Evaluating community measures on judicial citation-like networks.

June 12, 2022
INFOMNWSC

Abstract

With the emergence of large network datasets, the need to empirically measure properties of such networks is becoming more prevalent. One area that has received significant research attention is the community structure of real-world networks, with many measures proposed in the literature to measure such structure. In this study, we investigate three of these measures (edge ratio, modularity, and modularity density) and incorporate them into a modified Louvain algorithm to determine which measure most effectively captures the community structure of judicial citation networks. For this study, we employ synthetic random graphs created by a modified Lancichinetti–Fortunato–Radicchi benchmark model, designed to mimic the properties of the American appellate court citation network. We find that modularity and modularity density to be the most effective, yielding partitions with an average normalized mutual information score of 0.778 compared to the ground-truth partitions of the synthetic graphs.

1 Introduction

In the modern era, data is one of the most valuable resources available. Luckily it is also one of the most plentiful resources, but this also poses novel problems regarding the analysis and understanding of this large volume of data. Relational data can often be modelled as a Network to give more insight into the overall structure of the relations that are present, and much research has been performed in this field, which has come to be known under the name of Network Science. Multiple properties of networks have been investigated, such as the way in which random networks grow [7, 6, 1], how networks can be attacked and degrade [4, 19], and how information or viruses can spread in a network [9, 16].

Another property of networks is the emergence of communities. More precisely, communities can be described as partitions of the main graph such that the amount of edges within the nodes in each partition is higher than a certain degree compared to the edges connecting nodes inside the partition to the rest of the network. Multiple measures have been proposed to give a score to a partition of the network based on how well the partition divides the network into communities. For example, a partition of the network where the nodes inside each partition hardly have any edges connecting them might get a lower score then one where the opposite is true. Examples of such measures include Modularity [14, 11], Edge ratio [2], and statistical methods [TODO: Source for final version].

There exist many networks which have natural communities due to the nature of the network, such as families in a social network, cities in a road network, or research areas in a citation network. The existence of these multiple examples of communities, as well as the multiple measures that can be used to find suitable partitions raises the question of which measure is best suited for which type of network. In this paper, we will try to answer this question for one of these types of network: a juridical citation network; and multiple community measures. The goal is to find the type of community measure which can best be used in a simple community detection algorithm to find the ground truth communities of a network with the properties like that of the network in question, and explain why this measure is most suited for this problem.

We will start by discussing the dataset that will be considered, by giving some properties of the network. In the same section we will discuss how we will generate smaller networks networks with similar properties to use in the experiments. Section 2 will discuss the community measures that we will explore as well as the optimization algorithm that will be used to find the best partition according to these community measures. Section 4 will elaborate on our experimental setup and the method we used to compare the found partitions to the ground truth partitions. The results of our experiments will be given in Section 5, and a final discussion will be given in Section 6.

2 Dataset

2.1 Judicial citation network

In order to test the characteristics of our different measures, we must assess them as applied to a real-world network with ground truth communities.

In this study, we examine the citation network from appellate courts in the American federal judiciary.[1] The American judiciary is geographically arranged; "circuit courts of appeal" make law on a regional level, subject to appeal to the national Supreme Court. Because each circuit's precedent is binding only upon itself, courts of appeal most often cite cases from within their own region. However, citations to other circuit courts still occur for persuasion and comparison purposes. We study citations among only the circuit court layer, so as to avoid any hierarchical structure.

Thus, the groups of cases from each circuit court operate as natural ground-truth communities. Our dataset includes 12 circuit courts, and so 12 ground-truth communities. The network consists of approximately 745,000 cases (nodes) and 4.97 million citations (edges)

2.1.1 Community Structure

In order to make our study illuminative of the properties of different community measures, we must first ascertain that our network indeed demonstrates community structure.

We first observe a strong tendency of nodes to cite nodes within their own community. Though tendencies to cite to other circuits vary across the different circuit courts, on average over 70% of a circuit's citations go to other cases within its circuit.[2]

Additionally, we find that the ground-truth communities in the network have substantially higher community structure than randomly generated communities by several of our studied metrics: a modularity score of 0.63 (compared to 0.0 on average for random communities) and an edge ratio of 6.542 (compared to 0.43 for random communities).

