Spectral Analysis of Communication Networks Using Dirichlet Eigenvalues

Alexander Tsiatas*
Department of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0404
atsiatas@cs.ucsd.edu

Iraj Saniee
Mathematics of Networks
Alcatel-Lucent Bell Labs
600 Mountain Avenue
Murray Hill, NJ 07974
iis@research.bell-labs.com

Onuttom Narayan
Department of Physics
University of California, Santa Cruz
1156 High Street
Santa Cruz, CA 95064
narayan@physics.ucsc.edu

Matthew Andrews
Mathematics of Networks
Alcatel-Lucent Bell Labs
600 Mountain Avenue
Murray Hill, NJ 07974
andrews@research.bell-labs.com

ABSTRACT

Good clustering can provide critical insight into potential locations where congestion in a network may occur. A natural measure of congestion for a collection of nodes in a graph is its Cheeger ratio, defined as the ratio of the size of its boundary to its volume. Spectral methods provide effective means to estimate the smallest Cheeger ratio via the spectral gap of the graph Laplacian. Here, we compute the spectral gap of the truncated graph Laplacian, with the so-called Dirichlet boundary condition, for the graphs of a dozen communication networks at the IP-layer, which are subgraphs of the much larger global IP-layer network. We show that i) the Dirichlet spectral gap of these networks is substantially larger than the standard spectral gap and is therefore a better indicator of the true expansion properties of the graph, ii) unlike the standard spectral gap, the Dirichlet spectral gaps of progressively larger subgraphs converge to that of the global network, thus allowing properties of the global network to be efficiently obtained from them, and (iii) the (first two) eigenvectors of the Dirichlet graph Laplacian can be used for spectral clustering with arguably better results than standard spectral clustering. We first demonstrate these results analytically for finite regular trees. We then perform spectral clustering on the IP-layer networks using Dirichlet eigenvectors and show that it yields cuts near the network core, thus creating genuine single-component clusters. This is much better than traditional spectral clustering where several disjoint fragments near the network periphery are liable to be misleadingly classified as a single cluster. Since congestion in communication networks is known to peak at the core due to large-scale curvature and geometry, identification of core congestion and its localization are important steps in analysis and improved engineering of networks. Thus, spectral clustering with Dirichlet boundary condition is seen to be more effective at finding bona-fide bottlenecks and congestion than standard spectral clustering.

*At time of publication, at Google Inc., 1600 Amphitheatre Pkwy, Mountain View, CA 94043.

WWW'13, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory—*network problems*; C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*network topology*

Keywords

Spectral clustering, Dirichlet eigenvalues, Communication networks, Cheeger ratio

1. INTRODUCTION

An important problem in network analysis is partitioning nodes, also sometimes known as finding communities. This requires finding clusters of nodes that are inherently well-connected within themselves with sparser connections between clusters; when the clusters do not overlap, they define a partitioning of the graph. This problem is also closely related to finding network bottlenecks; if a graph has bottlenecks, then a good partition is often found by dividing the graph at its bottlenecks. Many real-world networks are truly vast, encompassing hundreds of thousands to billions of nodes and edges; for example, communication, social and biological networks. This scale produces serious computational challenges for detection of bottlenecks and communities: the large majority of algorithms are computationally too intensive to use at this scale on such graphs. Instead, one can study smaller sub-graphs of these networks; for example, the portion of a social network corresponding to one university or the portion of a communication network corresponding to one Internet service provider, and hope to derive properties of the larger graph from those of the smaller subgraphs. In this paper, we show how to define key properties of a graph, its Dirichlet spectral gap and eigenvectors which aid in clustering, and show how these closely relate to the spectral gaps and eigenvectors of its sub-graphs thus making identification of bottlenecks and points of congestion both more effective and scalable.

More precisely, spectral graph theory [3], the study of eigenvalues and eigenvectors of graph-theoretic matrices, is often used to analyze various graph properties. One might hope that the properties of a large sub-graph of a network will be representative of the properties of the entire network. Unfortunately, the properties of an expander graph depend on the conditions imposed at its (large) boundary. For example, the spectral gap of the graph Laplacian on a finite truncation of an infinite regular tree approaches zero as the size of the truncation is increased, even though the spectral gap of its infinite counterpart is non-zero. In this paper we show that, by contrast, if the spectral gap is calculated with *Dirichlet boundary conditions,* it approaches the infinite graph limit as the size of truncation is increased. Computation of a better spectral gap makes it possible to do spectral clustering more effectively thus making identification of bottlenecks and points of congestion more effective.

Motivated by this result, we compute the Dirichlet spectral gap for ten IP-layer communication networks as measured and documented by previous researchers in the Rocketfuel database [18]. We find that the Dirichlet spectral gap is much larger than the traditional spectral gap for these graphs. (Traditional spectral clustering uses the normalized Laplacian matrix $\mathcal{L}$ or some similar matrix; we use the matrix $\mathcal{L}_D$: the Laplacian restricted to the rows and columns corresponding to non-boundary nodes.) Moreover, unlike the traditional spectral gap, it does not trend downwards for larger networks. This indicates that the spectral gap for these networks viewed as sub-graphs of a much larger graph is away from zero.

There are precedents for treating networks essentially as subsets of an an overarching (infinite) graph; many network generation models [2, 6, 21] exhibit unique convergence properties (to power-law degree distributions or otherwise) as the size of the network grows to infinity. We also note that Dirichlet boundary conditions have been shown to be successful at mitigating other boundary-related issues in graph vertex ranking [5].

