Spectral Analysis of Internet Topologies

Christos Gkantsidis, Milena Mihail, Ellen Zegura
{gantsich, mihail, ewz}@cc.gatech.edu
College of Computing
Georgia Institute of Technology
Atlanta, GA

Abstract—We perform spectral analysis of the Internet topology at the AS level, by adapting the standard spectral filtering method of examining the eigenvectors corresponding to the largest eigenvalues of matrices related to the adjacency matrix of the topology. We observe that the method suggests clusters of ASes with natural semantic proximity, such as geography or business interests. We examine how these clustering properties vary in the core and in the edge of the network, as well as across geographic areas, over time, and between real and synthetic data. We observe that these clustering properties may be suggestive of traffic patterns and thus have direct impact on the link stress of the network. Finally, we use the weight of the eigenvector corresponding to the first eigenvalue to obtain an alternative hierarchical ranking of ASes and links between the ASes.

I. INTRODUCTION

Studying and modeling network topologies is necessary for protocol performance evaluation and simulation of a variety of network problems. Early modeling efforts focused around random graphs with relatively regular degree distributions [35], [40], [41], [6]. With the rapid growth of the network and the persistent effort of network measurement [12], [11], [32], real topology data started becoming available, in particular at the AS (Autonomous System) level, where it was first observed that the degree distribution of the AS level topology is actually consistently highly skewed [10]. Consequently, the research community has shown considerable interest in obtaining topology models that better resemble the real data [22], [5], [30], [1], as well as understanding the impact of such network topologies on the performance of network protocols [9], [38].

This new generation of synthetic Internet topologies models is strongly driven by the observed skewed statistics of the degree sequence and its evolution, and by even further observations of more detailed graph theoretic characteristics of the network. Most notably, following the natural intuition that, for example, geography must be relevant in the real Internet topology, [30] paid special attention to the “clustering” coefficient; the observation of the significance of geography has been also made in [29] and [2].

A major question remaining open is why does the network topology exhibit the observed behavior? What are the micro-primitives that drive the network’s growth? Notice that syntactic answers such as “growth with preferential connectivity” which have, on the one hand, the double advantage of (a) matching the macro-statistics of the network and (b) being efficient to simulate, have, on the other hand, the disadvantage of offering no suggestion on further semantics, such as link bandwidth, congestion, or traffic pattern. These latter semantics are very important in a network study. A first effort towards developing direct primitives can be found in the work of [16] and [25] who suggest that heavy tailed statistics can be the result of multi-objective optimization. In particular, [25] give a network growth model where heavy tailed statistics follow from an optimization tradeoff between cost and performance.

In this paper we revisit the issue of clustering. As opposed to previous work that has focused on the clustering coefficient, our starting point is the method of spectral filtering. This method examines the large eigenvalues of matrices related to the adjacency matrix, and looks for clusters in the eigenvectors associated with these eigenvalues. Indeed, the first reference to the large eigenvalues of the adjacency matrix of the AS Internet topology is the “eigenvalue power-law” which was reported together with the “degree power-law” in [10]. The connection between spectral filtering and graph connectivity, including clustering, has been extensively studied in discrete mathematics (e.g. see the books of [7], [28] and the further references that they point to), and has found very successful applications in information retrieval and data-mining where clusters represent groups of data with semantic proximity [20], [25], [27], [3], [24]. Practical experience suggests that spectral analysis might be better suited for data that lack regularity (thus it has been extensively used in computer science), while clustering coefficients are better suited for data that have stronger regularities (thus it has been extensively used by physicists who study lattices, crystals, etc.). Indeed, by definition, spectral filtering yields a large number of clusters, and it can be applied iteratively in subgraphs of a network. By contrast, it is not clear how to grow clusters around nodes with large clustering coefficient and this approach is not typical in information retrieval or data-mining.

Our contributions include:

- An explanation of the eigenvalue power-law of [10]: it is a consequence of the degree power-law. This implies that the eigenvectors associated with the largest eigenvalues of the adjacency matrix of the AS topology cannot express information beyond the degrees, and thus further normalizations are needed to retrieve non-trivial clustering properties.
- Adaptation of the spectral filtering method in the

1 Though a related approach called “k-means” is quite common: but we do not expand further on it, since we do not use it in this work.
context of the AS Internet topology, by (a) performing inverse frequency normalization via stochastic matrices, (b) considering similarity transformations and (c) considering the entire topology as well as subgraphs of the topology. As a result, we get non-trivial groupings of ASes with clear semantic proximities, such as geography and business interests. We note that without this adaptation, i.e., by considering the eigenvectors corresponding to the eigenvalues of the adjacency matrix as in [10], we get trivial groupings corresponding to the large ISPs and their customers: this is indeed a restatement of the highest degrees.

- The observation that the clustering properties (a) vary in the core and the edge of the network and across geographic areas, (b) are persistent and consistent over time, and (c) are not matched by synthetic Internet topology generators.
- Study of the connection between the information retrieved by spectral filtering and link stress (link stress can be thought of as a first approach towards congestion). In particular, we argue that the eigenvectors associated with the largest eigenvalues are suggestive of non-trivial intra-cluster traffic patterns that cause significant decrease in the link stress. The decrease is much more notable in the Internet than in any synthetic topology. (If on the other hand the traffic patterns become intracluster the link stress correspondingly increases). This reasoning is in line with [26], [16] which suggest that network characteristics should be studied in the context of the design problem they are trying to solve, and that such approaches have better hope of giving further network semantics.

