ANALYSIS OF INTERNET TOPOLOGY DATA

Johnson Chen and *Ljiljana Trajković*
Simon Fraser University
Vancouver, Canada
{hchenj, ljilja}@cs.sfu.ca

ABSTRACT

Discovering Internet topology is important for analyzing routing protocols and Internet robustness and resilience. Recent research results dealing with Internet topology, such as the discovery of power-laws and the application of normalized Laplacian analysis to Internet topology data, have increased the need for more complete datasets and their more rigorous interpretations. In this paper, we examine datasets from two sources: Route Views and RIPE. We show that each dataset may have a geographical bias and, consequently, consist of distinct routes. Furthermore, by employing normalized Laplacian analysis, we identify distinct cluster characteristics that could not be inferred by directly examining the collected data.

1. INTRODUCTION

In spite of the exponential growth of Internet, certain characteristics of Internet topology remain invariant. Better understanding of these invariants may contribute to further Internet research and development, such as new protocol designs.

Because of the large size of Internet, researchers frequently restrict analysis of Internet topology to the autonomous system (AS) level. An AS administers one or more networks that have a coherent routing policy. Instead of dealing with millions of hosts and routers, researchers can analyze Internet topology that consists of less than 65,000 ASs [1]. Collecting accurate AS data becomes of great importance when Internet information is reduced to tens of thousands ASs.

Unfortunately, due to the underlying complex mechanisms and the extremely large size of Internet, it may be impossible to acquire complete Internet AS data. Therefore, studies of Internet topology rely either on limited Internet AS data or on employing synthetic topology generators. Collected AS data can be categorized in two groups: data emanating from the Border Gateway Protocol (BGP) routing tables and IP addresses collected using a TCP utility called *tracerout*.

In this paper, we examine datasets from two sources: the Route Views project from the University of Oregon [2] and Réseaux IP Européens (RIPE) [3]. Both datasets are collected from BGP routing tables and have been extensively used by the research community. In this paper, we propose the notion of reverse pairs and use spectral metrics to analyze the Internet data. The two datasets that we used are large compared to those previously reported in research studies [4] - [7].

We provide a short survey of Internet topology research in Section 2 and a description of AS relationships in Section 3. We give a short introduction to spectral analysis in Section 4. Our results are presented in Section 5, while we conclude with Section 6.

2. INTERNET TOPOLOGY AND DATASETS

Until 1999, a usual approach to analyzing Internet topology was to use randomly generated graphs, where routers were represented by vertices and transmission lines by edges. A major breakthrough in exploring properties of the Internet topology on AS level was achieved by Faloutsos et al., [4]. Despite the apparent randomness of the Internet, by analyzing three snapshots of Internet topology they discovered several simple power-laws that govern Internet topology: node degree vs. node rank, degree frequency vs. degree, number of nodes within a number of hops vs. number of hops, and the eigenvalues of the adjacency matrix vs. their order.

Chang et al., [5] questioned the completeness of data used in [4] because the analysis relied exclusively on the BGP data from Route Views [2]. Hence, they suggested inclusion of additional data. After examining the "extended" dataset by including data from RIPE [3], the authors [5] arrived at a conclusion that departed from the original power-law discovery in Internet topology [4].

Eigenvalues associated with a network graph are closely related to important topological features, such as diameter of the network, presence of cohesive clusters, long paths and bottlenecks, and how random the network graph is. Network researchers have often explored these spectral properties. Vukadinović et al., [6] recently reported that the normalized Laplacian spectrum (*nls*) of Internet topology on AS level (AS graph) is invariant regardless of the exponential growth of the Internet. *nls* can also distinguish between AS graphs and synthetically generated graphs. In another approach, Mihail et al., [7]

used the eigenvectors corresponding to the largest eigenvalues of the Laplacian matrix to find clusters of ASs with certain characteristics, such as geographic locations or business interests.

Despite of various findings related to the BGP AS-level Internet topology, researchers usually rely on the two major available datasets: Route Views [2] and RIPE [3], as described in Table 1.

Table 1. Comparison of data sources.

	Route Views	RIPE
Faloutsos et al., [4]	yes	no
Chang et al., [5]	yes	yes
Vukadinović et al., [6]	yes	no
Mihail et al., [7]	yes	yes

3. AS RELATIONSHIPS

The BGP protocol allows each AS to choose its own administrative policy for selecting routes and propagating reachability information to other routers. However, these routing policies are constrained by the commercial agreements between domains. Hence, AS relationships [8] are an important factor and need to be considered when the analysis of Internet topology is based on data obtained from BGP routing tables.