There is a direct connection between the spectral gap and clustering in networks, through the Cheeger inequality. Spectral graph theory has led to many effective algorithms for finding cuts that result in a small Cheeger ratio, including spectral clustering [15, 17, 16, 20] and local graph partitioning algorithms [1]. These algorithms have been well-studied, both empirically [15, 17] and theoretically [15, 20]. Unfortunately, these algorithms can also exhibit some undesirable behavior. It has been shown empirically [12] that the "best" partitionings of many networks, as measured by the Cheeger ratio, result in cutting off nodes or subtrees near the boundary of the network. The resulting 'clusters' near the boundary actually consist of several disjoint fragments. Especially when viewed as subsets of larger networks, this kind of clustering is not particularly meaningful.

In this paper, we use *Dirichlet spectral clustering* to identify good cuts in the networks in the Rocketfuel database. We use the top two eigenvectors of $\mathcal{L}_D$, the normalized graph Laplacian with Dirichlet boundary conditions, to cut the network into two sections. We demonstrate that, compared to traditional spectral clustering, there is a substantial reduction in the average number of components resulting from the cut, without a significant increase in the Cheeger ratio. Instead of finding cuts near the boundaries of the networks, Dirichlet spectral clustering obtains cuts in the network core.

The Cheeger ratio of a cut is a well known indicator of the congestion across the cut; small Cheeger ratios are likely to be associated with bottlenecks. The emphasis on identifying core bottlenecks becomes more critical in the light of the recent observation that many real-world graphs exhibit large-scale curvature [10, 14]. It has been shown [10, 14] that such global network curvature leads to core bottlenecks with load (or betweenness) asymptotically much worse than flat networks, where "load" means the the maximum total flow through a node assuming unit traffic between every node-pair along shortest paths [14]. As such, it is important to find and characterize bottlenecks at the core rather than the fringes, where they do not matter as much. Our observations, suggest that Dirichlet spectral clustering may be more useful in this regard.

The rest of this paper is structured as follows: in Section 2, we give the theoretical justification for using Dirichlet eigenvalues [4] instead of the traditional spectrum for analyzing and clustering finite portions of much larger graphs. In Section 3, we then compare the spectral gap using Dirichlet eigenvalues to the traditional spectral gap on real, publicly-determined network topologies [18] that represent smaller portions of the wider telecommunications grid. In Section 4, we demonstrate how Dirichlet spectral clustering finds graph partitions that are more indicative of bottlenecks in the network core rather than the fringes.

2. SPECTRUM OF FINITE TREES: MOTIVATION FOR DIRICHLET SPECTRAL CLUSTERING

Throughout this paper, we analyze general undirected connected graphs G by using the normalized graph Laplacian $\mathcal{L}$, defined as in [3]. For two vertices x and y, the corresponding matrix entry is:

$$\mathcal{L}_{xy} = \begin{cases} 1 & \text{if } x = y, \\ -\frac{1}{\sqrt{d_x d_y}} & \text{if } x \text{ and } y \text{ are adjacent, and} \\ 0 & \text{otherwise,} \end{cases}$$

where d_x and d_y are the degrees of x and y. We denote by λ the *spectral gap*, which is simply the smallest nonzero eigenvalue of $\mathcal{L}$.

For any graph G and finite subgraph $S \subset G$, the *Cheeger ratio* $h(S)$ is a measure of the cut induced by S:

$$h(S) = \frac{e(S, \bar{S})}{\min(\mathrm{vol}(S), \mathrm{vol}(\bar{S}))}.$$

We use $e(S, \bar{S})$ to denote the number of edges crossing from S to its complement, and the *volume* $\mathrm{vol}(S)$ is simply the sum of the degrees of all nodes in S. The *Cheeger constant* h is the minimum $h(S)$ over all subsets S. The Cheeger constant and spectral gap are related by the following *Cheeger inequality* [3]:

$$2h \geq \lambda \geq \frac{h^2}{2}.$$

Both λ and h are often used to characterize expansion or bottlenecks in graphs. This inequality shows that they are both good candidates and gives the ability to estimate one based on the other.

For the infinite d-regular tree, the spectral gap and Cheeger constant have both been analytically determined [7, 13]. Using $\mathcal{L}$, the spectral gap is

$$\lambda = 1 - \frac{2}{d}\sqrt{d-1}, \quad (1)$$

and the Cheeger constant is $h = d-2$ [9]. Both of these values are nonzero, indicating good expansion. However, the Cheeger ratio for truncated d-regular trees (TdT) – those with all branches of the infinite tree cut off beyond some radius r from the center – approaches zero as the tree gets deeper. By cutting off any one subtree S from the root, there is only one edge connecting S to $\bar{S}$, and as the tree gets deeper, this ratio gets arbitrarily small. Using the Cheeger inequality, it follows that the $\lambda_{TdT} \to 0$ as $r \to \infty$. Thus, the standard spectral properties of finite trees do not approach the infinite case as they get larger; in fact, they suggest the opposite. This is problematic when making qualitative observations about networks and their expansion, necessitating another tool for spectral analysis of networks.

The main reason why the traditional spectral gap does not capture expansion well in large, finite trees is the existence of a boundary. This is also problematic in network partitioning algorithms; often times the "best" partition is a *bag of whiskers* or combination of several smaller cuts near the boundary [12]. In this paper, we will use Dirichlet eigenvalues to eliminate this problem.