[1] The data used in this study is sourced from the Free Law Project, an American non-profit dedicated to expanding access to legal data [17].

[2] A full table of normalized edge probabilities between circuit communities can be found in Appendix A.

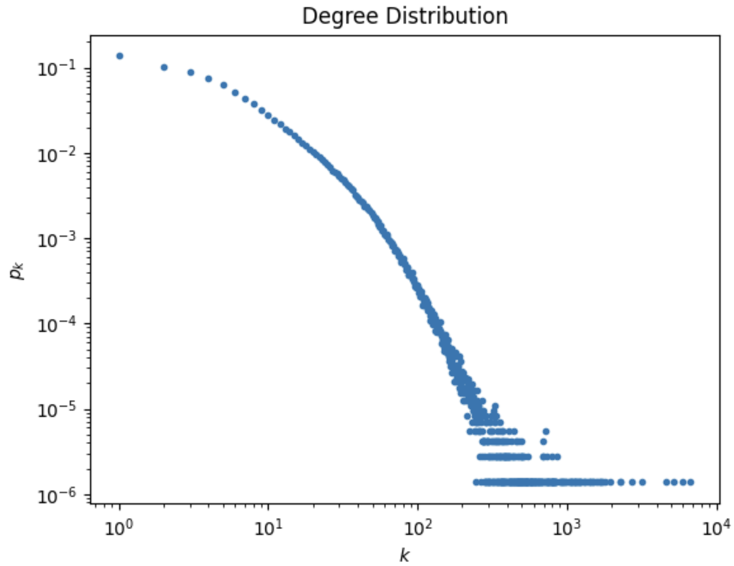

Figure 1: The degree distribution of the judicial citation network.

2.1.2 Degree Distribution

Though the prevalence of scale-free networks in the real world has been doubted in recent years (TODO cite scale free networks are rare paper), we find strong evidence that the judicial citation network we study is such a network.

Figure 1 shows the degree distribution of the network, strongly suggesting the presence of a power-law distribution. Of course, visual inspection alone of such a plot is insufficient to distinguish a true power-law distribution from other similarly-presenting distributions like log-normal and exponential distributions.

Thus, we apply Kolmogorov-Smirnov tests [12] to assess the statistical significance of the power-law fit against exponential and log-normal fits. We find that the power-law conclusion is statistically significant for $p < 0.01$ compared to these alternatives ($p = 1.30 \times 10^{-23}$ and $p = 1.28 \times 10^{-13}$, respectively).

On these bases, we conclude that the judicial citation network exhibits a power-law degree distribution.[3] The power-law fit gives a degree exponent of $\alpha = 3.57$.

2.2 Modified LFR Benchmark Model

Though the purpose of our study is to analyze the application of community detection measures to the network structure embodied by judicial citations, it is insufficient to simply study the single real network in its totality for several reasons.

First, applying our community detection algorithms and measures to the full judicial citation network is computationally infeasible in the time we have available. Second, studying only one

[3]Comparing a pure power-law fit to a truncated power-law fit, we do not find that one explains the degree distribution over the other statistically significantly ($p = 0.56$ for the pure power-law hypothesis). For the purposes of our LFR benchmark graphs discussed in 2.2, however, this distinction is not salient.

network limits the applicability of this study to only the instant network rather than a broader class of structurally-similar citation networks. Finally, we can make more robust conclusions about the properties of our different measures by assessing them against a set of networks rather than a single one.

We therefore require a way of creating smaller (i.e. fewer nodes and edges) networks that are structurally representative of our real network. Thus, we turn to the generation of random graphs that imitate both the structural and community properties of the judicial citation network. Multiple graph generation models with ground-truth communities exist, including the stochastic block model [8] and the Lancichinetti–Fortunato–Radicchi (LFR) benchmark model [10].

Our analysis in 2.1.2 indicates that our network follows a power-law degree distribution. We thus use LFR benchmark graphs in this study, because it imitates not only the community structure of our graph but also the scale-free degree distribution.

