes and the number of intra-community edges of community 1, respectively. Similarly, let n_2 and a_2 be the number of nodes and the number of intra-community edges of community 2. In order to express the number of inter-community edges of both the communities in terms of their respective intra-community edges, we consider four positive constants x_1, y_1, x_2 and y_2. Let $x_1 a_1$ represent the number of inter-community edges going from community 1 to community 2 and $y_1 a_1$ represent the number of inter-community edges going from community 1 to rest of the network. Similarly, let $x_2 a_2$ represent the number of inter-community edges going from community 2 to community 1 and $y_2 a_2$ represent the number of inter-community edges going from community 2 to rest of the network. Let N represent total number of nodes in the network.

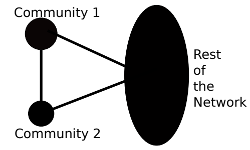


Figure 1: Outline of a network with at least 3 communities

We will consider two different partitions of this network. Partition A, in which 1 and 2 are considered as two different communities and Partition B, in which 1 and 2 are considered as a single community. The partition of the rest of the network can be done in anyway but identical in partitions A and B. Let N_A and N_B be the network scores of partitions A and B respectively. Let N_0 be the network score of rest of the network.

$$N_A = N_0 + \frac{n_1}{N}\left(\frac{a_1}{a_1 + x_1 a_1 + y_1 a_1}\right)\left(\frac{a_1}{\binom{n_1}{2}}\right) + \frac{n_2}{N}\left(\frac{a_2}{a_2 + x_2 a_2 + y_2 a_2}\right)\left(\frac{a_2}{\binom{n_2}{2}}\right)$$

$$N_B = N_0 + \frac{n_1 + n_2}{N}\left(\frac{a_1 + x_1 a_1 + a_2}{a_1 + x_1 a_1 + a_2 + y_1 a_1 + y_2 a_2}\right)\left(\frac{a_1 + x_1 a_1 + a_2}{\binom{n_1+n_2}{2}}\right)$$

For 1 and 2 to be separate communities, N_A should be greater than N_B, i.e.,

$$\left(\frac{n_1 a_1^2}{(a_1 + x_1 a_1 + y_1 a_1)\binom{n_1}{2}}\right) + \left(\frac{n_2 a_2^2}{(a_2 + x_2 a_2 + y_2 a_2)\binom{n_2}{2}}\right) > \left(\frac{(n_1 + n_2)(a_1 + x_1 a_1 + a_2)^2}{(a_1 + x_1 a_1 + a_2 + y_1 a_1 + y_2 b_2)\binom{n_1+n_2}{2}}\right)$$

Note that the above inequality does not depend on any parameter which is related to size of the network. This means that CEIL score does not suffer from resolution limit. □

A similar analysis for modularity leads to the parameters of communities on one side of the inequality with number of edges in the network on the other side [Fortunato and Barthélemy, 2007] which means that the decision to split or merge two communities will depend on the size of the network.

4 Comparison of scoring functions

Yang and Leskovec [Yang and Leskovec, 2012] compared the performance of several scoring functions using perturbation experiments and reported that Conductance and Triad Participation Ratio are the best performers. Modularity is the widely used definition. So, we compare CEIL score against these three scoring functions.

The networks with ground-truth communities which we have used are listed in table 1. We obtained them from `http://snap.stanford.edu/data`

Table 1: Networks with ground-truth communities

Networks	Nodes	Edges
LiveJournal	3,997,962	34,681,189
Youtube	1,134,890	2,987,624
DBLP	317,080	1,049,866
Amazon	334,863	925,872

4.1 Community Goodness Metrics

Following the same notation as 2.2 of this paper, the goodness metrics that we choose are as below.

$$\text{Internal density}(s) = \frac{a_s}{n_s\,(n_s - 1)}$$

$$\text{Separability}(s) = \frac{a_s}{b_s}$$

$$\text{Clustering Coefficient}(s) = \frac{\text{Number of closed triplets}}{\text{Number of connected triplets}}$$

We rank the ground-truth communities based on goodness metrics as well as scoring functions. We measure the correlation of the ranks given by scoring functions and goodness metrics. We use spearman's rank correlation coefficient to find the correlation between the ranks [Myers and Well, 2003].