AS relationships can be classified into three groups: customer-provider, peers, and siblings. An AS sets its export policies according to its relationships with the neighboring ASs. AS relationships can be translated into the following rules that govern BGP export policies [9], [10]: an AS can export to a customer or to a sibling its routes and the routes of its customers, providers, or peers. However, when an AS exports to a provider or to a peer, only its own routes and its customer routes can be exported. Hence, in Internet routing, AS relationships are a major factor that contributes to the difference between the path from a source to a destination and the return path to the source.

4. SPECTRUM OF A GRAPH

A graph G (V, E) is a set of vertices V connected by a set of edges E. An Internet AS graph represents a set of ASs connected via logical links. The number of edges incident to a node in an undirected graph is called the degree of the node. In digraphs (directed graph with a set of nodes connected by a set of directed links), *indegree* and *outdegree* of a node indicate how many links are directed to and out of the node, respectively. Two nodes are called *adjacent* if they are connected by a link.

A network can be represented by the adjacency matrix A(G):

$$A(i,j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases}.$$

Associated with A(G) is a diagonal matrix D(G) with row-sums of A(G) as the diagonal elements. D(G) indicates the connectivity degree of each node. The Laplacian matrix is defined as L(G) = D(G) – A(G).

Eigenvalues of a matrix M are defined as numbers λ satisfying $Mx = \lambda x$ for a non-zero vector x. Vector x is called an eigenvector of the matrix M belonging to eigenvalue λ. The collection of all eigenvalues is called a *spectrum*. The eigenvalues of L(G) are closely related to certain graph invariants. For example, the spectrum of L(G) contains 0 for every connected graph component.

The normalized Laplacian matrix NL(G) is defined as [11]:

$$NL(i,j) = \begin{cases} 1 & \text{if } i = j \text{ and } d_i \neq 0 \\ -\dfrac{1}{\sqrt{d_i d_j}} & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}.$$

where d_i and d_j are the degrees of nodes i and j, respectively.

5. SPECTRAL ANALYSIS OF INTERNET TOPOLOGY

We analyzed data from the Route Views and RIPE datasets, collected on a typical day in May 2003. Each dataset consists of 65,000 ASs. In each dataset, we only considered the first 30,000 assigned ASs by the Internet Assigned Numbers Authority (IANA) [1]. Most of the remaining 35,000 ASs have node degree zero. This reduced the computational complexity without compromising the accuracy of our results.

5.1. Description of datasets and analysis of raw data

A brief summary of the analyzed data is given in Table 2. In a directed AS graph, a pair of ASs can be represented either as one or two pairs of ASs, depending on whether the link is unidirectional or bi-directional. 15,369 probed ASs in Route Views are also found in RIPE. 29,477 connected AS pairs in the directed graph of Route Views also exist in the RIPE dataset. The two datasets possess very similar characteristics. As shown in Table 2, the two datasets contain routing information from overlapping sets of ASs, with only 0.3% difference. For the directed AS graph, 85% of AS links from Route Views are identical to 84% of AS links in RIPE.

Table 2. Number of ASs and AS pairs in the two datasets.

Dataset	No. of ASs	AS pairs (undirected graph)	AS pairs (directed graph)
Route Views	15,418	34,057	34,878
RIPE	15,433	34,274	35,225

Table 3 shows the assigned numbers (ASN) of ASs found in the Route Views and RIPE datasets. Fourteen of twenty ASs in both datasets are identical. 70% of the core ASs (ASs with the largest degrees) are identical between the two datasets.

Table 3. Twenty ASs with largest node degrees.

