Correlation Analysis of Complex Network Metrics on the Topology of the Internet

Alberto Garcia-Robledo and Arturo Diaz-Perez
Information Technology Laboratory
Cinvestav-Tamaulipas
Cd. Victoria, Mexico
{algarcia, adiaz}@tamps.cinvestav.mx

Guillermo Morales-Luna
Computer Science Department
Cinvestav-IPN
Mexico City, Mexico
gmorales@cs.cinvestav.mx

***Abstract*—We present an experimental study on the linear relationship between a rich set of complex network metrics, to methodologically select a subset of non-redundant and potentially independent metrics that explain different aspects of the topology of the Autonomous System view of the Internet. We followed a data-driven approach based on (1) a correlation study of different properties of evolving Internet networks, and (2) the validation of a non-redundant set of metrics by evaluating the performance of supervised and unsupervised machine learning techniques. We confirm pair-wise metric correlations observed in other types of networks and identify sets of highly correlated metrics that may reveal patterns specific to the topology of the Internet.**

Keywords—Internet; complex networks; autonomous systems; correlation analysis; machine learning

I. The Autonomous System View of the Internet

The past decade witnessed an increasing interest in the reconstruction of a variety of technological, social, and biological phenomena using complex networks. This interest has been motivated by factors such as a growing availability of ever-increasing computing resources and the discovery of common properties in real-world networks across different disciplines. Complex networks have been used to model the macroscopic structure of massive technological phenomena, including the topology of the Internet at the Autonomous System (AS) level.

An AS is a group of millions of host IP addresses (routing prefixes) that share common routing policies. ASes interact with each other through a massive network of thousands of links to form an AS Network (ASN). An ASN communicates millions of host IP addresses across the world. From a technological perspective, the measurement of the Internet may be considered a big data problem, and involves topology measurements for the discovery (and inference) of AS connections that, viewed as a whole, reveal important details of the functional organization and evolution of the Internet [1].

A variety of Internet measurement initiatives have been proposed in the form of complex network metrics that characterize different topological aspects of ASNs for a wide variety of purposes [2]. Unfortunately, the selection of a definitive set of metrics for the characterization of ASNs is hampered by the existence of a wide variety of redundant metrics that unintentionally explain the same aspects of complex networks.

There are efforts [3], [4], [5], [6], [7] that study the pair-wise correlation between complex network metrics in order to select a subset of non-redundant metrics, i.e. metrics that hold little correlation to each other. However, it is not clear if metrics correlation patterns inferred from datasets of networks with very different sizes and from different application domains are valid for the characterization of ASNs. We believe that the characterization of ASNs should be performed using independent (uncorrelated) metrics. These metrics should be size-independent and obtained exclusively from ASN datasets.

The objective of this paper is to present an experimental study on the correlation between widely used complex network metrics. We identify potentially independent metrics that explain different aspects of the AS view of the Internet. We followed a data-driven approach based on: (1) measurements of different properties on real-world ASNs and (2) the evaluation of the "expressiveness" of uncorrelated metrics on machine learning techniques.

II. Related Work

Currently, there is no a consensus on a definitive set of metrics that provide a "complete" characterization of real-world complex networks like the Internet. Nonetheless, there are methodological efforts [3], [4], [5], [6], [7] that, given an initial set of redundant metrics, allow us to obtain a subset of potentially non-redundant metrics with similar "descriptive capacity" than the initial set.

These efforts follow a data-driven approach that investigate possible pair-wise relationships between the considered measurements on an ensemble of complex networks from one or different application domains. Most of these efforts [3], [4], [5], [6] conclude that many widely-used metrics are strongly correlated and identify sets of similar network properties by looking at correlation heat maps. In addition, some works [3], [5], [8] use techniques borrowed from the data mining realm, such as principal components analysis (PCA), to identify further metric correlation patterns and discover classes of similar networks in spaces of reduced dimensionality.