Table 2: Spearman's rank correlation coefficient for density

Networks	LiveJournal	Youtube	DBLP	Amazon
Modularity	-0.3751	-0.9017	-0.2313	-0.9070
Conductance	0.1963	0.5762	0.1736	-0.5676
TPR	0.4386	-0.5124	0.4052	0.4714
CEIL score	**0.5363**	**0.8279**	**0.7034**	**0.9474**

Table 3: Spearman's rank correlation coefficient for separability

Networks	LiveJournal	Youtube	DBLP	Amazon
Modularity	0.0600	-0.4854	0.0687	0.5791
Conductance	**1.0000**	**1.0000**	**1.0000**	**1.0000**
TPR	-0.0482	-0.4782	-0.0240	-0.3891
CEIL score	0.9002	0.9192	0.7954	-0.3513

From tables 2, 3 and 4, we have the following conclusions. Modularity does not correlate well with goodness metrics. Conductance absolutely correlates with separability but it doesn't correlate well with other two goodness metrics. Triad participation ratio overall has the second best correlation with density and clustering coefficient but it do not correlate well with separability. CEIL score has the highest correlation with density, second highest correlation with separability except in Amazon network and highest correlation with clustering coefficient except in Youtube network. In the amazon network, communities with high internal density have low separability and vice versa. This is the reason for the negative correlation of CEIL score with separability in Amazon network. This correlation experiment also shows that the poor correlation of conductance with density is due to the fact that it do not consider the feature 'number of nodes forming the community'. Similarly, the poor correlation of triad participation ratio with separability is due to the fact that it does not consider the feature 'number of inter community edges incident on the community'. Since, CEIL score takes into account all the features, it correlates well with all the goodness metrics.

Table 4: Spearman's rank correlation coefficient for clustering coefficient

Networks	LiveJournal	Youtube	DBLP	Amazon
Modularity	-0.1991	0.6738	-0.0559	-0.8675
Conductance	0.1754	-0.4093	0.1433	-0.5316
TPR	0.3142	**0.8707**	0.3399	0.4148
CEIL score	**0.4841**	-0.4278	**0.6332**	**0.9152**

4.2 Perturbation Experiment

Perturbation experiments were introduced for comparative study of different scoring functions [Yang and Leskovec, 2012]. In these experiments, ground-truth communities are perturbed using few perturbation techniques to degrade their quality. A good scoring function is expected not only to give high scores for ground-truth communities and is also expected to give low scores for perturbed communities.

We consider a ground-truth community s and do the NODESWAP, RANDOM, EXPAND and SHRINK perturbations on the community as described in [Yang and Leskovec, 2012].

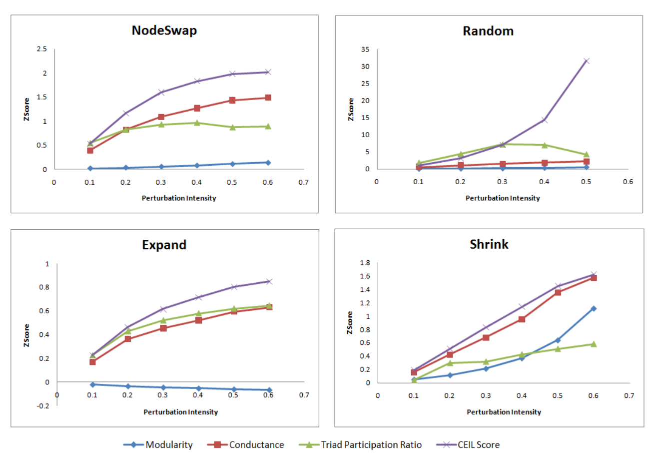


Figure 2: Z-score given by various scoring functions as a function of perturbation intensity for the LiveJournal network.

Figure 2 shows the plot of Z-score (normalized difference between the scores of unperturbed and perturbed communities) as a function of perturbation intensity in the LiveJournal network. CEIL score performs significantly better in RANDOM and considerably better in NODESWAP and EXPAND. We obtained similar performance in Youtube and DBLP datasets also. In SHRINK, we obtained mixed results as CEIL score does better in LiveJournal network while modularity does better in Youtube network and conductance does better in DBLP network. Since the rank correlation between density and separability is negative in Amazon network unlike the other networks and the ground-truth communities have poor density and better separability, Conductance performs well in all the different perturbations except RANDOM in which CEIL score performs better.

5 CEIL Algorithm

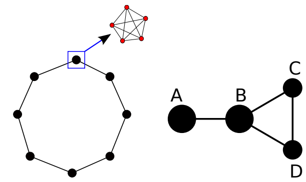
A greedy approach to find communities by efficiently maximizing an objective function is already proposed in [Blondel *et al.*, 2008]. Since it is the fastest known heuristic, we use the same method to maximize CEIL score. In this section, we provide a complete description of the algorithm. A pseudocode is omitted due to lack of space. The algorithm has two phases. In the first phase, we assign each node to its own community. Then, we consider each node in the network in a sequential manner, remove it from its original community and add it either to the community of one of its neighbors or back to the original community, whichever will result in the greatest increase in CEIL score of the network. The newer properties of a community when a node n is added to the community is calculated as,

$$
\begin{aligned}
a_s &= a_s + \text{intra}_n + \text{incident}_{n,s} \\
\text{deg}_s &= \text{deg}_s + \text{deg}_n \\
n_s &= n_s + n_n
\end{aligned}
$$

where intra_n is the number of intra-community edges in the community represented by node n, $\text{incident}_{n,s}$ is the sum of weights of the edges incident from node n to community s, deg_s is the sum of degree of all nodes in the community s, deg_n is the sum of degree of all nodes in the community represented by node n and n_n is the number of nodes in the community represented by node n.