Our implementation of the LFR benchmark algorithm differs in one key respect from the original paper, however. The reference algorithm also applies a power-law distribution to the sizes of the different communities in the network. In our analysis of the real citation network, we find no statistical evidence for the hypothesis that community sizes follow a power-law distribution.[4] Thus, we instead fix the relative sizes of the communities to the proportions observed in the real citation graph.

For our experiment, we will assess our various community measures against 10 synthetic graphs created by this process using the methodology outlined in the next section.

3 Community Measures and Algorithm

We will test a total of 4 different community measures. The measures that will be tested are: Modularity [14]; Edge Ratio [2]; Modularity Density [5]; and TODO FINAL MEASURE. The four measures will be explained and discussed briefly in the next subsections. The section will end with a short discussion of the algorithm used to find a partition of the network based on a given community measure.

3.1 Modularity

Modularity is one of the most commonly used metrics for measurings community structure in a network introduced by Newman [14]. The method for weighted networks can be defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m} \right] \delta(c_i c_j)$$

Where $A_{i,j}$ is the weight of an edge connecting i and j, k_i is the degree of i, c_i is the community of i, and $\delta(u, v)$ is a function which returns 1 if both u and v share the same assigned community, and 0 otherwise.

Since our network is directed, we will use a modified version [11] defined as follows:

$$Q = \frac{1}{m} \sum_{i,j} \left[A_{i,j} - \frac{k_i^{in} k_j^{out}}{m} \right] \delta(c_i c_j)$$

[4] Indeed, given that the circuit courts are created by acts of the legislature to cover regions of roughly similar size, there is no practical mechanism by which such a distribution could be created.

Where k_i^{in} is the degree of incoming edges of i, and k_j^{out} is the degree of outgoing edges of j.

The idea of modularity is to compare the amount of edges to the expected amount of edges. I.e. if two vertices have a high degree, we expect there to be a bigger probability of an edge between them then if they both had a low degrees. In our domain, if one node has many outgoing citations, and another has many incoming, we would expect them to be connected within the same community.

3.2 Edge Ratio

Edge ratio is a measure introduced by Baumes et. al. [2] which measures the communication within a community compared to outside of the community. The edge ratio of a community C is defined as:

$$W(C) = \frac{E(C)}{E(C) + E(C_{out})}$$

where $E(C)$ is the number of edges contained in the community C and $E(C_{out})$ is the number of edges from a node inside the community to a node outside the community, or from a node outside the community to a node inside the community.

A community with a high Edge Ratio score must have many more edges inside the community relative to those going in or out. This is also what one would expect in the case of citations within juridical faculties in our network. We therefore expect this to be a good measure to find communities close to the ground truth.

3.3 Modularity Density

The third measure we will investigate is called Modularity Density. It was introduced by Chen et. al.[5] as an extension to Modularity. The measure uses the modularity measure from Newman and adds a Split Penalty (SP) to the score. The split penalty is the fraction of edges between communities defined as:

$$SP = \sum_{i,j|i \neq j} \left[\frac{|E_{c_i,c_j}|}{|E|} \right]$$

where E_{c_i,c_j} is the sum of the weights of all arcs from the community of i to the community of j.

By implementing this split penalty, it is expected that communities with little communication between them are merged less frequently and those that do are merged more frequently.

3.4 TODO: Final measure

We are still working on the final measure and will include it in the final report

3.5 The Louvain method

The Louvain method for community detection was introduced by Blondel et. al. at the University of Louvain [3]. The algorithm optimizes modularity by moving nodes between communities such that modularity is increased. This is done until no improvement can be made, at that stage the found communities are converted to nodes and the process repeats as long as the modularity increases. A simplified version of the algorithm can be written as follows:

1. Let each $n \in G$ be a community in C

2. For each n in random order, move it to the cluster of a neighbour for which the modularity increase is the largest after moving, if one exists.
3. Repeat step 2 until no improvement is found.
4. Let each community be a node in the new graph G', and add edges between nodes in the new graph if there was an edge between nodes in both communities in the original graph.
5. Return to step 1 with this new graph G'
6. Repeat steps 1-5 until no more improvement can be found.

We will use the same method with the other measures mentioned in this section, as it is possible to optimize our communities in the same way irrespective of the particular measure that is used.