These works provide evidence that many of the metrics used for the study of the evolution, modeling, and reduction

978-1-4799-2546-9/13/$31.00 ©2013 IEEE

of ASNs could be redundant. For example, in [3] is found that many metrics are highly correlated on Barabási-Albert networks. These networks are generated from a widely-known preferential attachment model used to replicate the scale-free topology of the Internet. Likewise, in [5] and [7] are investigated metric correlations in a variety of technological, social, biological, and linguistic networks; and sets of highly correlated metrics are found.

Unfortunately, it is not completely clear if the metric correlations found in [5], [7] on ensembles of networks with very heterogeneous structure can be directly applied to ASN datasets. In [3] is shown that correlation patterns for a network model do not necessarily agree with the patterns that arise when different models are put together. Likewise, in [6] is warned against mixing network ensembles from different applications domains to extract global metric correlation patterns. Moreover, in [4] are documented analytical results on metric correlations on Erdös-Rényi, Barabási-Albert, and Watts-Strogatz networks, and it is concluded that the degree distribution have influence on the metric correlations.

On the other hand, in [4] is shown that metric correlation patterns change with the size and density of graphs. Likewise, in [6] is shown that size-dependent metrics distort the correlations between metrics. ASN datasets are composed of snapshots of the same complex phenomena taken at different times. This introduces the need for the normalization of size-dependent metrics in order to analyze ensembles of networks with varying size. However, some of the current works [5], [7], [8] do not consider that some of the studied metrics are size-dependent.

We believe that there is a need for the methodological study on the linear dependences of complex network metrics performed exclusively on ASNs datasets in order to: (1) confirm (or refute) metric correlation patterns observed in other studies, and (2) discover metric relationships that help us to distinguish ASNs from networks of other application domains.

III. Selection of the Dataset and Metrics

We based our study on the RV/RIPE dataset of ASNs [9]. The dataset includes Border Gateway Protocol (BGP) AS-paths obtained from raw BGP table dumps from two major publicly available collectors: Route Views (RV) and RIPE. The dataset provides 51 ASNs corresponding to the evolution of the Internet from January 1998 to January 2010. As we will develop in more detail in Section VI, the study of the RV/RIPE dataset is interesting because it models a very-specific ASN phenomenon: the Internet growth settlement since 2001 due the telecoms market crash [9].

We studied the correlation patterns of 19 connectivity, clustering, distance, centrality, hierarchical, and scaling complex network metrics. The selected metrics have been used to study different aspects of the Internet topology, such as its evolution during the last decade [9], its modeling through ASN network models that abstract its "essence" [10], its robustness and breakdown against failures [11], and its simplification for the acceleration of the simulation of Internet protocols [12].

Scaling metrics are the Pearson correlation of the log-log plot of the rich-club, average clustering coefficient, average neighbor connectivity, and cumulative degree scaling with the degree of vertices. These scalings have been used to characterize the hierarchical structure, scale-freeness, and disassortativity of ASNs [2]. We considered the average of the vertex betweenness, clustering coefficient, closeness, neighborhood connectivity, and eigenvector centrality metrics over all vertices. We also assessed the average of the edge betweenness centrality over all edges. Likewise, we considered the average of the hierarchical clustering coefficient metric over all vertices at distances 2 and 3. We also measured the density, assortativeness, density, average path length, and the diameter of ASNs. Finally, we considered the rich-club of the 5% of the richest vertices.

TABLE I.
Normalization functions f_μ of size-dependent complex networks metrics, where n is the number of vertices of an ASN [6].

Metric μ	f_μ
Average neighbor connectivity	$(n-1)$
Diameter	$(n-1)$
Average vertex betweenness centrality	$2/(n-1)(n-2)$
Average edge betweenness centrality	$2/n(n-1)$
Maximum degree	$(n-1)$
Average path length	$(n-1)$

Let $G=(V, E)$ be a network or graph with vertex set V and edge set E. Let $n=|V|$ be the number of vertices of G. Let $\mu(G)$ be the value of the metric μ evaluated on the graph G. Let s be the number of ASNs in the dataset. Let $(n_1, n_2, ..., n_s)$ be the sizes of the network in the dataset.