With the updated a_s, deg_s and n_s, the newer score and hence the increase is calculated. In a similar way, the decrease in the score of a community when a node is removed from it is calculated. We repeat this process iteratively until there is no increase in the score given by the scoring function. At this time, the scoring function has reached its local maxima and the first phase ends.

In the second phase, we construct an induced graph of the network by using the community labels of nodes obtained from the first phase. Each community in the first phase is represented by a node in the induced graph. Number of nodes in the community, sum of degree of all nodes in the community and the number of intra-community edges are all preserved in the induced graph by associating them with the respective node. The weight of an edge between two nodes in the induced graph is equal to the sum of weights of all the edges between those two communities in the original graph. The second phase ends after the construction of the induced graph.



(a) COC (b) POC

Figure 3: COC - Circle Of Cliques and POC - Pair Of cliques

We keep track of the scores of all the communities, i.e., update the scores of the communities as and when a change (addition or deletion of nodes) is made to the community. This will help in faster calculations of the increase or decrease of the community score whenever a change is made to the community. The size of the graph is reduced drastically after the second phase that even a million nodes graph gets reduced to thousands of nodes in few iterations. This contributes significantly in reducing the running time of the algorithm.

The induced graph obtained as output of the second phase is given as the input to the first phase. The two phases are thus iterated until there is no increase in the score.

In weighted networks, we calculate the number of edges as the sum of weights on all the edges. CEIL score of the network takes the low value of 0 but the high value is dependent on weights on the edges of the network. Nevertheless, the relative ordering of communities in a network will not get affected and so we can use CEIL algorithm in weighted networks also. By calculating only the edges going out from the nodes belonging to a community while calculating the number of intra- and inter-community edges, we can extend the CEIL algorithm to directed graphs also. We note that we cannot apply CEIL algorithm without modifications to find overlapping communities. But, CEIL score can still be used to rank the overlapping communities.

6 Empirical Validation

In this section, we compare CEIL algorithm to representative algorithms that exhibit good scaling behavior [1].

6.1 Experimental demonstration of resolution limit

In [Fortunato and Barthélemy, 2007], synthetic networks with specific structural properties were used to demonstrate the resolution limit in modularity. We use similar networks to show that the CEIL algorithm finds the expected communities.

Circle of cliques : Figure 3(a) shows an example network consisting of a circle of cliques. In the figure each dot correspond to a clique. The line between any two dots is a single edge connecting the corresponding cliques. Consider such a network of 30 cliques with each clique having 5 nodes. CEIL algorithm detected 30 communities with each clique being a separate community. Louvain method was shown to give

[1] Source code of the CEIL algorithm can be found in `github.com/vishnu1512/CEIL`.

only 15 communities with two adjacent cliques belonging to a single community [Fortunato and Barthélemy, 2007].

To show that CEIL algorithm finds communities in networks irrespective of the number of communities, we repeated this experiment with 300, 3000 and 30000 cliques of size 5. In all the cases, CEIL algorithm was able to find each of the cliques as a separate community. To show that CEIL algorithm finds communities irrespective of the size, we kept the number of communities in this network to a fixed value of 10 and generated this network with size of cliques as 50, 500 and 1000. In all these networks, CEIL algorithm was able to find the correct communities.

Two Pairs of cliques : To show that CEIL algorithm finds communities that differ in size, we generated a network with a pair of big sized cliques and a pair of small sized cliques. In figure 3(b), A and B are cliques of size 20, i.e., 20 nodes and 190 edges while C and D are cliques of size 5. The line between any two dots is the single edge connecting two cliques. CEIL algorithm gave 4 communities with each clique being a community. Louvain method was shown to give only 3 communities [Fortunato and Barthélemy, 2007]. We kept the size of the two small cliques as constant at a size of 5 and generated three networks with the size of the big cliques as 200, 2000 and 5000. In each of these cases, CEIL algorithm is able to find the four cliques as four different communities.

Note that we can construct numerous such examples where modularity maximization fails due to the resolution limit while maximization of CEIL score does not fail. The above experiments also show that CEIL algorithm will be able to identify communities irrespective of their size and number provided that a strong community structure is present in the network.