4 Experimental Setup

Experiments were carried out for each measure on 10 networks generated with the same 10 random number generator seeds. Networks were generated by the process described in 2.2 to contain 5000 nodes with average degree $\langle k \rangle = 6.78$.

The following procedure was carried out for each combination of network and measure. For the given measure, the global score computed on the network using the ground truth partition was calculated to be compared later. Afterwards, the Louvain algorithm using the given measure was applied to the network with a gain threshold of 10^{-13} to prevent numerical errors (a node will only be moved to a different community if the gain from moving is larger then the threshold). The final global score on the partition found with the algorithm was calculated for the given measure.

Finally, the partition found using the Louvain algorithm with the given measure was compared to the ground truth partition of the network by calculating the Normalized Mutual Information score (NMI). The formula for NMI is defined as follows:

$$NMI(C_1, C_2) = \frac{2 * I(C_1, C_2)}{[H(C_1) + H(C_2)]}$$

Where $H(C)$ is the entropy of the clusters in C and $I(C_1, C_2)$ is the mutual information between C_1 and C_2 defined as $I(C_1, C_2) = H(C_1) - H(C_1|C_2)$.

Experiments were carried out in python using the NetworkX[13] and scikit-learn[18] libraries for the graph data structure and implementation of some algorithms.

5 Results

Measure	NMI Mean	NMI Variance
Edge Ratio	0.4391	1.447×10^{-6}
Modularity	0.7782	0.0215
Modularity density	0.7783	0.0214

Table 1: Summarized results of the experiment. For each measure, the mean and variance of the NMI scores across the synthetic graphs are shown.

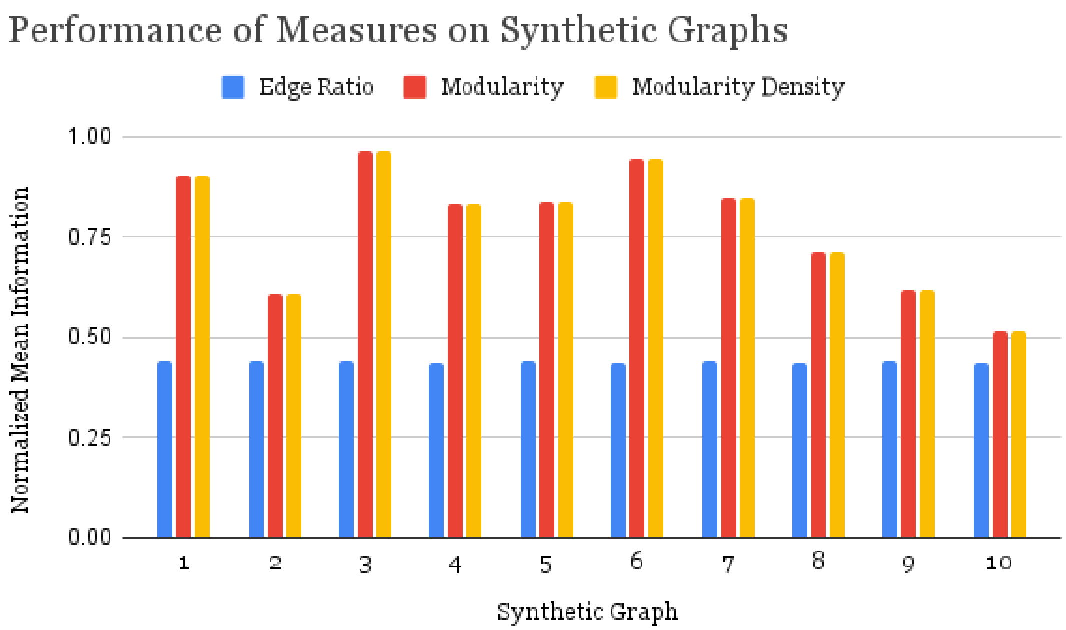

Figure 2: Results of the experiment. For each synthetic graph, the Normalized Mutual Information (NMI) score is calculated between the ground-truth communities and the communities generated by the algorithm utilizing the given measure.

Table 1 and Figure 2 give the performance of our various community measures across the synthetic graphs used for testing.

As the results show, edge ratio optimization performs the weakest of our three tested measures, owing to its rudimentary approach. On the other hand, the modularity and modularity density measures show equally strong performance, with average NMI scores of 0.7782 and 0.7783, respectively.