Rank of degree	Route Views	RIPE	Rank of degree	Route Views	RIPE
1	701	701	11	6461	4589
2	1239	1239	12	4513	6461
3	7018	7018	13	4323	8220
4	3561	209	14	16631	3303
5	1	3561	15	6347	13237
6	209	3356	16	8220	6730
7	3356	3549	17	3257	4323
8	3549	702	18	4766	3257
9	702	2914	19	3786	16631
10	2914	1	20	7132	6347

According to AS relationships [8], a route from a source to a destination and a route back to the source may differ. We call *reverse pair* a pair of ASs (A, B) present in both datasets with a link (A→B) existing only in one and the reverse link (B→A) exists only in the other dataset. Since most participating sites in the Route Views project are located in North America, the collected AS routes tend to have a North American AS origin. Hence, these routes have a North American biased AS relationship. Similarly, the AS routes found in the RIPE dataset, have a European bias. A consequence of this geographical bias is that certain routes in the first dataset may never be found in the second. These routes are called *unknown routes*. If the ASs forming these unknown routes in the first dataset are also present in the second dataset, reverse pairs can be located in the intersection of the two datasets. Locating the remaining unknown routes would require internal information from ASs. Note that the number of ASs forming unknown routes is at least equal to the number of ASs in reverse pairs. Hence, reverse pairs may be used to indicate the geographical bias of Internet topology datasets.

We found 558 reverse pairs in the two datasets that we analyzed. They represent approximately 1.60% and 1.58% of all AS pairs in Route Views and RIPE, respectively. With approximately 15% distinct AS pairs in the two datasets, the number of reverse pairs is not negligible. Our analysis suggests that data from the Route Views and RIPE datasets may indicate geographical bias and, consequently, consist of distinct routes. This characteristic of the datasets should be considered when analyzing Internet topology.

5.2. Spectral of AS Internet topology

Eigenvectors of a Laplacian matrix corresponding to small eigenvalues are often used to partition data [12]. The second smallest eigenvalue of a Laplacian matrix is called "algebraic connectivity" [13] and it is closely related to the connectivity characteristic of the normalized Laplacian matrix [11]. After normalizing the Laplacian matrix, elements of the eigenvector corresponding to the largest eigenvalues tend to be positioned close to each other if they correspond to AS nodes with similar connectivity patterns constituting clusters [11], [14].

We calculated the second smallest and the largest eigenvalues in the two datasets that we analyzed. Each element (weight) of an eigenvector corresponds to one of 30,000 ASs. We sorted the elements of each eigenvector in ascending order. Hence, every element of the two eigenvectors is indexed. The connectivity status is equal to 1 if the AS is connected to another AS or zero if the AS is isolated or is not present in the routing table.

Figure 1 shows that eigenvectors corresponding to the second smallest eigenvalues tend to capture characteristics that can be directly obtained from the raw data. Among the 30,000 ASs, ~15,000 ASs with a degree larger than zero (connected ASs) are grouped together as shown in Figures 1(a) and (c). In Figure 1(b), two large clusters are visible in the Route Views dataset. In Figure 1(d), values of 1 and 0 are tightly interwoven because the RIPE dataset has a large number of relatively smaller clusters compared to Route Views. However, information directly available from the analyzed datasets (number of probed ASs, number of connected AS pairs, set of core ASs, and degree distribution of ASs) is insufficient to identify the highly connected clusters revealed by the analysis of the eigenvector corresponding to the second largest eigenvalue.

6. CONCLUSIONS

In this paper, we analyzed two Internet topology datasets. We propose the notion of reverse pairs and use it as a new metrics to analyze the datasets. Our findings suggest that each of the two datasets collected from two distinct sources may have geographical bias. We also employed spectral analysis to explore the differences between the two datasets and to find their distinct clustering features.