- A detailed and efficient AS ranking method according to the 1st eigenvector of a suitably defined stochastic matrix, which has strong correlation with other known hierarchical assignments [17]. This approach is an adaptation of the pagerank used by Google [39]. An adaptation of the same method for ranking links between ASes, The found rankings are highly correlated with link stress under uniform traffic. A further adaptation of the method to obtain groups of ASes that correspond to seemingly highly stressed cuts.

The balance of the paper is as follows: In Section II we cover necessary primitives from linear algebra and highlight the intuition behind the spectral filtering method. We also introduce normalizations and similarity transformations, and discuss their suitability and necessity for graphs with skewed statistics, like the Internet topology. In Section III we describe the spectral filtering results for the AS Internet topology, and give the qualitative nature of the information retrieved by the eigenvectors. In Section IV we give an application of the information retrieved by the eigenvectors in terms of defining non-trivial traffic patterns that deviate from uniform traffic. In Section III we give a method of ranking ASes and links between ASes that is highly correlated with hierarchical assignments.

II. SPECTRAL ANALYSIS OF MATRICES ARISING FROM GRAPHS

In this Section we give a high level overview of the intuition and the primitives of spectral filtering. We discuss the basics of eigenvalues and eigenvectors of matrices, some useful transformations and normalizations, and why the eigenvectors corresponding to the large eigenvalues contain information relevant to clustering. This motivates the processing that we will do to the eigenvectors of the AS Internet topology in Section III. We also give an explanation of the eigenvalue power-law of [10] as a restatement of the Zipf with exponent 1 rank-degree distribution; this serves as additional motivation for the processing in Section III, in the sense that without this processing the spectral method does not give non-trivial information.

A. Eigenvalues and Eigenvectors of a Matrix A

Let $G(V, E), |V|=n$, be an undirected graph and let $A$ be its adjacency matrix: $a_{ij}=1$ if $(i, j) \in E$, $a_{ij}=0$ otherwise. Since $G$ is undirected, $A$ is symmetric $a_{ij}=a_{ji}$. In general, the $(i,j)$-th entry of a symmetric matrix can be thought of as a measure of the correlation between parameters $i$ and $j$. Let $\vec{c}$ be an $n$-dim real vector; $\vec{c}$ can be thought of as a function on the vertices of $G$. We say that $\vec{c}$ is an eigenvector of $A$ with eigenvalue $\lambda$ if and only if $\vec{c}A = \lambda \vec{c}$. It is a well known fact of linear algebra that every $n \times n$ real symmetric matrix $A$ has a spectrum of $n$ orthonormal eigenvectors $\vec{e_1}, \vec{e_2}, ..., \vec{e_n}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ [14], [36]. In general, each eigenvalue $\lambda_i$ of multiplicity $k$ is associated with a unique invariant subspace, any $k$ orthonormal vectors of which can be eigenvectors. Thus, the eigenvectors are unique up to degeneracies related to equal eigenvalues.

How do the above algebraic notions of eigenvectors and eigenvalues relate to the structural properties of the graph? Below we list a few facts which build the intuition behind the spectral filtering method (the statements are straightforward, though some of the proofs to which we point are quite technical).

(a) The largest eigenvalue $\lambda_1$ of a $d$-regular graph is $d$ and the corresponding eigenvector assigns uniform weights to all vertices [7], [21].

(b) All eigenvalues $\lambda_i, 2 \leq i \leq n$, of a uniformly random $d$-regular graph (i.e. every $d$-regular graph on $n$ vertices is equally likely) are $|\lambda_i| \leq O(\sqrt{d})$, almost surely [7].

(c) The eigenvalues $\lambda_i, 1 \leq i \leq n$, of a graph with $m$ edges and maximum degree $d$ are bounded by $|\lambda_i| \leq \min\{\sqrt{m}, d\}$ [21].

(d) The spectrum of the union of vertex disjoint graphs is the union of their spectra [7], [21].

(e) If $A$ and $B$ are the adjacency matrices of not necessarily disjoint graphs with eigenvalues $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_n$ and $\beta_1 \geq \beta_2 \geq ... \geq \beta_n$, then the eigenvalues of their union $C = A + B$ are $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_n$ with $\alpha_i - \beta_n \leq \gamma_i \leq \alpha_i + \beta_1, 1 \leq i \leq n$ [14], [36]. In addition, the corresponding invariant subspaces of $C$ follow from the invariant subspaces of $A$ and $B$ perturbed by no more than the maximum invariant subspace of $B$ [14], [15].
The intuition behind the spectral filtering method is that, if we take the union of two vertex disjoint regular random graphs \( A_1 \) and \( A_2 \) and connect them with a few random edges \( B \), then, combining Facts (a) through (e) above, the spectrum of \( C = A_1 + A_2 + B \) will have \( \gamma_1 \approx \gamma_2 \approx d \) (corresponding to the largest eigenvalues of \( A_1 \) and \( A_2 \)) and \( \gamma_i = O(\sqrt{d}) \), \( 2 < i \leq n \). Furthermore, we expect to identify the vertices of \( A_1 \) and \( A_2 \) by examining the eigenvectors corresponding to the first two eigenvalues. See Figure 1. Indeed, the second eigenvector assigns mostly large negative weights on \( A_1 \) and mostly large positive weights on \( A_2 \).