6.2 Four Community Network

To show that CEIL algorithm performs well on graphs with different densities, we generated synthetic networks consisting of four communities as described in [Danon *et al.*, 2005]. The density of edges in the resulting networks were 25.19%, 12.59%, and 6.29%. In the network with density 25.19%, we were able to recover the 4 ground-truth communities whenever the average number of intra-community links per node (Z_{in}) was greater than the number of inter-community links per node (Z_{out}). In the network with density 12.59%, we were able to identify the communities fully when Z_{out} was 0 and 1. We obtained a rand index of 0.9922 and 0.93 when Z_{out} was 2 and 3. In the network with density 6.29%, we were able to obtain a rand index of 0.9506 when Z_{out} was 0. We note that the ability to recover the communities goes down as the density of edges in the network decreases.

6.3 Real World Graphs

We ran the CEIL algorithm, the Louvain modularity maximization approach and the Label Propagation algorithm [Raghavan *et al.*, 2007] in real world graphs. In the label propagation algorithm, each node in the network is initially assigned an unique label. Then, nodes in the network are visited in a random order and are assigned the label which majority of its neighbors have. Ties are broken uniform randomly. This process will be repeated until all the labels of all the nodes in the network agree with the majority of nodes in their neighbourhood.

Table 5: Rand index

Networks	Louvain method	Label Propagation	CEIL algorithm
Youtube	0.8957	0.6915	**0.9959**
DBLP	0.9702	0.9818	**0.9828**
Amazon	0.9910	**0.9953**	0.9938

Table 6: Running Time

Networks	Louvain method	Label Propagation	CEIL algorithm
Youtube	321.25s	**30.03s**	395.25s
DBLP	134.39s	571.69s	**77.77s**
Amazon	81.17s	**10.76s**	80.68s

From table 5, we see that CEIL algorithm captures the ground-truth communities better than Louvain method as measured by rand index [Rand, 1971]. Though the label propagation algorithm gives a better match to ground-truth communities in Amazon network, it performs poorly in Youtube network.

We note from table 6 that the running time of CEIL algorithm is on par with the Louvain method. Only difference between Louvain method and CEIL algorithm is the scoring function which is maximized. The computation of both the scoring functions using the pre-computed parameters requires O(1) time. So, the empirical running time of both the algorithms is the same. But both the algorithms generate different number and size of communities depending on the topology of the networks. This is the reason for small differences in the running time of the algorithms. Label propagation appears to have a much better running time in two of the networks. This is because we have stopped the algorithm as soon as the 95% of the nodes in the network agree with the labels of at least half of it's neighbors and have not run it to completion.

7 Conclusion

In this paper, we have characterized the necessary features required for a good community scoring function and have shown that both the internal and external properties of the communities have to be taken into account in order to design a good scoring function. To our knowledge, this is the first work to propose a scoring function having explicit components to represent the internal and external score and then combining them to get the actual community scoring function.

Earlier algorithms to find communities were resolution limit free but were non-scalable. Several methods addressed the problem of scalability but each one of them suffered from its own limitations. CEIL algorithm addresses both the problem of scalability and resolution limit. So, we believe that CEIL algorithm is a better alternative to find communities in large networks.

8 Acknowledgements

This work was funded by Ericsson Research India PO 9201813513.

References

[Blondel *et al.*, 2008] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, page P10008, 2008.

[Brandes *et al.*, 2006] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner. Maximizing modularity is hard. *ArXiv Physics e-prints*, 2006.

[Danon *et al.*, 2005] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Staistical Mechanics*, 2005.

[Fortunato and Barthélemy, 2007] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, pages 36–41, 2007.

[Myers and Well, 2003] J. L. Myers and A. D. Well. *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, 2003.

[Newman and Girvan, 2002] M. E. J Newman and Michelle Girvan. Mixing patterns and community structure in network. arXiv.org, 2002.

[Newman, 2004] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, page 066133, 2004.

[Ovelgnne and Geyer-Schulz, 2010] Michael Ovelgnne and Andreas Geyer-Schulz. Cluster cores and modularity maximization. In *ICDM Workshops*, pages 1204–1213. IEEE Computer Society, 2010.

[Ovelgonne and Geyer-Schultz, 2013] Michael Ovelgonne and Andreas Geyer-Schultz. An ensemble learning strategy for graph clustering. *American Mathematical Society*, 2013.

[Palla *et al.*, 2005] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, pages 814–818, 2005.

[Pons and Latapy, 2006] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, pages 191–218, 2006.

[Raghavan *et al.*, 2007] Usha N. Raghavan, Reka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *ArXiv Physics e-prints*, 2007.

[Rand, 1971] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, pages 846–850, 1971.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[Yang and Leskovec, 2012] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *International Conference on Data Mining*, pages 745–754. IEEE Computer Society, 2012.