We normalized the measurements following the procedure of Bounova *et al.* [6] as follows. The measurements of the i^{th} size-dependent metric $(\mu_i(G_1), \mu_i(G_2), ..., \mu_i(G_s))$ are normalized by the expression $(\mu_i(G_1)/f_{\mu i}(n_1), \mu_i(G_2)/f_{\mu i}(n_2), ..., \mu_i(G_s)/f_{\mu i}(n_s))$, where $f_{\mu i}$ is a normalization function of the metric μ_i that depends on n_i. In this way, the measurements become independent of the network size. Table I shows a list of size-dependent metrics μ used in this study and the functions f_μ used for their normalization. We only considered the simple, undirected, and unweighted version of the giant component of each ASN.

IV. Correlation Study of Complex Network Metrics

Let $M=(\mu_1, \mu_2, ..., \mu_m)$ be the vector of the $m=19$ selected complex network metrics. We obtained an $s \times m$ feature matrix X; where the i^{th} row of X is the feature vector $X_{i,*}=(\mu_1(G_i), \mu_2(G_i), ..., \mu_m(G_i))$ that characterizes the i^{th} ASN G_i in the dataset. We standardized X by subtracting to each matrix value the corresponding column mean (centering) and then by dividing each matrix value by the corresponding column standard deviation (normalization).

Let $X_{*,j} = (\mu_j(G_1), \mu_j(G_2), ..., \mu_j(G_s))$ be the j^{th} column of X, i.e. the values of the metric μ_j evaluated on every ASN in the dataset. We assessed the degree of pairwise dependence between the m metrics by calculating an $m \times m$ symmetric correlation matrix, C; where the matrix value $C_{j,j'} = |\mathrm{COR}(X_{*,j}, X_{*,j'})|$ is the absolute Pearson correlation coefficient between the columns $X_{*,j}$ and $X_{*,j'}$. Thus, $C_{i,j}$ quantifies the linear dependence between the metrics μ_i and μ_j.

Fig. 1 shows a bar plot of the average pair-wise correlation for each metric, i.e. the average of the correlation values between a given metric and any other metric. We observed that the assortativity coefficient was the metric that presented the lowest correlation values to other metrics, with an average of 0.1893 +/- a standard deviation of 0.1054. This is consistent with existing results [6] that show very low pair-wise metric correlations for the assortativity coefficient on ensembles of random Erdös-Rényi and Barabási-Albert networks. This may suggest a connection between random networks and RV/RIPE ASNs. In [6] is concluded that degree correlation metrics, such as the assortativity, are useful in distinguishing network topologies. According to our results, the conclusion in [6] may apply to ASNs too.

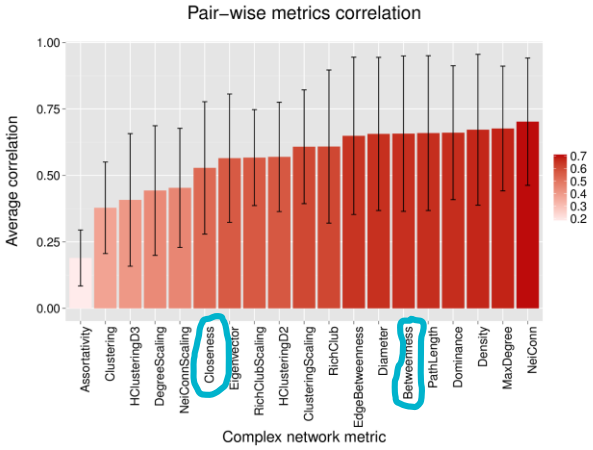

Figure 1. Bar plot of average pair-wise metric correlations. Error bars denote the metric average correlation +/- a standard deviation.