In general, the eigenvectors corresponding to large eigenvalues tend to capture global characteristics of the graph and its semantics, such as groups of nodes \( S \subset V \) for which the ratio

\[
\frac{\text{edges inside } S}{\text{edges incident to } S} = \frac{|\{(i, j) \in E : i \in S, j \in S\}|}{|\{(i, j) \in E : i \in S, j \in V\}|} \quad (1)
\]

is small, indicating clusters of relatively high connectivity and, thus, presumably further semantic proximity, not necessarily otherwise expressed in the data (the deep theory of “expander” graphs supporting this claim can be found, for example in [7], [28]). In addition, finding a set \( S \) that minimizes the above ratio is an \( NP \)-hard problem, and thus the spectral method is also an efficient approximation algorithm [33]. Eigenvectors corresponding to small eigenvalues tend to capture noise, or local characteristics that are explicit or can be easily computed from the data.

In broad lines, the spectral filtering method for an \( n \times n \) symmetric matrix \( A \) proceeds as follows:

**STEP 1:** Compute the \( k \) largest eigenvalues of \( A \) together with the corresponding eigenvectors. The parameter \( k \) depends on the application and the instance, but it is always one to two orders of magnitude smaller than \( n \).

**STEP 2:** For each \( i \), \( 1 \leq i \leq k \), let \( \vec{e}_i \) be the eigenvector associated with \( \lambda_i \). Sort the vertices according to the weight assigned by \( \vec{e}_i \). A typical profile of the sorted vertices is in Figure 2. Cut towards the most positive end and towards the most negative end, with special preference to sharp jumps, if they exist (a good example of a sharp jump can be found in Table II). These groups are candidates for clustering and/or semantic proximity.

**B. Similarity Transformation** \( \text{SIM}(A) = A \cdot A^T \)

Now suppose that \( G(V, E), |V| = n \), is a directed graph, and thus the adjacency matrix \( A \) is no longer symmetric. \( A \) is no longer guaranteed to have a complete real spectrum, and the notion of clustering is not well defined either. Let \( A^T \) be the transpose of \( A \), i.e., \( a^T_{ij} = a_{ji} \). Notice that the product \( A \cdot A^T \) is a symmetric matrix. Notice further that its \( (i, j) \)-th entry is \( \sum_{k=1}^{n} a_{ik} a_{kj} \), measuring the number of nodes that \( i \) and \( j \) point to in common. In the case where the nodes represent ASes and edges are directed from customers to their providers, the above sum relates \( i \) and \( j \) to the number of their common providers. Similarly, the product \( A^T \cdot A \) relates \( i \) and \( j \) to the number of their common customers. The transformation \( A \cdot A^T \) is very common in spectral analysis. Depending on the application, it is called self-adjoint, co-citation, co-variance, or similarity transformation. Here we shall use the notation \( \text{SIM}(A) = A \cdot A^T \).

**C. Stochastic Normalization**

The intuition behind the spectral filtering method that we gave in the previous paragraphs referred to regular graphs. Indeed, in practice, the spectral filtering method has been found to deteriorate rapidly when the frequencies of non-zero entries vary substantially [24], which is certainly the case with the very skewed degrees of Internet topologies. Inverse frequency normalization is a general approach to restore spectral filtering in such cases.

In its simplest form, inverse frequency normalization divides each entry \( a_{ij} \) with the sum \( \sum_j a_{ij} \) of the entries of the corresponding row, thus obtaining a matrix where all the rows add up to 1. Notice that this is now a stochastic matrix, in the sense that it describes the transition prob-
abilities of a Markov chain in the natural way. Like symmetric matrices, stochastic matrices also have a complete spectrum of real eigenvalues and eigenvectors (i.e., they span the n-dimensional space), moreover, the range of the eigenvalues is conveniently normalized in \((-1, 1)\) \([14]\). If, in addition, we make all diagonal entries \(a_{ii} = 1/2\) and multiply all other entries by \(1/2\), the range of the eigenvalues shifts to \((0, 1)^2\). For any matrix \(A\), we denote its stochastic normalization \(N(A)\). In what follows, we may apply the stochastic normalization to either \(A\) or \(\text{SIM}(A)\), thus getting \(N(A)\) or \(N(\text{SIM}(A))\).

**D. Faloutsos’ Eigenvalue Power-Law**

[10] examined the spectrum of the adjacency matrix of the AS Internet topology, without performing any normalization or other transformation. They reported a power-law on the twenty or so largest eigenvalues of this matrix with exponent between .45 and .5.

We observe that Faloutsos’ eigenvalue power-law is a direct consequence of the degree sequence power-law along the lines of Facts (d) and (e) of Section II.A, in the following sense:

**STEP 1:** Decompose an undirected AS topology \(A\) as \(A = F + E\), as follows. Initially \(F\) is the set of vertices that have the \(k\) highest degrees, and let \(d_1, d_2, \ldots, d_k\) be these degrees. Initially \(F\) contains no edges. Let \(E\) be the entire AS topology graph. Now we will remove some edges of \(E\) and add them to \(F\), so as to create \(k\) disjoint stars in \(F\). We do this by the following process: For each vertex \(v\) that is not in \(F\), if \(v\) is incident to \(k\) vertices in \(F\), pick one of these vertices \(u\) with probability proportional to the degree of \(u\) in the entire graph, make the edge \(\{v, u\}\) incident to the vertex \(u \in F\) and remove the edge \(\{v, u\}\) from \(E\). Notice that \(F\) is now a set of vertex disjoint stars with degrees \(d_1', d_2', \ldots, d_k'\), and \(E\) is the initial AS topology where all edges belonging to the stars have been removed.