Dirichlet eigenvalues are the eigenvalues of a truncated matrix, eliminating the rows and columns that are associated with nodes on the graph boundary. We will use a truncated normalized graph Laplacian, $\mathcal{L}_D$, a submatrix of $\mathcal{L}$. This is different from simply taking the normalized Laplacian of an induced subgraph, as the edges leading to the boundary nodes are still taken into account; it is only the boundary nodes themselves that are ignored. We define the *Dirichlet spectral gap* to be the smallest eigenvalue of $\mathcal{L}_D$. This version of the graph Laplacian was first introduced in [4] to analyze local cuts on graphs.

Using Dirichlet eigenvalues, it is also possible to obtain a *local Cheeger inequality* [4] for the sub-graph S. First, the *local Cheeger ratio* is defined [4] for a set of nodes $T \subset S$ as

$$H(T) = \frac{e(T,\bar{T})}{\mathrm{vol}(T)};$$

because the boundary nodes are excluded from S in the definition of [4], the set T cannot contain any boundary nodes of S. The local Cheeger ratio $H(T)$ is the appropriate quantity when S is a sub-graph of a larger graph. Note that there is no min in the denominator; this is because the local Cheeger ratio is specific to a subgraph, and it does not make sense to take into account the rest of the graph beyond the boundary of that subgraph. The *local Cheeger constant* h_S for S is then defined as the minimum of $H(T)$ for all $T \subset S \setminus \partial(S)$. The local Cheeger inequality obtained in [4] is

$$h_S \geq \lambda_S \geq \frac{h_S^2}{2},$$

where λ_S is the Dirichlet eigenvalue of the normalized Laplacian restricted to the rows and columns corresponding to nodes in S. This inequality indicates a relationship between local expansion and bottlenecks.

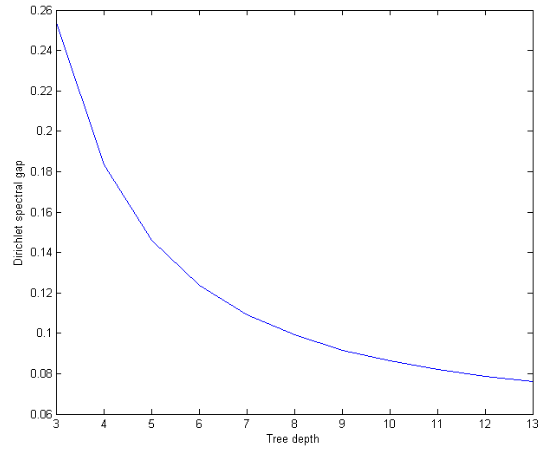


Figure 1: Dirichlet spectral gap for successively larger 3-regular trees, showing convergence to a nonzero value. The limit, as estimated by the Cheeger inequality, is $\lambda \approx 0.057$.

The use of Dirichlet eigenvalues requires that the boundary of the graph S be defined. If S is a tree, the leaf nodes are a natural choice. When S is actually a finite truncation of a larger graph, the boundary can be defined as the set of nodes that connect directly to other nodes outside the truncation; for the Rocketfuel data [18], we will use the nodes with degree 1 which presumably connect outside of the subnetwork.

We first use Dirichlet eigenvalues on d-regular trees as prototypical evidence for their effectiveness in capturing true spectral properties on real-world networks. There is empirical evidence in Figure 1, showing that the Dirichlet spectral gap for 3-regular trees indeed converges to a nonzero value as tree depth increases, contrasting with the traditional spectral gap which converges to zero. This is made rigorous in the following theorem:

THEOREM 1. *For finite d-regular trees of depth L, the Dirichlet spectral gap converges to the true spectral gap (1) of the infinite tree as L approaches infinity.*

PROOF. To derive the Dirichlet spectral gap for finite trees using the leaves as the boundary, we will solve a recurrence that arises from the tree structure and the standard eigenvalue equation

$$\mathcal{L}_d \vec{x} = \lambda \vec{x}. \quad (2)$$

Let T be a d-regular tree of depth $L+1$; the $(L+1)$st level is the boundary. We first consider eigenvectors $\vec{x}$ which have the same value at every node at the same depth within T; these eigenvectors are azimuthally symmetric. We can represent each such eigenvector $\vec{x}$ as a sequence of values $(x_0, x_1, \ldots, x_L)$, where x_i is the uniform value at all nodes at depth i, similar to the analysis of the infinite-tree spectral gap appearing in [7]. Using this eigenvector form for $\vec{x}$ in (2) leads to the recurrence:

$$x_i - \frac{1}{d}x_{i-1} - \frac{d-1}{d}x_{i+1} = \lambda x_i, 2 \leq i \leq L. \quad (3)$$

At the leaves of the tree, we have the Dirichlet boundary condition:

$$x_{L+1} = 0, \tag{4}$$

and at the root of the tree we have the boundary condition

$$x_0 - x_1 = \lambda x_0. \tag{5}$$

We can solve (3) using the characteristic equation:

$$\frac{d-1}{d} r^2 - (1-\lambda) r + \frac{1}{d} = 0,$$

whose roots can be written as

$$r_{1,2} = \frac{1}{\sqrt{d-1}} e^{\pm i\alpha} \tag{6}$$

with

$$\lambda = 1 - \frac{2}{d}\sqrt{d-1}\cos\alpha. \tag{7}$$

Since λ has to be real, either the real or imaginary part of α must be zero. Substituting the first boundary condition (4) yields a solution to (3) with the form

$$x_n = A(r_1^{n-L-1} - r_2^{n-L-1}). \tag{8}$$

for some constant A and $r_{1,2}$ given in (6). Using (5), the condition for eigenvalues is

$$\frac{\tan\alpha}{\tan(L+1)\alpha} = -\frac{d-2}{d} \qquad 0 < \alpha < \pi. \tag{9}$$

Since $\tanh x / \tanh(L+1)x > 0$ for all real x, there are no imaginary solutions to Eq.(9). Therefore all the $L+1$ solutions are real. From Eq.(7), the corresponding $L+1$ eigenvalues are all outside the infinite-tree spectral gap.