On the other hand, we observed that the average neighbor connectivity was the most correlated metric (0.7023 +/- 0.2395), followed by the maximum degree (0.6765 +/- 0.2345), and the density (0.6718 +/- 0.2836). Note in Fig. 1 that the top-8 most correlated metrics are normalized versions of size-dependent metrics (with the exception of the central point dominance). This suggests that the size of the networks still has influence on the characterization of the ASNs, even after the normalization of the measurements and the standardization of the feature matrix X.

Fig. 2a shows a heat map visualization of the correlation matrix C. By using a hierarchical clustering algorithm we permuted the rows and columns of the heat map so that the most correlated metrics were placed close together. The heat map shows a block of seven highly correlated metrics that potentially express the same topology aspect of the evaluated ASNs. Fig. 2b shows a weighted graph representation of C, in which is clearer that distance metrics are highly correlated correlations.

In order to better observe the most correlated metrics, Fig. 2c shows the heat map for $0.9 \leq C_{i,j} \leq 1$. The heat map shows two clusters of highly correlated metrics. Fig. 2d is the graph representation of the heat map in Fig. 2c, which offers other perspective of the two clusters of highly correlated metrics.

The three-vertex cluster represents metrics related to the most central vertex of ASNs. The maximum degree and the central point dominance reflect the maximum connectivity and the maximum betweenness of any vertex in the network, respectively. On the other hand, most of the metrics in the seven-vertex cluster describe either density or shortest-path properties of ASNs. This is in contrast to existing results [6] that conclude that density and distance metrics form two orthogonal groups of highly correlated metrics.

Some of the observed strong correlations in Fig. 2d are consistent with existing analytic and experimental results. For example, it is known that the average path length and the diameter are related [6]. Likewise, it can be analytically shown that the normalized average vertex betweenness centrality is linearly proportional to the normalized average path length [6]. On non-Internet networks, Jamakovic [5] found that several distance metrics (average betweenness and average path length) are strongly correlated. Likewise, Jamakovic found that the density is mildly correlated to the rich-club coefficient.

However, the observed strong correlation between the rich-club coefficient, the distance metrics, and the eigenvector centrality may reveal important ANS-specific patterns. The rich-club acts as a super-hub that provides many routing paths between the club members in the Internet topology, causing the average path length (and in consequence the diameter) to be very small [13]. As a super-traffic hub, the rich-club members are likely to show high betweenness centrality, handling many of the shortest routes that connect non-hub vertices. In fact, the structure of complex systems, like the Internet, may be determined by the connectivity between the rich-club vertices [14].

We obtained a reduced list of $r = 12$ non-redundant metrics $M_r \subset M$ as follows. First, we selected from Fig. 2d all isolated vertices. Then, from each connected component we selected the metric with the highest average correlation, according to Fig. 1. In this way, we selected the maximum degree from the three-vertex cluster and the average path length from the seven-vertex cluster.

We additionally selected the average eigenvector centrality because it did not show significant correlation to other metrics inside its cluster, besides the rich-club coefficient. In addition, the average eigenvector centrality showed mild average correlation in general.