**STEP 2:** Notice that the eigenvalues of a star of degree \(d\) are \(\pm \sqrt{d-1}\) and 0 with multiplicity \(d-1\) \([21]\). Thus, by Fact (d) of Section II.A, the largest eigenvalues of \(F\) are \(\sqrt{d_1'}, \sqrt{d_2'}, \ldots, \sqrt{d_k'}\). Also, by Fact (e) of Section II.A, the largest eigenvalues of \(A = F + E\) cannot be perturbed by more than the largest eigenvalues of \(E\).

**STEP 3:** For typical AS topologies, we have found that the above procedure, for \(k = 100\), gives \(d_i' = d_i, 1 \leq i \leq k\), hence the largest eigenvalues of \(F\) are close to \(\sqrt{d_1'}, \sqrt{d_2'}, \ldots, \sqrt{d_k'}\), and the largest eigenvalues of \(E\) are in the worst case strictly smaller than \(\sqrt{d_1'}\) and on the average 1/5 of \(\sqrt{d_1'}\). Now by Fact (e) of Section II.A, the largest eigenvalues of \(A = F + E\) can be understood to be close to \(\sqrt{d_1'}, \sqrt{d_2'}, \ldots, \sqrt{d_k'}\). Hence, for graphs where the largest degrees follow Zipf with exponent close to 1, as [10] reported for the AS Internet, the largest eigenvalues follow a power-law with exponent close to .5, also as [10] reported for AS the Internet. (We also refer to [34], [19], [23] for formal analysis of stochastic models of power-law random graphs).

We may now conclude that by looking at the eigenvectors corresponding to the largest eigenvalues examined in [10] we should not hope to get information beyond the ASes of highest degree and their customers. Indeed, in experiment, we have found these eigenvectors to be highly concentrated on the large ISPs. Therefore, to obtain more interesting clusters, we will need the processing discussed in III.

**III. SPECTRAL ANALYSIS OF AS INTERNET TOPOLOGY**

In this Section we describe the spectral analysis that we performed on AS Internet topologies. We discuss the used data, the processing, the behavior of large eigenvalues, and the resulting groups of ASes from the corresponding eigenvectors. We show that connectivity and clustering varies in the core and the edge of the network, as well as across different geographic areas. On the other hand, the clustering is consistent over time.

**A. Data Used, Transformations and Normalizations**

We have used topology data from two sources. The first source is the data of [18] which collect BGP routing information from many routers in the Internet and combine all the routing tables to reconstruct the undirected AS topology. Using the heuristics in [17], we also have the information of whether an edge of the undirected topology corresponds to a customer-provider or a peering relationship. Finally, [17] give a method to assign the ASes to the levels of a 5-level hierarchy. The most important ASes, like big ISPs in the core of the Internet, are assigned to level 1. The smallest ASes are assigned to level 5. The topological data from this effort dating on April 6, 2002 are the ones used most in our study. We should note that perhaps the most complete set of data is in [37]. We did not use this data because it was not easy to annotate it with the AS relationship and hierarchy information of [18]. We do not believe that this affects the results of our study, in the sense that the collection process is such that missing links would quite likely strengthen the clustering findings.

The second set of data is from [11]. Though this data is far less complete, it has the advantage that it spans the time period of 1997 to date. We have thus used this data to study the evolution of clustering over time. [11] does not contain information about the relationships between the ASes. We have used the algorithm of [13] to infer AS relationships\(^3\).

An AS topology without AS relationships corresponds to an undirected graph with a symmetric adjacency matrix \(A\) in the natural way. For such a topology we perform spectral analysis on the stochastic normalization \(N(A)\). An AS topology with customer-provider or peer relationships corresponds to a directed graph \(A'\), where \(a_{ij}' = 1\) and \(a_{ji}' = 0\) \(^2\) on the other hand, the eigenvectors of stochastic matrices are not necessarily orthogonal, and sometimes additional normalizations that rectify orthogonality and necessary for good results. In our analysis this did not turn out to be necessary. We also note that there are many further normalization methods, including so-called Laplacians and divisions by logarithmic or other functions of \(\sum_j a_{ij}\), but, again, we did not use them in our analysis.

\(^2\) On the other hand, the eigenvectors of stochastic matrices are not necessarily orthogonal, and sometimes additional normalizations that rectify orthogonality and necessary for good results. In our analysis this did not turn out to be necessary. We also note that there are many further normalization methods, including so-called Laplacians and divisions by logarithmic or other functions of \(\sum_j a_{ij}\), but, again, we did not use them in our analysis.

\(^3\) In addition to customer-provider and peering, [13] includes sibling relationships: to be consistent with our first set of data, we replace sibling relationships with peering relationships.
if and only if \( i \) is a customer of \( j \) and \( j \) is a provider of \( i \), and \( a_{ij}^* = a_{ji}^* = 1 \) if and only if \( i \) and \( j \) are peers (in all other cases the entries are 0). For such a topology we perform spectral analysis on the stochastic normalization \( N(SIM(A')) \).

If we perform spectral analysis starting from the entire undirected graph \( A \) or directed graph \( A' \) we find that the clusters indicated by the eigenvectors associated with the large eigenvalues correspond to groups of nodes assigned levels 3, 4 and 5 of the hierarchy of [17], thus are away from the core of the network. This is intuitive, since we expect the edge of the network to have more areas with higher connectivity inside the area and relatively lower connectivity to the rest of the network, along (1) of Section II. Similarly, we expect that the core of the network is better connected, and thus the ratio (1) of Section II is consistently higher in the core.