We also observe that modularity and modularity density exhibit a much larger degree of variance compared to edge ratio, with its NMI scores on the synthetic graphs ranging from 0.51 to 0.96. Nonetheless, both of the former measures outperformed edge ratio optimization on all tested graphs.

6 Discussion

In this paper we performed a comparative study of multiple community detection measures to determine which was most suitable to find the ground truth communities on a judicial citation-like networks using the Louvain optimization algorithm. We compared 3 measures and concluded that the modularity and modularity density measures yielded partitions which were the most similar to the ground truth partitions. As the authors of the modularity density measure write in their original paper, modularity density is free from bias and better at separating weakly connected communities compared to the original modularity measure. However, since the network we have investigated contains communities which are not weakly connected, it is not surprising that the measures performed similarly.

It is slightly surprising to see that the Edge Ratio metric only yields NMI scores which are just slightly over half as good as the NMI scores of the modularity and modularity density measures, but we speculate this might be partly caused by the fact that the Edge Ratio measure does not take into account the directionality of the network, which we believe is an integral aspect of a citation network.

6.1 Future work

This paper has provided a comparative study of multiple community measures, more specifically how well these measures can be used to find communities close to the ground truth communities in a judicial citation network. In interesting area for further research is to investigate how well these measures perform the same task on other citation networks, or real-world networks in general. To this extend it would seem beneficial to investigate how alike the network we studied is to other citation networks.

Furthermore, our research relied on a modification of the Louvain optimization algorithm to find the best communities, but there exists many other algorithms that may be modified similarly and might yield better results. Some examples of these algorithms include a greedy approach by Newman [15], or the Leiden algorithm [20].

Finally, the metrics that we have implemented and tested are far from exhaustive of the wide range of community measures that have been proposed in the field, and a more elaborate investigation of more of these measures could possibly find measure that are more fitting for this dataset than the ones that we have investigated.

References

[1] Réka Albert and Albert-lászló Barabási. "Statistical mechanics of complex networks". In: *Rev. Mod. Phys* (), p. 2002.

[2] Jeffrey Baumes et al. "Finding communities by clustering a graph into overlapping subgraphs". In: Jan. 2005, pp. 97–104.

[3] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. DOI: 10.1088/1742-5468/2008/10/p10008. URL: https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008.

[4] Duncan S Callaway et al. "Network robustness and fragility: Percolation on random graphs". In: *Physical review letters* 85.25 (2000), p. 5468.

[5] Mingming Chen, Tommy Nguyen, and Boleslaw Szymanski. "On Measuring the Quality of a Network Community Structure". In: Sept. 2013, pp. 122–127. DOI: 10.1109/SocialCom.2013.25.

[6] Paul Erdos and Alfred Renyi. "On the evolution of random graphs". In: *Publ. Math. Inst. Hungary. Acad. Sci.* 5 (1960), pp. 17–61.

[7] P. Erdös and A. Rényi. "On Random Graphs I". In: *Publicationes Mathematicae Debrecen* 6 (1959), p. 290.

[8] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps". In: *Social Networks* 5.2 (1983), pp. 109–137. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(83)90021-7. URL: https://www.sciencedirect.com/science/article/pii/0378873383900217.

[9] William Ogilvy Kermack and Anderson G McKendrick. "A contribution to the mathematical theory of epidemics". In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), pp. 700–721.

[10] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. "Benchmark graphs for testing community detection algorithms". In: *Phys. Rev. E* 78 (4 Oct. 2008), p. 046110. DOI: 10.1103/PhysRevE.78.046110. URL: https://link.aps.org/doi/10.1103/PhysRevE.78.046110.

[11] E. A. Leicht and M. E. J. Newman. "Community Structure in Directed Networks". In: *Physical Review Letters* 100.11 (Mar. 2008). DOI: 10.1103/physrevlett.100.118703. URL: https://doi.org/10.1103%2Fphysrevlett.100.118703.

[12] Frank J. Massey. "The Kolmogorov-Smirnov Test for Goodness of Fit". In: *Journal of the American Statistical Association* 46.253 (1951), pp. 68–78. ISSN: 01621459. URL: http://www.jstor.org/stable/2280095 (visited on 06/12/2022).