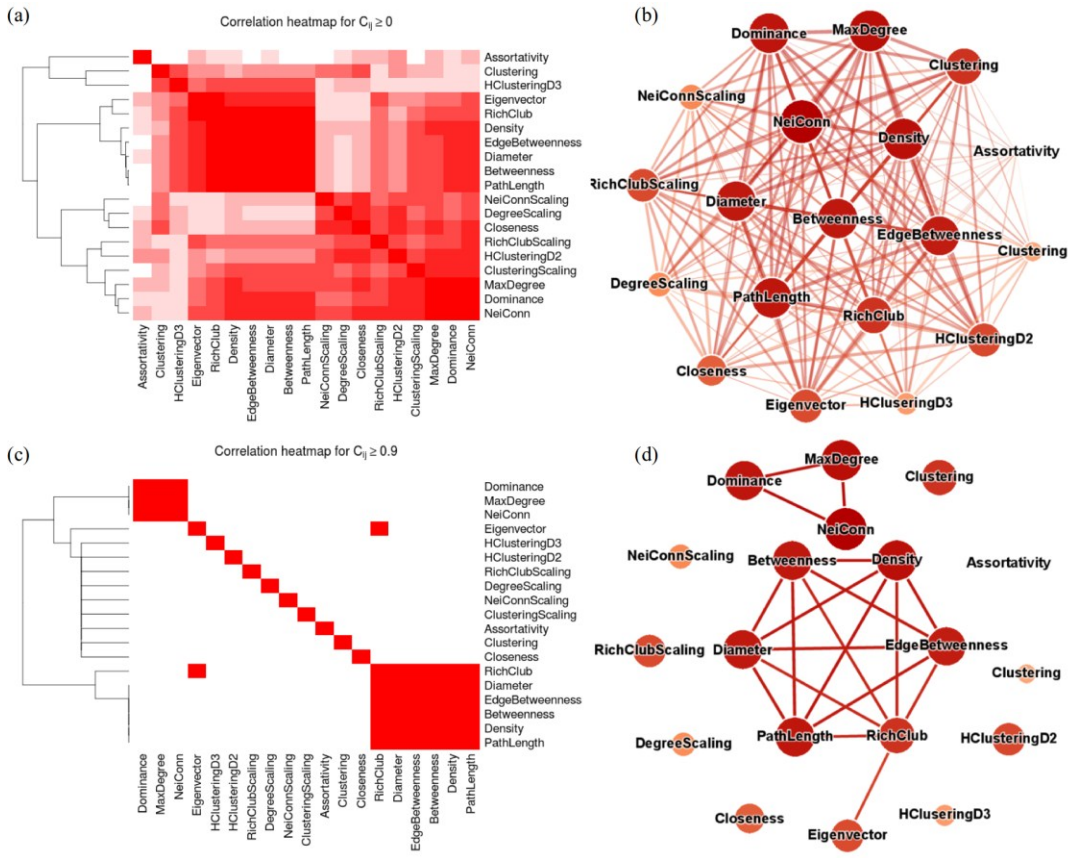

Figure 2. Left: heat maps of absolute pair-wise metric correlations on the RV/RIPE dataset. Right: weighted graph representation of the heat maps. Figs. (a) and (b) correspond to all correlation values: $0 \leq C_{i,j} \leq 1$. Figs. (c) and (d) correspond to strong correlation values: $0.9 \leq C_{i,j} \leq 1$. In Fig. (a), white colored cells denote correlation values near to 0, while red colored cells denote correlation values near to 1. In Fig. (c), white colored cells denote correlation values of $C_{i,j} < 0.9$, while red colored cells denote correlation values of $C_{i,j} \geq 0.9$. In the graphs, vertices represent metrics and edges are weighted with pair-wise correlation values. Edge color and width are proportional to the pair-wise correlation. Vertex color and size are proportional to the average pair-wise correlation.

V. Machine Learning Verification of Non-redundant Metrics

Recall that X denotes the feature matrix. Let $X_r = X_{*, Mr}$ denote the sub-matrix of X that includes only the columns of X corresponding to the metrics in M_r. Let X_r be the reduced feature matrix. Non-redundant metrics in M_r should provide roughly the same amount of information about ASNs than the full set of metrics M. To test this claim, we studied the "similarity" of the behavior of supervised and non-supervised machine learning techniques when trained with the feature matrices X and X_r, separately. We grouped, classified, and visualized in reduced metric spaces the selected ASNs and compared the results.

We experimented with two unsupervised techniques, K-means and the Ward algorithm [15], to find a varying number of clusters in the ASN dataset. Basically, what we wanted was to quantify how well we can recover the clusters found with X by training the clustering algorithms using the reduced feature matrix X_r.

We generated clusterings for different number of clusters: $K = (2, 3, 4, 5, 6, 7, 8)$. For the K-means algorithm we used the Euclidean distance and performed 100 different clusterings using the K-means++ criteria [16] to initialize the centroids. We compared the clusterings obtained with X and X_r using the adjusted Rand index (ARI), and the adjusted mutual information (AMI) similarity metrics.