To capture the clustering properties of the core of the network we have to explicitly isolate the core from the edge and analyze the core alone. A first method to do this we use the information about the hierarchy of [18]. In this case, we assume that the core of the AS topology is composed of all the ASes that are assigned level 1 in the topology of equal to or less than 4. When this information is not available, we can iteratively prune all the nodes in the graph that have degree one or two. As in the previous paragraph, the topology of the core of the network can be either directed or undirected. We call these graphs Core\((A)\) and Core\((A')\). As above, we perform spectral analysis to \( N(\text{Core}(A)) \) and \( N(\text{SIM}(\text{Core}(A'))) \).

### B. Results for the Entire AS Topology

Figure 3 shows the largest eigenvalues of the AS topology of [18]. We have considered the adjacency matrices of the topology without and with AS relationships, both for the entire network and for the core. The point to notice in this graph is that the eigenvalues are quite high, indicating the existence of clusters in the underlying topology. The point to also notice is the drop in the eigenvalues between the entire topology and the core of the network. This is to expected because the core was constructed by removing small ISP's which tend to cluster more.

Next we give some representative groups of nodes corresponding to the highest weights assigned by eigenvectors corresponding to large eigenvalues. The first example was taken using the \( N(\text{SIM}(\text{Core}(A))) \). The group corresponds to the largest eigenvalue, which is 1.0. In Table I, we list the members of the group that take the highest weights in the eigenvector.

In Table II, we give a group of ASes that belong to Chinese providers. This was taken from the eigenvector of \( N(\text{SIM}(\text{Core}(A))) \) that corresponds to the 6th largest eigenvalue with value 0.8363. Notice that the clusters of relatively big ASes in Tables I and II (levels 1 through 4 of the hierarchy) appear in prominent positions when we examine the core of the topology. As we shall see below, such clusters do not appear when we examine the entire topology.

#### Table I

A sample of a group of ASes found in the \( N(\text{Core}(A)) \) topology. This cluster is taken using the eigenvector which corresponds to the highest eigenvalue. The ASes in this group are big ISP providers, mostly in North America and Europe. The weights of the eigenvector did not show a sharp jump. In such a case we would use a heuristic of taking the first 50 to 100 entries (depending on the application).

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
AS & Weight & Level & Description & Country \\
\hline
3303 & 0.1096 & 2 & Global Crossing & CH \\
209 & 0.1032 & 3 & Elonet & US \\
5511 & 0.0986 & 3 & France Telecom, Worldwide & FR \\
3549 & 0.0986 & 1 & Global Crossing & US \\
3582 & 0.0983 & 3 & University of Oregon & US \\
4513 & 0.0972 & 3 & Global Crossing & US \\
6461 & 0.0967 & 3 & Primary AS for Abroad Network & US \\
1808 & 0.0917 & 2 & AOL Transit Data Network & US \\
1299 & 0.0916 & 1 & TeliaNet Global Network & SE \\
3356 & 0.0907 & 1 & Level 3 Communications & US \\
70 & 0.0897 & 1 & Alcatel & FR \\
3001 & 0.0896 & 1 & Cable & Wireless (CW) & US \\
6305 & 0.0889 & 2 & Broadband & US \\
8018 & 0.0884 & 2 & Carrier1 Autonomous System & US \\
4565 & 0.0883 & 2 & Epipe & US \\
1239 & 0.0867 & 4 & SprintLink Backbone & US \\
6079 & 0.0866 & 3 & RCN Backbone AS & US \\
6259 & 0.0862 & 3 & Fiber Network Solutions, Inc. & US \\
2407 & 0.0852 & 2 & INET & JP \\
2914 & 0.0846 & 1 & Vero & US \\
2628 & 0.0846 & 2 & NO Communications, Inc. & US \\
254 & 0.0842 & 2 & DIGEX, Inc. & US \\
5459 & 0.0840 & 3 & London Internet Exchange Ltd. & GB \\
5650 & 0.0840 & 2 & Electric Lightwave, Inc. & US \\
\hline
\end{tabular}
\caption{Table I}
\end{table}
TABLE II

Another group of AEs found in the N(Core(A)) topology. This group was found in the eigenvector corresponding to the 6th largest eigenvalue. The last entry (AS 4058) does not belong to the group. We have included it to indicate a typical sharp jump suggestive of where to cut a group.

In Table III we give a group of AEs that belong to Greek academic institutions. This was taken from the eigenvector of N(SIM(A)) that corresponds to the 2nd eigenvalue with value 0.9539. Notice that this cluster of rather small AEs (levels 4 and 5 of the hierarchy) appears in prominent position when we examine the entire topology.

We should note that the three examples presented here are typical. We chose to include the particular examples wanting to give one cluster from each constient. For example, in Table III we gave the second eigenvector because the first corresponded to Core and we had already given China in Table II.

A natural question arises: do these clusters correspond to bad cuts, in the sense of (1) of Section II? We have measured the corresponding ratio (and its adaptations for directed graphs) and have found that, indeed, as we approach and overcome the natural jump of the eigenvector profile, the ratio (1) tends to increase and decrease sharply. However, we have not found consistent criteria (as is common with information retrieval heuristics, performance evaluation is subjective or indirect), and since we intend to give qualitative rather than quantitative results, we do not expand on this issue. In Section IV, where we conduct an application specific experiment, we make our definitions more precise.