We now consider eigenvectors which are zero at all nodes up to the k'th level with $L > k \geq 0$. The eigenvector is non-zero at two daughters of some k'th level node and the descendants thereof. We assume azimuthal symmetry inside both these two sectors. The eigenvalue condition for the parent node at the k'th level forces the eigenvector to be opposite in the two sectors. Inside each sector, (3), (6), (7) (4) and (8) are still valid. However, (5) is replaced by the condition $x_k = 0$, from which $\sin(L+1-k)\alpha = 0$. There are $L-k$ real solutions to this equation, corresponding to eigenvalues that lie outside the infinite-tree spectral gap, each with degeneracy $d^k(d-1)$. The total number of eigenvalues we have found so far is

$$L + 1 + \sum_{k=0}^{L-1} d^k(d-1)(L-k) = \frac{d^{L+1}-1}{d-1} \tag{10}$$

i.e. we have found all the eigenvalues. As L gets larger, the smallest α approaches 0, showing that the Dirichlet spectral gap converges to the spectral gap of the infinite tree (1) as the depth approaches infinity. □

This derivation shows that Dirichlet eigenvalues capture the expansion properties of trees much better than the traditional spectral gap which has been shown to approach zero for large finite trees. This behavior on trees suggests that Dirichlet eigenvalues are a good candidate for use in analyzing real-world networks. Such analysis appears in Section 3.

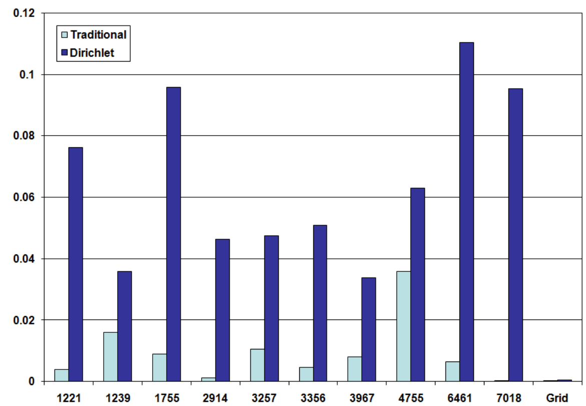


Figure 2: Comparison of traditional and Dirichlet spectral gaps in Rocketfuel data as well as the 2-dimensional Euclidean grid.

3. SPECTRUM OF ROCKETFUEL NETWORKS

Our work is motivated by derivation of scalable methods for clustering of large graphs. As an example, we study clustering of a series of datasets representing portions of network topologies using Rocketfuel [18]. Rocketfuel datasets are publicly-available, created using `traceroute` and other networking tools to determine portions of network topology corresponding to individual Internet service providers. Even though like most measured datasets, the Rocketfuel networks are not free of errors (see for example [19]), they provide valuable connectivity information at the IP-layer of service provider networks across the globe. Because the datasets were created in this manner, they represent only subsets of the much larger Internet; it becomes impossible to determine network topology at certain points. For example, corporate intranets, home networks, other ISP's, and network-address translation cannot be explored. The networks used range in size from 121 to 10,152 nodes.

Because of the method of data collection, the Rocketfuel datasets contain many degree-1 nodes that appear at the edge of the topology. In actuality, the network extends beyond this point, but the datasets are limited to one ISP at a time. As such, for these networks, it makes sense to view these degree-1 nodes as the boundary of a finite subset of a much larger network. Using this boundary definition, we compute the Dirichlet spectral gaps of these graphs and compare with their standard counterparts, as shown in Table 1 and Figure 2. It is apparent that the Dirichlet spectral gaps are much larger than the traditional spectral gaps for all the networks, implying a much higher degree of expansion than one would traditionally obtain. The spectral gaps for a two-dimensional square Euclidean grid are also shown; the grid is known to be a poor expander, and accordingly even the Dirichlet spectral gap is very small.

Figure 3 shows the same data, plotted as a function of the number of nodes N in each network. We see that the traditional spectral gap keeps decreasing as N is increased, whereas the Dirichlet spectral gap does not.

Since Figure 3 compares different networks, possibly with different properties, we confirm the result by computing the

Table 1: Structural and spectral properties of Rocketfuel datasets.