Figs. 3a and 3b show the values of the ARI and AMI scores for each clustering algorithm when varying the number of clusters K. Positive values of both ARI and AMI scores, near to 1, denote a significant level of similarity. Note that the similarity is consistent across all values of K. The fact that we were able to retrieve the clusterings using X_r suggests that we do not lose much information if we use the non-redundant metric set in M_r.

Regarding supervised techniques, we used CART and Naive Bayes to find a varying number of classes in the ASN dataset using the feature matrices X and X_r, separately. Basically, what we wanted was to assess how similar is the performance of two different supervised techniques trained separately with X and X_r.

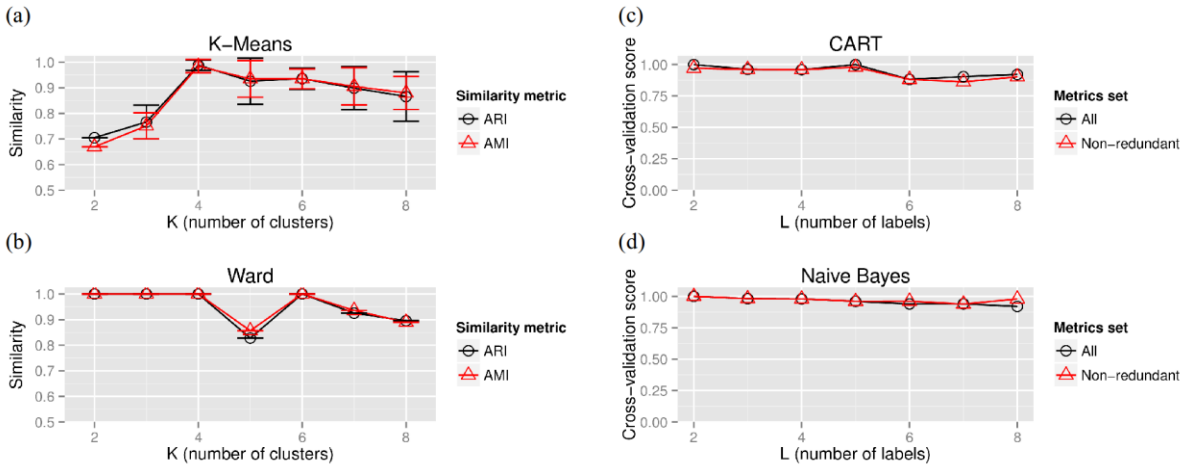

Figure 3. Left: ARI and AMI similarity scores for the *K*-means (a) and hierarchical (b) clusterings of the RV/RIPE ASNs, when varying of the number of clusters, *K*. Error bars denote the average similarity score +/- a standard deviation. Right: cross-validation scores of the CART (c) and Naive Bayes (d) classification of the RV/RIPE ASNs using *M* (all metrics) and M_r (non-redundant metrics), when varying the number of labels, *L*.

We labeled each graph with artificial labels obtained with the *K*-means algorithm. Then, we used cross-fold validation to quantify the average performance of the classifiers trained with *X* and X_r. Figs. 3c and 3d show that the two classifiers obtained a precision greater than 75% across different number of labels *L* when trained either with *X* or X_r. The similarity of the performance of classifiers provides more evidence on the fact that we do not lose much information if we use the reduced set of metrics M_r to characterize ASNs.

VI. PCA Projection and Visualization of ASNs

PCA [17] is a non-supervised technique for reducing the number of features (dimensions) while retaining most of the original variability in the data. Features are linearly combined into orthogonal variables called principal components (PCs). PCA is a useful tool to visualize multivariate data in two or three dimensions.