C. Results specific to Geography

Is the Internet topology homogeneous across the entire globe? Do the same connectivity patterns apply everywhere? The first synthetic models of Internet topologies which emphasized the principle of preferential connectivity [5], [22] were implicitly making such homogeneity assumptions. Recently, these assumptions have been challenged, most notably in [2], [29] who show strong correlation between the placement of AEs and routers with geography as well as economic development. We second and strengthen these finding, by observing that different geographic parts of the network exhibit different connectivity patterns.

We have used the data of [31] to assign AEs to continents. We constructed three graphs for the continents of North America (NA), Europe (EU) and Asia (AS)\footnote{It is possible that some AEs are present in more than one continent. We treated such AEs as belonging to only one continent. However, their number is very small, and the results are not affected.}. We included AS relationships, thus obtaining non-symmetric adjacency matrices ANA for North America, AEU for Europe and AAS for Asia. In Figure 4 we give the largest eigenvalues of N(SIM(ANA)), N(SIM(AEU)) and N(SIM(AAS)).

We also give the plots for the spectrum of the corresponding cores N(SIM(Core(ANA))), N(SIM(Core(AEU))) and N(SIM(Core(AAS))), the point to notice is that, both in the entire topology and in the core, North America exhibits less clustering than Europe and Asia. This can be understood intuitively by thinking of the network in North America as being at a later evolutionary stage, and hence is more connected.

D. Spectrum Consistency over Time

Is the spectral behavior of the Internet topology consistent over time? See Figure 5. We have used snapshots from [11] taken six months apart and found consistent behavior.
of the largest eigenvalues of $N(A)$. This confirms the intuitive belief that the spectrum is a robust characteristic of a topology. Figure 5 refers to the entire AS topology without AS relationships. We have observed similar behavior in the evolution of the AS topology with AS relationships, as well as the core of the topology, and when restricted to specific continents.

IV. Impact of Spectral Analysis on Performance and Traffic Primitives

Walt is the significance of the information retrieved by the spectral analysis of Section III? What is the significance of the eigenvectors associated with the large eigenvalues? The main difficulty in answering this question is in deciding which metric to pick and examine its correlation with clustering. In general, there is no consensus on the metrics by which Internet topologies should be evaluated. One approach is to include detailed graph properties [5], [22], [30], while another approach is to use metrics that distinguish graphs with heavy tailed degree sequences as opposed to more regular topologies and may be correlated with further coarse characteristics of the network [9], [38].

Our approach is closer to the latter, and influenced from the proposal of [26], [16] that topology properties should be studied in connection to the functionality of the network. In particular, we shall study the correlation of the information retrieved from the eigenvectors of Section III to the performance of a primitive experiment that studies the “congestion” in the network.

For an undirected (without AS relationships) topology, suppose that we send one unit of traffic along a minimum hop (shortest) path from each node to every other node. This induces a stress for each link defined as the total number of paths going through the link. We study the mean link stress and the maximum link stress. The maximum link stress can be thought of as an indicator of congestion. It is reasonable to expect that, in general, topologies with higher principal eigenvalues, and thus worse cuts (in the sense of (1) of Section III), should tend to exhibit worse link stress behavior. Therefore, we may tend to conclude that, because of its strong clustering, the AS Internet topology will be exhibiting worse link stress behavior than if it were more homogeneous and less clustered.

We computed the average and max link stress of the AS Internet and several synthetic topologies from Brite (BA,GLP,Waxman [22], [4], [30], [35]) Inet [5], and PLRG [1]. Indeed, all the synthetic topologies exhibited smaller average link stress than the average link stress of the AS Internet, and all synthetic topologies, except Inet [5], exhibited smaller max link stress than the max link stress of the AS Internet.

Why then is the Internet strongly clustered? And what are the clusters associated with the large eigenvectors suggestive of? Here we propose a plausible explanation related directly to the parameter affecting link stress, namely, the amount of traffic between any two nodes. In particular, we
observe that, as this traffic shifts from uniform to intracluster, the maximum link stress decreases strikingly more in the AS Internet than in any synthetic topology. We define intracluster traffic as follows: If \( n \) is the size of the topology, consider the \( .05n \) largest eigenvalues of \( N(A) \), and the eigenvectors associated with each such eigenvalue. Consider the nodes \( H_1 \) and \( H_2 \) that are assigned the highest \( .25n \) positive and the highest \( .25n \) negative weights in each such eigenvector. Each AS which appears in \( H_1 \) or \( H_2 \) for at least one examined eigenvector will be assigned to the cluster of the positive or negative end of the first eigenvector in whose \( H_1 \) or \( H_2 \) it appeared. In this way we assign each AS to at most one cluster. We say that a traffic pattern is \( \alpha \% \) intraclustered if each node sends \( \alpha \% \) of its traffic exclusively inside the cluster that it belongs, and \( 1-\alpha \% \) of its traffic uniformly to all nodes (thus uniform traffic is 0\% intraclustered). We also adjust the definition of link stress, by weighting shortest hop paths that use the link with the percentage that the particular path takes under the particular traffic pattern. See Figure IV. We have indicated the drop in the max link stress as the traffic pattern shifts from uniform to 100\% intraclustered. The interesting observation is that in the case of the AS Internet the max link stress drops by 50\%, while in every synthetic topology the drop is between 10\% and 25\%.