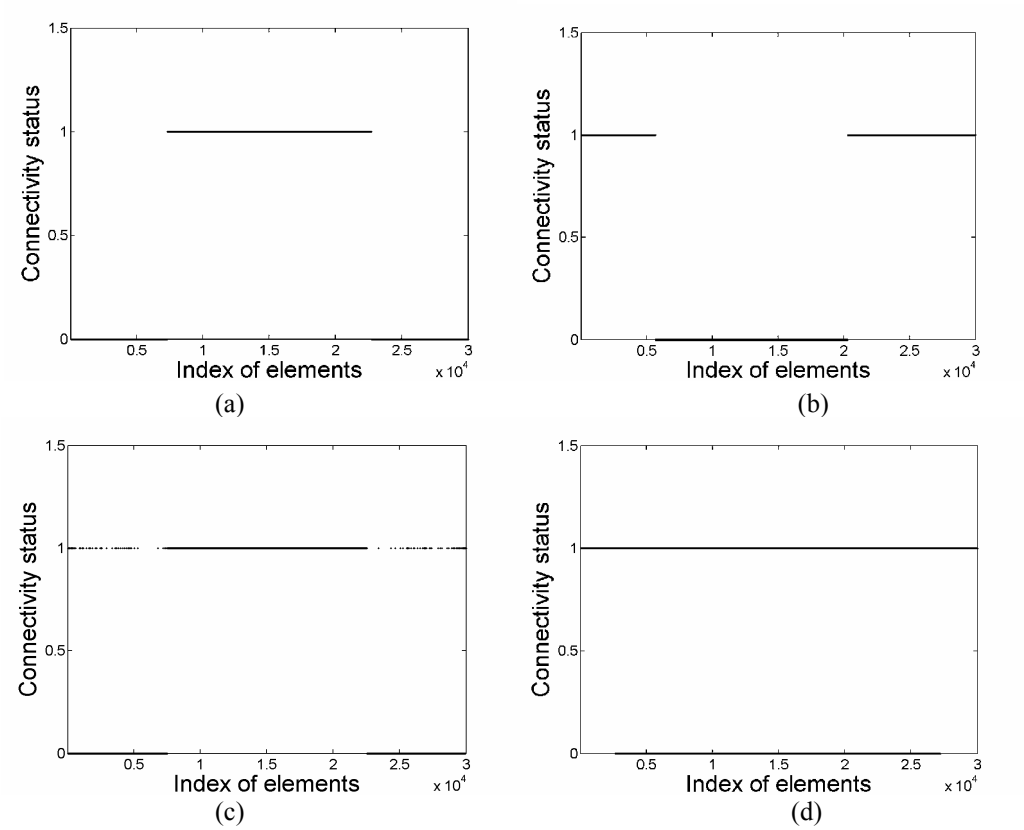

Figure 1. Spectral views of AS connectivity in two datasets: Connectivity status of elements of the eigenvector corresponding to the second smallest (a) and the largest eigenvalue (b) in Route Views, and the second smallest (c) and the largest eigenvalue (d) in RIPE. Because of the sample size (30,000 points), the seemingly connected lines in (d) are actually composed of many disconnected segments. Eigenvectors corresponding to distinct eigenvalues can capture various characteristics of the analyzed data.

7. REFERENCES

[1] Autonomous System Numbers: http://www.iana.org/assignments/as-numbers.

[2] Route Views project: http://www.routeviews.org.

[3] Réseaux IP Européens: http://www.ripe.net/ris.

[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," *Proc. of ACM SIGCOMM'99*, Cambridge, MA, Aug. 1999, pp. 251-262.

[5] H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Towards capturing representative AS-level Internet topologies," *Proc. of ACM SIGMETRICS 2002*, New York, NY, June 2002, pp. 280-281.

[6] D. Vukadinovic, P. Huang, and T. Erlebach, "On the spectrum and structure of Internet topology graphs," in H. Unger et al., Eds., *Innovative Internet Computing Systems, LNCS 2346*. Berlin: Springer-Verlag, 2002, pp. 83-96.

[7] M. Mihail, C. Gkantsidis, and E. Zegura, "Spectral analysis of Internet topologies," *Proc. of Infocom 2003,* San Francisco, CA, Mar. 2003, vol. 1, pp. 364-374.

[8] L. Gao, "On inferring autonomous system relationships in the Internet*," IEEE/ACM Transactions on Networking*, vol. 9, no. 6, pp. 733-745, Dec. 2001.

[9] G. Huston, "Interconnection, peering and settlements-Part II," *Internet Protocol Journal*, June 1999: http://www.cisco.com/warp/public/759/ipj_2-2/ipj_2-2_ps1.html.

[10] C. Alaettinoğlu, "Scalable router configuration for the Internet," *Proc. IEEE International Conference on Computer Communications and Networks*, Washington DC, Oct. 1996.

[11] F. R. K. Chung, *Spectral Graph Theory*. Providence, Rhode Island: Conference Board of the Mathematical Sciences, 1997, pp. 2-6.

[12] A. Pothen, H. Simon, and K.-P. Liou, "Partitioning sparse matrices with eigenvalues of graphs," *SIAM Journal of Matrix Analysis,* vol. 11, no. 3, pp. 430-452, July 1990.

[13] M. Fiedler, "Algebraic connectivity of graphs," *Czech. Math. J.*, vol. 23, no. 2, pp. 298-305, 1973.

[14] J-P. Benzécri, *Correspondence Analysis Handbook.* New York: Marcel Dekker, 1992, part I.