Descriptive non-redundant metrics should allow us to discriminate among different ASN datasets, and this should be verifiable by using PCA visualizations. To validate this claim, we characterized ASNs from other real-world datasets: CAIDA [18] (January 2004 to November 2007) and DIMES [19] (October 2007 to September 2010). In addition, we characterized synthetic ASNs generated by INET-3.0 [20]. For the INET-3.0 ASNs, we generated a dataset of 51 ASNs with the same number of vertices than the ASNs in the RV/RIPE dataset and used the default values for the other model parameters.

Fig. 4a shows the PCA projection into the first two PCs of all datasets using *X*. We observed that the datasets grouped into separated clusters of ASNs. Note that the artificial INET-3.0 ASNs was the most separated cluster. Fig. 4b shows that the PCA projection using X_r successfully preserved the cluster organization induced by the two first PCs.

The AS-level Internet has gone through two growth phases: an initial exponential phase up to mid/late-2001, followed by a slower exponential growth thereafter [9]. This trajectory change, caused by the well-known telecoms market crash in 2001 [21], can be easily observed in the PCA visualization, as shown in Fig. 4c. Fig. 4d shows that the PCA projection using X_r successfully preserved the pre-2001 and post-2001 differentiation of the RV/RIPE ASNs.

Recall from Section III that we only considered normalized versions of size-dependent metrics. However, the spatial distribution of RV/RIPE ASNs in Figs. 4c and 4d suggests that the size of ASNs, characterized by size-dependent metrics, still plays an important role for the characterization of this kind of evolving networks.

Finally, note that the organization of the ASN datasets into separated clusters in Fig. 4a suggests that the combination of ASNs from different sources may not represent a good statistical ensemble of complex networks. In consequence, a correlation analysis should be performed separately for each dataset [6]. An exception may be the RV/RIPE and the CAIDA ASNs, that showed some degree of overlapping when characterized by *X*. However, further study is required.

VII. Conclusions

In the presented work we revealed clear relationships between subsets of complex network metrics on the RV/RIPE ASN dataset. We found that distance and density metrics are highly correlated on ASNs. Likewise, metrics related to the dominance of the most important vertex also showed strong linear dependencies. These correlation patterns may reflect specific and important details about the Internet topology. We also found that size-dependent metrics, even when normalized to be size-independent, still play an important role in expressing the "essence" of the structure of the Internet.

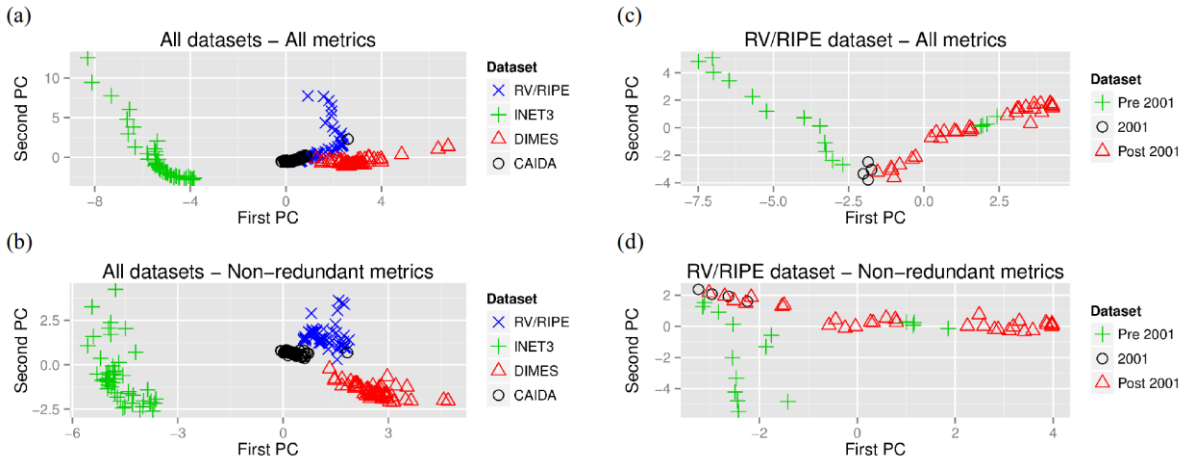

Figure 4. Left: PCA visualization of the RV/RIPE, DIMES, CAIDA, and INET-3.0 datasets using the first two PCs obtained with (a) M and (b) M_r. Right: PCA visualization of ASNs from the RV/RIPE dataset by using the first two PCs obtained with (c) M and (d) M_r.