We therefore propose that the information retrieved from the eigenvectors associated with the largest eigenvalues may be suggestive of intracluster traffic patterns. We propose to use the clusters suggested by these eigenvectors as one meaningful way to generate traffic patterns that deviate from uniform traffic. One additional remark is due. It may be thought that the decrease in link stress under intracluster traffic patterns is a straightforward consequence of shorter min-hop paths that would be used in an intraclustered traffic pattern, See Table V. For each node, define its expected hop distance as the expected hop distance of the node from every other node under a specific traffic pattern. Notice that both in the Internet and in the synthetic topology produced by Inet, the drop in the average expected hop distance is not nearly as striking as that of the max link stress. We therefore conclude that the drop in the link stress is a result of a better distribution of the weighed shortest paths rather than a mere decrease of their length. Thus the intracluster traffic pattern is indeed non trivial.

Of course, similarly we can define intercluster traffic patterns. These patterns create corresponding sharp increases in the link stress, and can be used for pessimistic (worst case) analysis.

Finally, for completeness, in Figure 6 we give the largest eigenvalues of the AS Internet topology, as well as similar graphs generated by Inet [5], Waxman, growth with preferential connectivity according to Barabasi-Albert and the improved GLP heuristics [22], [30] which explicitly tries to capture better clustering (all the above for the same number of nodes as the Internet topology), and the power law random graph (PLRG) model of [1] (for the specific degree sequence of the Internet topology). We give the spectrum of both the entire AS topology and the core (recall that the core of synthetic topologies where there is no other indication of hierarchy is obtained by iterative pruning). Overall, the Internet tends to have higher eigenvalues. However, BA and Waxman have higher eigenvalues than the Internet in the core. We believe that this is due to high regularity of these models, and have found that the corresponding eigenvectors suggest clusters whose size is much smaller than those of the Internet. Thus in this case the seeming shift is a result of the normalization in the denomenator of (1) in Section II. Also, PLRG has has higher eigenvalues in the entire topology (though it does have lower eigenvalues in the core). We examined the corresponding eigenvectors of PLRG, but did not find a conclusive explanation for this behavior. We do not include the figures, for lack of space.

### Table V

<table>
<thead>
<tr>
<th></th>
<th>Internet Max Link Stress</th>
<th>Internet Avg. Exp. Hop Dist</th>
<th>Inet Max Link Stress</th>
<th>Inet Avg. Exp. Hop Dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>213,380</td>
<td>3.3744</td>
<td>496,261</td>
<td>2.4690</td>
</tr>
<tr>
<td>20%</td>
<td>195,260</td>
<td>3.2856</td>
<td>477,400</td>
<td>2.7161</td>
</tr>
<tr>
<td>40%</td>
<td>177,010</td>
<td>3.1965</td>
<td>466,720</td>
<td>2.6892</td>
</tr>
<tr>
<td>60%</td>
<td>158,820</td>
<td>3.1076</td>
<td>435,950</td>
<td>2.6564</td>
</tr>
<tr>
<td>80%</td>
<td>130,530</td>
<td>3.0187</td>
<td>425,180</td>
<td>2.6106</td>
</tr>
<tr>
<td>100%</td>
<td>122,440</td>
<td>2.9297</td>
<td>414,410</td>
<td>2.5757</td>
</tr>
</tbody>
</table>

Indicating the drop in max link stress and average expected hop distance, as the traffic shifts from uniform to intraclustered. The trend is the same for all other synthetic topologies that we considered.

### V. Ranking by the First Eigenvector

The “significance” of an AS, or its position in a hierarchy, is a subjective matter, in the sense that ASes are never explicitly or implicitly assigned such rankings. There is relatively good agreement about the “top” and “bottom” of a hierarchy (for example, an ISP that has only peers and no provider is almost surely very big and hence at the top of the hierarchy), while an AS that has no costumers or peers and only one or two providers is almost surely very small and hence at the bottom of the hierarchy). In two separate efforts, [13] and [17] gave heuristics to assign hierarchical levels to ASes, after inferring AS relationships and taking into account several non-trivial further characteristics.

In this Section we observe that a different heuristic, based on the weights assigned to the ASes by the first eigenvector of a suitably defined modification of the directed AS graph (i.e., after AS relationships have been inferred), is highly correlated with the hierarchy of [17]. The proposed heuristic is an adaptation of the pagerank method used by Google to infer quality of Web pages. The analogy is natural. Both the directed AS topology and the WWW are directed graphs. In the WWW, a hyperlink pointing from a page \( i \) to a page \( j \) indicates an endorsement of importance from \( i \) to \( j \). In the Internet, an edge pointing from a customer \( i \) to a provider \( j \) can be thought of as a similar endorsement of importance, while in peers the endorsement becomes mutual.
indicating the drop of the max link stress, as the traffic shifts from uniform to intracorporate. The AS Internet exhibits twice as much drop as any synthetic topology. We note that these numbers refer to the core of the network. The behavior was similar when we did the same experiment in the whole network, and in each specific continent.

The ranking method is the following. Let \( A' \) be the directed adjacency matrix. For each node \( i \) define the out-degree of \( i \) as \( d_{out}(i) = |\{ j : a_{ij} = 1 \}|. \) Now consider the stochastic matrix

\[
P(A) : p_{ij} = \begin{cases} 
\frac{\alpha}{d_{out}(i)} & \text{if } a_{ij} = 1 \\
\frac{1 - \alpha}{n - d_{out}(i)} & \text{if } a_{ij} = 0 
\end{cases}
\]

The above stochastic matrix represents a random walk on the directed graph \( A' \), where with probability \( \alpha \) we go to a provider or peer chosen uniformly at random, and with probability \( 1 - \alpha \) we jump to a uniformly random node from the set of all nodes (the latter step is a standard correction to avoid degeneracies pertaining to sinks).