Dataset ID	Nodes	Edges	Traditional spectral gap	Dirichlet spectral gap
1221	2998	3806	0.00386	0.07616
1239	8341	14025	0.01593	0.03585
1755	605	1035	0.00896	0.09585
2914	7102	12291	0.00118	0.04621
3257	855	1173	0.01045	0.04738
3356	3447	9390	0.00449	0.05083
3967	895	2070	0.00799	0.03365
4755	121	228	0.03570	0.06300
6461	2720	3824	0.00639	0.11036
7018	10152	14319	0.00029	0.09531
Grid	10000	19800	0.00025	0.00050

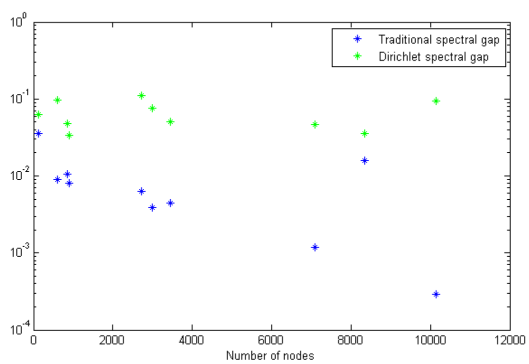


Figure 3: Comparison of traditional and Dirichlet spectral gaps across Rocketfuel networks.

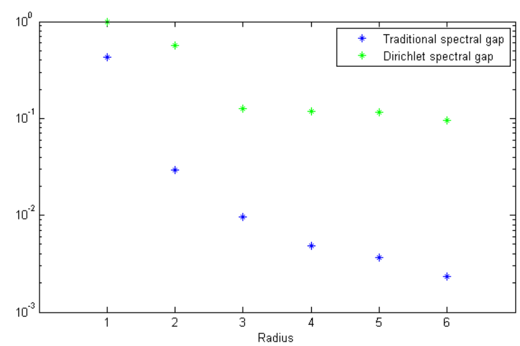


Figure 4: Comparison of traditional and Dirichlet spectral gaps in successively larger subgraphs, grown from the center of mass of dataset 7018.

spectral gap for subgraphs of different sizes drawn from a single network. All the nodes that are within a distance r of the center of mass of a network are included in a subgraph, with r varying between 1 and the maximum possible value for the network. In Fig. 4 shows the results for the largest of the Rocketfuel networks, dataset 7018 containing over 10,000 nodes. For a subgraph of radius r, the boundary is defined as all the nodes which i) have edges connecting them to nodes in the graph that are outside the subgraph or ii) connect to the outside world, i.e. that have degree 1 in the full dataset. As in Fig. 3, in Fig. 4, the traditional spectral gap keeps decreasing as r is increased, but the Dirichlet spectral gap does not.

4. SPECTRAL DECOMPOSITION

One important application of the eigendecomposition of a graph is spectral clustering or partitioning [15, 17]. The problem is to group the nodes into partitions, clusters, or communities that are inherently well-connected within themselves, with sparser connections between clusters. This is closely related to finding bottlenecks; if a graph has a bottleneck, then a good partition is often found by dividing the graph at the bottleneck. See [16] for a general survey of graph clustering.

It is often desirable for a network partition to be balanced, and finding bottlenecks near the core or center of mass of a network is often more useful than simply clipping small subsets of nodes near the boundary. But according to [12], using the Cheeger ratio as a metric on real-world data, the "best" cuts larger than a certain critical size are actually "bags of whiskers" or combinations of numerous smaller cuts. Because many graph clustering algorithms, including spectral clustering, try to optimize for this metric, the resulting partitions often slice numerous smaller cuts off the graph, which is not always useful. For our Rocketfuel data, we know that the boundary of the network is imposed by the method of data collection. Thus, by eliminating the boundary from graph clustering, we can more easily find partitions that are more evenly balanced, and bottlenecks that are closer to the core of the network.

To do this, we use standard spectral clustering techniques from [15], but instead of using the normalized graph Laplacian $\mathcal{L}$, we use the truncated Dirichlet version $\mathcal{L}_D$. The eigenvectors used for clustering will therefore not include components for the degree-1 boundary nodes, but we can assign them to the same side of the partition as their non-boundary neighbor nodes. Specifically, we compute the first two eigenvectors of $\mathcal{L}_D$ and cluster the nodes based on their components in these eigenvectors using k-means. For each node, we compute the distance to both centers and sort the nodes based on the difference. For a partition of size k, we take the top k nodes.

We follow the experiments of Leskovec et al. in [12] by using both traditional spectral clustering and Dirichlet spectral clustering to find cuts of different sizes. Specifically, we find Dirichlet cuts of all possible sizes, and then we find cuts using traditional spectral clustering for those same sizes after adding boundary nodes back in. Thus, for each network of N nodes, we calculate $N-B$ cuts, where B is the number of boundary nodes.

For each cut, we measure the Cheeger ratio h and the number of components c. Ideally, a logical cut would split the network into exactly $c = 2$ components, but as Leskovec et al. demonstrated, as cut size increases, spectral clustering and other algorithms that optimize for h yield cuts with many components. This is precisely the problem we are trying to avoid using Dirichlet clustering, and our results show that Dirichlet clustering is effective in finding cuts with fewer components. Furthermore, even though our algorithm is not specifically optimizing for h, it does not find cuts that have significantly worse values for h while finding cuts with far fewer components.

We outline some aggregate data in Table 2. For several datasets, we count the number of cuts in four different categories, comparing the Dirichlet Cheeger ratio and number of components (h_D and c_D) with traditional spectral clustering (h_T and c_T). It is evident that Dirichlet clustering finds cuts with fewer components than traditional spectral clustering ($c_D \leq c_T$) for most cut sizes, indicating that while spectral clustering optimizes for Cheeger ratio, it often "cheats" by collecting whiskers as one cut. In addition, despite the use of Cheeger ratio optimization, Dirichlet clustering sometimes finds cuts with better Cheeger ratio as well. In the last two columns for each dataset, we give the difference in h and c averaged out over all cut sizes. It turns out that the Cheeger ratios, on average, are not drastically different between the two methods, and Dirichlet clustering gives cuts with far fewer components.