[13] *NetworkX documentation*. URL: https://networkx.org/.

[14] M. E. J. Newman. "Analysis of weighted networks". In: *Phys. Rev. E* 70 (5 Nov. 2004), p. 056131. DOI: 10.1103/PhysRevE.70.056131. URL: https://link.aps.org/doi/10.1103/PhysRevE.70.056131.

[15] M. E. J. Newman. "Fast algorithm for detecting community structure in networks". In: *Physical Review E* 69.6 (June 2004). DOI: 10.1103/physreve.69.066133. URL: https://doi.org/10.1103%2Fphysreve.69.066133.

[16] Romualdo Pastor-Satorras and Alessandro Vespignani. "Epidemic spreading in scale-free networks". In: *Physical review letters* 86.14 (2001), p. 3200.

[17] Free Law Project. *Bulk Data — CourtListener*. 2022. URL: https://www.courtlistener.com/api/bulk-info/ (visited on 05/06/2022).

[18] *Scicit-Learn documentation*. URL: https://scikit-learn.org/stable/.

[19] Toshi Tanizawa et al. "Optimization of network robustness to waves of targeted and random attacks". In: *Physical review E* 71.4 (2005), p. 047101.

[20] V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific Reports* 9.1 (Mar. 2019). DOI: 10.1038/s41598-019-41695-z. URL: https://doi.org/10.1038%2Fs41598-019-41695-z.

A Edge Probabilities Between Communities

Citation Influence / *Citing Circuit*	*Cited Circuit* 10th Circuit	11th Circuit	1st Circuit	2nd Circuit	3rd Circuit	4th Circuit
10th Circuit	72.72%	1.70%	3.73%	3.21%	2.50%	1.31%
11th Circuit	1.30%	77.20%	2.73%	2.53%	1.99%	0.99%
1st Circuit	1.77%	1.42%	75.80%	4.13%	2.51%	1.16%
2nd Circuit	1.47%	1.08%	3.87%	75.02%	3.01%	1.28%
3rd Circuit	2.32%	1.90%	5.34%	5.66%	65.36%	1.68%
4th Circuit	2.84%	2.54%	5.08%	5.27%	3.66%	59.87%
5th Circuit	2.60%	2.04%	4.63%	5.48%	6.09%	2.01%
6th Circuit	2.29%	1.92%	4.27%	4.41%	3.11%	1.68%
7th Circuit	2.11%	1.75%	4.01%	4.50%	3.06%	1.40%
8th Circuit	2.40%	1.58%	3.73%	3.90%	2.64%	1.51%
9th Circuit	2.56%	1.84%	4.57%	5.68%	3.54%	1.65%
DC Circuit	1.14%	0.71%	2.35%	3.29%	1.78%	1.02%

Citation Influence / *Citing Circuit*	5th Circuit	6th Circuit	7th Circuit	8th Circuit	9th Circuit	DC Circuit
10th Circuit	1.75%	1.83%	3.17%	2.41%	2.01%	3.66%
11th Circuit	4.87%	1.28%	2.20%	1.34%	1.29%	2.26%
1st Circuit	1.65%	1.58%	2.91%	1.82%	1.71%	3.54%
2nd Circuit	1.53%	1.59%	2.80%	1.67%	1.79%	4.89%
3rd Circuit	2.31%	2.21%	3.96%	2.29%	2.33%	4.63%
4th Circuit	2.56%	2.70%	4.48%	2.93%	2.44%	5.63%
5th Circuit	60.04%	2.41%	3.78%	2.75%	2.48%	5.69%
6th Circuit	2.06%	67.30%	4.14%	2.58%	2.21%	4.03%
7th Circuit	1.76%	2.14%	70.63%	2.44%	2.01%	4.19%
8th Circuit	1.88%	2.03%	3.53%	70.87%	2.01%	3.91%
9th Circuit	2.27%	2.20%	3.80%	2.51%	62.77%	6.62%
DC Circuit	1.18%	1.09%	1.95%	1.25%	1.24%	82.99%

Table 2: Relative citation frequency from each community to all other communities.