Let \( \pi(v) \) be the stationary probability of the stochastic matrix \( P(A) \). Google assigns to Web pages pagerank quality \( \pi(v) \). By analogy, we assign to each AS hierarchical weight \( \pi(v) \). In Figure 7 we compare the hierarchy of [17] to our hierarchical weight \( \pi(v) \). We have used \( \alpha = .95 \); the results are similar for any \( .9 \leq \alpha \leq .99 \). To plot the graph, we have grouped the ASes by their level in the hierarchy. Then, we sort the ASes in each group by their weight in \( \pi(v) \) and plot the weights in decreasing order. Observe that we use logarithmic scale for both axes.

There is important correlation between the weights assigned to the ASes and their level in the hierarchy. Nodes assigned by [17] in high levels have higher values in \( \pi(v) \). Also, the weights assigned to the ASes of a group are in general higher than the weights assigned to ASes that belong in groups of lower level. One noticeable exception is the weights assigned to levels 4 and 5. ASes in these levels have very small degree and they cannot be easily separated by the page rank method. At first glance it seems that there is an “anomaly” in the figure, since there are some ASes that are assigned higher weights than ASes which belong to higher levels. We argue that this could be a problem of the subjective nature of hierarchical assignment, and/or the heuristic used by [17] to assign ASes to levels. We will discuss two examples that make this point. The largest weights in levels 2 and 3 have a very high value which is comparable to the weights assigned to nodes in level 1. These weights correspond to the ASes of Tiscali Intl Network (AS number 3257) and of AboveNet (AS number 6461) respectively. We believe that they had to be assigned in the highest level. This is justified by their degrees in the
adjacency matrix, which are 330 and 585 respectively, and by the reputation they have as big ISP providers.

We extend the above method to obtain an assignment of significance to links. If \( n \) is the number of ASes and \( m \) is the number of links of the undirected AS topology, let \( N=n^2 \) be the number of pairs of ASes and associate with each such pair a shortest path between their endpoints. We may now consider the \( m \times N \) traffic matrix \( T \), where each row corresponds to a shortest path and there is a 1 on the columns of the links used by the path. Using the SVD method, which is a generalization of the decomposition into eigenvalues and eigenvectors for non-square matrices, we can compute the left eigenvector of \( T \) that corresponds to the largest eigenvalue. Just like pagerank, this eigenvector gives an order of importance to links. Links that get higher values are associated with links that accept more traffic and thus are candidates to be places of congestion. Observe that this statement was made without making any assumption about the traffic between any two ASes.

To find the correlation between the importance assigned to links and the amount of traffic they receive we did the following experiment\(^5\). We assumed that between each pair of ASes there is some amount of traffic flowing drawn from a uniform distribution that takes values between 0 and 2 traffic units. After performing shortest path routing and assigning loads to links, we have ordered the links by their load. We are interested to find the relation between this ordering and the ordering given by the weights in the eigenvector. In Figure 8 we depict this relationship. There is a point in \((i, j)\) when a link is in \(i\)-th position sorted by the load and in \(j\) position sorted by the weight in the eigenvector. Indeed it is easy to observe that there is strong correlation between the importance of the link and the amount of traffic it receives. The correlation coefficient in this case is 0.8594 indicating this strong correlation.

\(^5\)For this example we have used an induced graph of the real topology which includes all the ASes in levels 1 and 2 as assigned by [17]. Memory and processing limitations did not allow us to work with bigger matrices.

In addition, it is possible to use the left eigenvectors to identify clusters of related links that form a cut in the original adjacency matrix. The links in the cut carry traffic between areas in the Internet that are not well connected and thus they are candidates to be points of congestion. As a simple example we give Figure 9, where we draw a cluster of links (cluster in the same sense as the clusters defined earlier for ASes) taken from the left eigenvector which corresponds to the second largest eigenvalue. Intuitively, we expect that indeed the trans-Atlantic links to carry a lot of traffic and thus be points of congestion as indicated. We have observed similar clusters using the other eigenvectors, and also in positions that seem intuitively natural (across Central and Eastern Europe, across the Pacific, etc). It is still an open question to us how the clusters observed in the AS topology related with the link clusters.

VI. RELATED WORK AND SUMMARY

Spectral filtering is a well known and very established information retrieval method. We studied the adaptation of spectral filtering in the AS Internet topology. We found that the information retrieved corresponds to groups of nodes with obvious semantic proximities. We found that the clustering behavior varies in the core and in the edge of the network, and across different geographic areas. We suggested one useful application of these clusters, as sug-
gesting non trivial inatcluster and intercluster traffic patterns, and that such traffic patterns affect the stress on elements of the network. Finally, we adapted a different spectral method, namely that of Google's pagerank, to obtain an alternative and more detailed hierarchical characterization of nodes, as well as an assignment of “significance” to links. Our study proves that spectral filtering methods can be successful in processing Internet topologies.

Beyond information retrieval, spectral methods have found great applicability in information compression, via the technique of low rank approximations [3, 25, 27]. Examining if such low rank approximations apply to in the networking context (e.g. speed-up simulations) is a very important question. We believe that our study is a first step in this direction.

Finally we should mention that the first reference to to the high end of the spectrum of Internet topologies is due to [10], and another interesting study of the spectrum of the Internet can be found in [8] who discuss properties of the entire spectrum and relate them to certain structural properties of the Internet graph.

REFERENCES