Along with our aggregate data, we illustrate each individual cut for several of our Rocketfuel datasets in Fig. 5. (A few of the datasets were too large for accurate numerical computation and are therefore not shown.) For each cut size, we plot a point corresponding to the difference in Cheeger ratio h and the number of components c between Dirichlet and traditional spectral clustering. It should be clear that for the majority of cut sizes, Dirichlet clustering finds cuts with far fewer components, but there is generally little change in Cheeger ratio. This can be seen in the large variation on the c-axis with much smaller discrepancies on the h-axis. In other words, Dirichlet clustering avoids finding "bags of whiskers" while still maintaining good separation in terms of h, despite not explicitly optimizing for h.

We further visualize Dirichlet spectral clustering for two Rocketfuel data sets, shown in Figures 6 and 7. In both cases, one side of the partition is colored blue, and the other side is colored red. Notice that for these graphs, Dirichlet spectral clustering separates red and blue nodes much better than traditional spectral clustering as expected.

It is clear that using Dirichlet eigenvalues improves the partition by ignoring the boundary, alleviating the tendency to find "bags of whiskers" without drastically changing the Cheeger ratio. Although traditional spectral clustering does not always fail, there is clear evidence that Dirichlet spectral properties are an important tool in the analysis of real-world networks.

5. DISCUSSION

Our results show evidence that eigenvalues of the normalized graph Laplacian can provide rich information about real-world networks when Dirichlet boundary conditions are applied. We find that the Dirichlet spectral gap computed for several IP-layer networks is much larger than the traditional spectral gap, and is likely to go to a finite limit as the size of the network is increased. Rigorous analysis for infinite d-regular trees suggests that this may be the same as the spectral gap of a communication network that is a smaller section of something much larger. Spectral clustering using Dirichlet eigenvalues yields much better clustering than traditional methods.

The spectral decomposition using Dirichlet eigenvalues also suggests a connection to large-scale negative curvature [10, 11, 14] in the Rocketfuel data. Traditional negatively curved graphs such as trees and hyperbolic grids generally exhibit poor connectivity and core congestion. Standard clustering often yields combinations of smaller cuts near the periphery of the graph, but using Dirichlet clustering, we can see that there tend to be bad larger-scale cuts as well in the Rocketfuel datasets, in the graph interior. The presence of these larger-scale cuts is a hallmark of negative curvature or hyperbolicity [8], suggesting that Dirichlet spectral clustering may yield different behavior for hyperbolic and flat networks. The hyperbolic grids themselves are also suitable for further analysis, building from our study of regular trees. Many properties such as the spectral gap remain open questions.

With some evidence of a connection between global negative curvature, the spectral gap, and expansion, it would be interesting to empirically compare the hyperbolicity δ, the Cheeger constant h, and the traditional and Dirichlet spectral gaps of Rocketfuel and other real-world networks as well as well-known network models. From this, it could be possible to classify various networks based on these properties.

6. ACKNOWLEDGMENTS

This work was performed during an internship of A.T. at Bell Labs, Murray Hill, New Jersey and was supported by AFOSR Grants Nos. FA9550-08-1-0064 (for O.N., as a consultant with Bell Labs) and FA9550-11-1-0278.

7. REFERENCES

[1] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using PageRank vectors. In *Proc. IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 475–486, Berkeley, California, Oct. 2006.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.

[3] F. Chung. *Spectral Graph Theory.* AMS Press, Providence, RI, 1997.

[4] F. Chung. Random walks and local cuts in graphs. *Linear Algebra Appl.*, 423(1):22–32, May 2007.

[5] F. Chung, A. Tsiatas, and W. Xu. Dirichlet PageRank and trust-based ranking algorithms. In *Proc. Workshop on Algorithms and Models for the Web Graph (WAW'11)*, pages 103–114, Atlanta, Georgia, May 2011.

[6] P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci., Ser. A*, 5:17–61, 1960.

Table 2: Average Cheeger ratios (h) and number of components (c) for Dirichlet and traditional spectral clustering of several Rocketfuel datasets. Extreme values are visible in Fig. 5.

	Number of cuts in each category:							
Dataset	$c_D \leq c_T$ $h_D \leq h_T$	$c_D \leq c_T$ $h_D > h_T$	$c_D > c_T$ $h_D \leq h_T$	$c_D > c_T$ $h_D > h_T$	Avg $c_D - c_T$	Avg $h_D - h_T$	Avg c_T	Avg h_T
1221	49	197	0	6	-28.9	0.0506	36.8	0.0829
1239	538	362	0	30	-75.1	0.0127	83.4	0.1326
1755	32	91	0	14	-4.5	0.0545	7.9	0.1210
2914	224	819	0	323	-107.3	0.0565	125.8	0.1639
3257	49	67	0	35	-12.3	0.0370	20.0	0.1386
3356	182	315	3	41	-34.6	0.0388	45.6	0.1895
3967	24	137	3	129	-3.2	0.1423	9.2	0.1215
4755	15	6	0	6	-12.3	-0.0970	15.4	0.3460
6461	111	199	0	73	-13.4	0.0148	19.7	0.0999
7018	157	465	12	273	-54.3	0.0403	81.4	0.0735

[7] J. Friedman. The spectra of infinite hypertrees. *SIAM J. Comput.*, 20(5):951–961, Oct. 1991.