We also provided experimental evidence that the absence of highly redundant metrics do not remarkably affect the final results in tasks such as clustering, classification, and PCA visualization of ASNs. Finally, PCA projections using redundant and non-redundant measurements suggested that a separated correlation analysis may be needed for every different ASN dataset.

ACKNOWLEDGMENT

The authors acknowledge to the General Coordination of Information and Communications Technologies (CGSTIC) at Cinvestav for providing HPC resources on the Hybrid Cluster Supercomputer "Xiuhcoatl", that have contributed to the research results reported within this document.

REFERENCES

[1] K. Cho, "Internet measurement and big data", *Internet Infrastructure Review*, vol. 15, no. 1, pp. 31–34, 2012.

[2] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2007.

[3] L. Costa, P. Villas Boas, F. Silva, and F. Rodrigues, "A pattern recognition approach to complex networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2010, no. 11, p. P11015, 2010.

[4] C. Li, H. Wang, W. de Haan, C. J. Stam, and P. Van Mieghem, "The correlation of metrics in complex networks with applications in functional brain networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 11, p. P11018, 2011.

[5] A. Jamakovic and S. Uhlig, "On the relationships between topological measures in real-world networks," *Networks and Heterogeneous Media*, vol. 3, no. 2, pp. 345–359, 2008.

[6] G. Bounova and O. de Weck, "Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous Graph Ensembles," *Physical Review E*, vol. 85, no. 1, p. 016117, 2012.

[7] V. Filkov, Z. Saul, S. Roy, R. D'Souza, and P. Devanbu, "Modeling and verifying a broad array of network properties," *Europhysics Letters*, vol. 86, no. 2, p. 28003, 2009.

[8] L. Costa, F. Rodrigues, G. Travieso, and P. Boas, "Characterization of complex networks: a survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[9] A. Dhamdhere and C. Dovrolis, "Twelve years in the evolution of the Internet ecosystem," *IEEE/ACM Transactions on Networking*, vol. 19, no. 5, pp. 1420–1433, 2011.

[10] S. Zhou and R. Mondragon, "Accurately modeling the Internet topology," *Physical Review E*, vol. 70, no. 6, p. 066108, 2004.

[11] B. Mirzasoleiman, M. Babaei, M. Jalili, and M. Safari, "Cascaded failures in weighted networks," *Physical Review E*, vol. 84, no. 4, p. 046114, 2011.

[12] V. Krishnamurthy, M. Faloutsos, M. Chrobak, J. Cui, L. Lao, and A. Percus, "Sampling large Internet topologies for simulation purposes," *Computer Networks*, vol. 51, no. 15, pp. 4284–4302, 2007.

[13] S. Zhou and R. Mondragon, "The rich-club phenomenon in the Internet topology," *IEEE Communications Letters*, vol. 8, no. 3, pp. 180–182, 2004.

[14] X. Xu, J. Zhang, and M. Small, "Rich-club connectivity dominates assortativity and transitivity of complex networks," *Physical Review E*, vol. 82, no. 4, p. 046117, 2010.

[15] J. Ward Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[16] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," in *Proceedings of the 2007 Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[17] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 2002.

[18] CAIDA, "The CAIDA AS relationships dataset, January 2004 to November 2007."

[19] Y. Shavitt and E. Shir, "DIMES: let the Internet measure itself," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 15, pp. 71– 74, 2005.

[20] J. Winick and S. Jamin, "Inet-3.0: Internet topology generator," tech. rep., University of Michigan, 2002.

[21] *Telecoms: The Great Telecoms Crash*. The Economist, July 18th 2002.