[8] M. Gromov. Hyperbolic groups. In S. Gersten, editor, *Essays in Group Theory*, pages 75–263. Springer, New York, 1987.

[9] O. Häggström, J. Jonasson, and R. Lyons. Explicit isoperimetric constants and phase transitions in the random-cluster model. *Ann. Probab.*, 30(1):443–473, Jan. 2002.

[10] E. Jonckheere, P. Lohsoonthorn, and F. Bonahon. Scaled Gromov hyperbolic graphs. *J. Graph Theory*, 30(2):157–180, Feb. 2008.

[11] E. Jonckheere, M. Lou, F. Bonahon, and Y. Baryshnikov. Euclidean versus hyperbolic congestion in idealized versus experimental networks. *Internet Math.*, 7(1):1–27, 2011.

[12] J. Leskovec, K. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proc. International Conference on the World Wide Web (WWW'08)*, pages 695–704, Beijing, China, Apr. 2008.

[13] B. D. McKay. The expected eigenvalue distribution of a large regular graph. *Linear Algebra Appl.*, 40:203–216, Oct. 1981.

[14] O. Narayan and I. Saniee. Large-scale curvature of networks. *Phys. Rev. E*, 84(6), Dec. 2011.

[15] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS'02)*, pages 849–856, Vancouver, Canada, Dec. 2002.

[16] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, Aug. 2007.

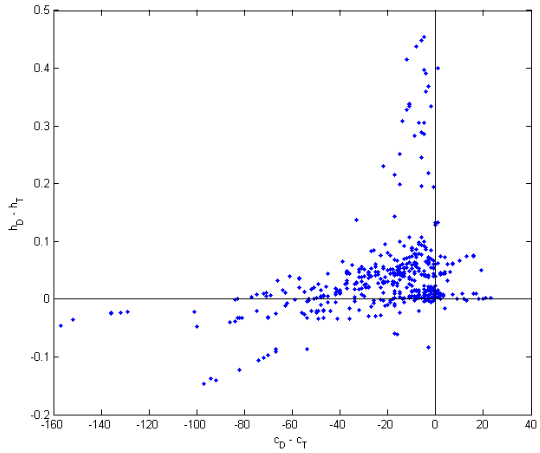
[17] J. Shi and J. Malik. Normalized cuts and image segmentation. 22(8):888–905, Aug. 2000.

[18] N. Spring, R. Mahajan, and D. Wetherall. Measuring ISP topologies with Rocketfuel. In *Proc. ACM SIGCOMM Conference (SIGCOMM'02)*, pages 133–145, Pittsburgh, Pennsylvania, Aug. 2002.

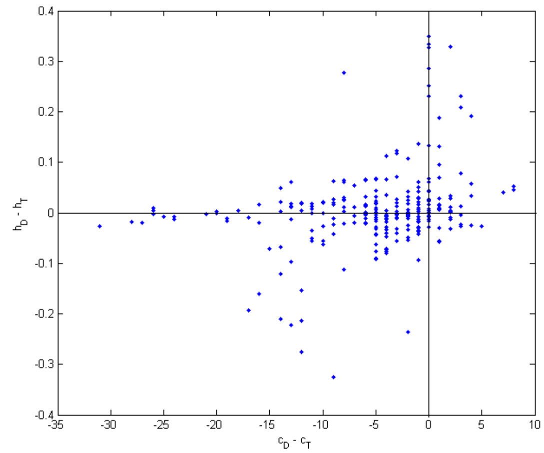
[19] R. Teixeira, K. Marzullo, S. Savage, and G. M. Voelker. In search of path diversity in ISP networks. In *Proc. Internet Measurement Conference (IMC'03)*, pages 313–318, Miami, Florida, Oct. 2003.

[20] U. von Luxburg. A tutorial on spectral clustering. *Statist. Comput.*, 17(4):395–416, Dec. 2007.

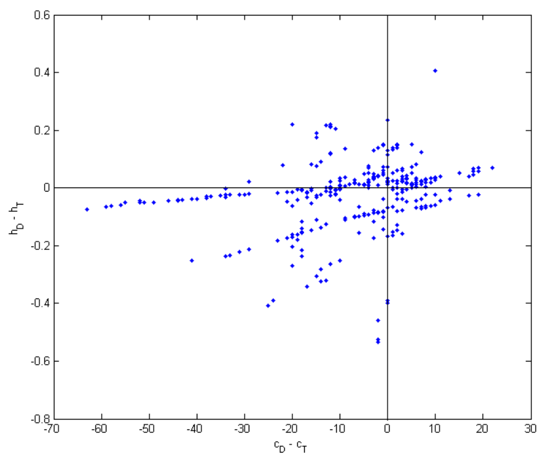
[21] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.



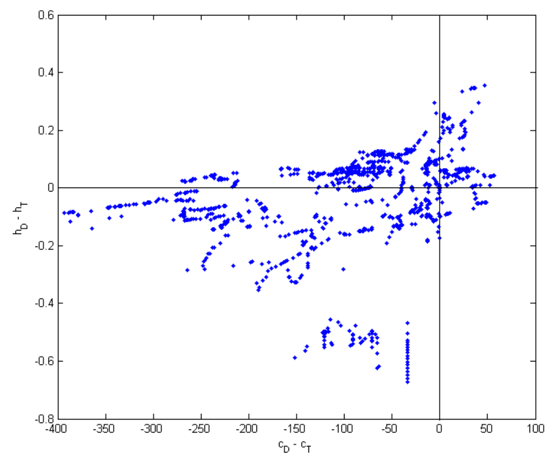
(a) Dataset 1221



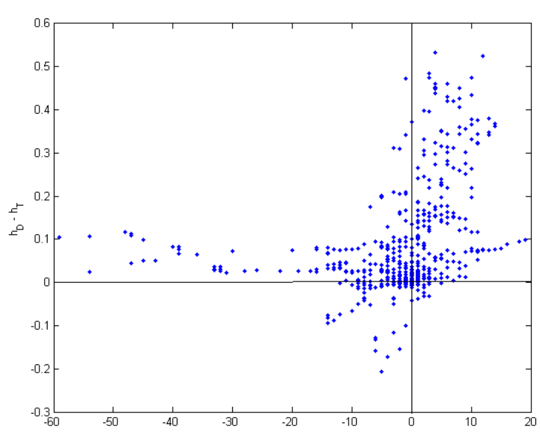
(b) Dataset 1755



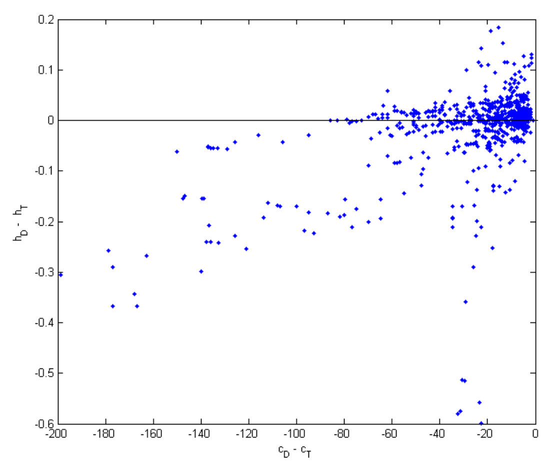
(c) Dataset 3257



(d) Dataset 3356



(e) Dataset 3967



(f) Dataset 6461

Figure 5: Comparison of Cheeger ratio h and number of components c for cuts for various datasets using Dirichlet (D) and traditional (T) spectral clustering. Each point represents one possible cut size; in general, Dirichlet clustering yields many fewer components without sacrificing much in Cheeger ratio.

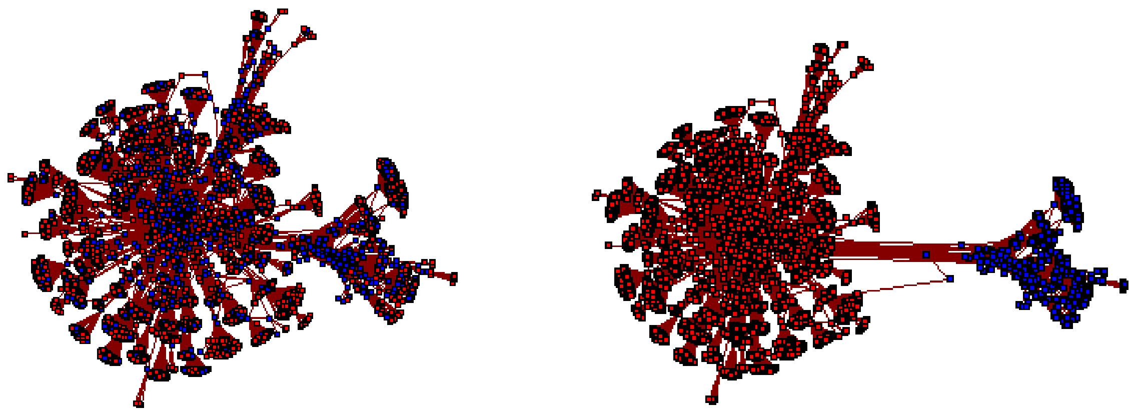


Figure 6: Partition of Rocketfuel dataset 3356 using standard (left) and Dirichlet (right) spectral clustering.

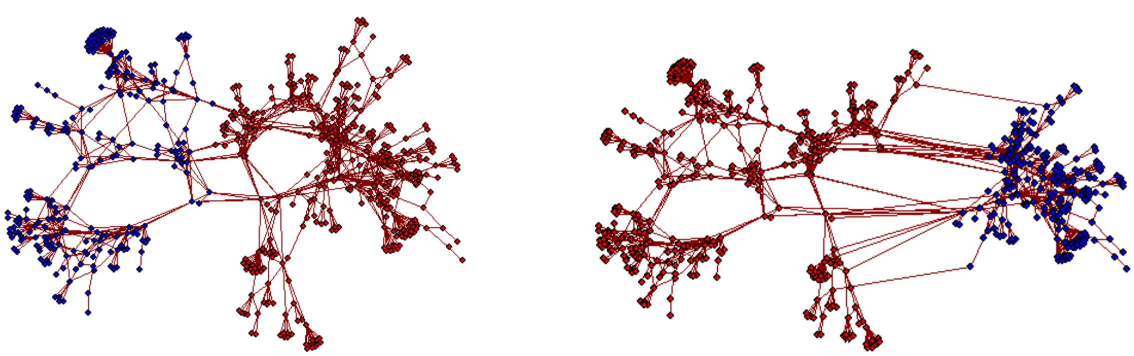


Figure 7: Partition of Rocketfuel dataset 1755 using standard (left) and Dirichlet (right) spectral clustering.