DOI: 10.15849/IJASCA.240330.08

Int. J. Advance Soft Compu. Appl, Vol. 16, No. 1, March 2024
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Enhancing the Accuracy of Internet Autonomous System Clustering Using K-Means Algorithm

Sajidah Shahadha Mahmood

Department of Radio and Television Journalism, Collage of Mass Media, University of Al Iraqia, Baghdad, Iraq.
Email: sajidah.sh.mahmood@aliraqia.edu.iq

Abstract

Internet autonomous system (AS) graph can gain more insight of how the Internet infrastructure is evolving. To construct this AS, the relations between ASes have to be harvested. Subsequently, graph theory has to be leveraged to study the characteristics of the constructed graph. One of the main properties of the AS graph is the type of relations between the ASes and their classification. In this work, AS graph has been constructed from two data sources; locking glass servers and PeeringDB. Subsequently, Gephi has been utilized to extract the characteristic of the constructed graph. Finally, the extracted characteristics have been fed into a K-mean model that clusters the ASes in the graph into three tiers according to the relation's inferred. Our results show that K-mean can infer the tier of each AS with an accuracy of 88%. Moreover, our results have shown that the Eigen value graph metric can be utilized as the clustering features without other features.

Keywords*: Graph theory, K-means, Eigen Value, Graph Metrics, Centrality Metrics, Autonomous System (AS)*

1 Introduction

Internet is defined as a complex massive network of networks. These networks are connected with different fiber optics and satellites' connections all over the globe. Internet is the infrastructure of Internet of things (IoT) and Internet of everything (IoE) paradigms [1]. It carries different applications, data and information all over the world. To gain more insight of the Internet infrastructure, its development and its future, this massive network should be modeled [2]. Modeling the Internet has many methods and layers. It can be modeled as overlay applications as in peer-to-peer applications, such as, PPTV [3], crypto-currencies [4] and VoIP [5]. It can be modeled as a number of networks 'companies' that provided services to other networks and users. These networks or enterprises are called autonomous systems (ASes). These ASes are under the administration of a single entity and they offer different services to the users. Internet service providers (ISPs) are an example of these ASes. Each one of these ASes has a unique identifier number, called ASN. This number is utilized to route the traffic

Received 10 November 2023; Accepted 19 February 2024

between these ASes since the exterior gateway protocol (EGP) does not leverage IP address for routing. The de facto EGP protocol in the Internet is the boarder gateway protocol (BGP) that creates an AS path between any two ASes over the Internet [6]. The ASes over the Internet are classified even as customers or providers. The relations between these ASes define one AS as a customer for the service of another AS. However, another type of relations between ASes occurs, the P2P relations. In this relation, a free agreement is signed between the ASes to allow the traffic to move free of charge between them. These relations classified the ASes over the Internet into three main tiers. Understanding these tiers is important information to know where to peer and where to connect in the Internet. However, there is no complete list that shows the tiers that an AS belongs to.

A third and final method to model the Internet is at the router level. This level is a complex level and it can gain more insight of the technical details of the Internet and how to route the data inside the ASes and between them. However, it is hard if not impossible to create a complete router model of the Internet.

To study the AS model of the Internet, the relations between ASes have to be harvested. Subsequently, graph theory has to be utilized to study the properties of the constructed graph [7]. These properties have shown the small world phenomenon of the Internet, relation types and even how the Internet is converting to a flat structure [8].

In this work, we attempted to cluster the ASes in the AS graph into their tiers based on unsupervised K-means machine learning algorithm. An AS graph has been constructed from two main sources; locking glass servers and PeeringDB. Subsequently, Gephi has been utilized to analyze the constructed graph to extract five different graph metrics; degree, triangles, closeness, betweenness and Eigen value. Finally, the calculated graph metrics have been used as features to K-means algorithm to cluster the ASes into one of three tiers. Our contribution in this work can be summarized as follows:

- Constructing an AS graph. Python has been utilized to telnet 32 locking glass servers to save the BGP dumps located in these servers. Subsequently, another Python code has been written to harvest the data from Peering DB to enhance the accuracy of the constructed graph
- Gephi has been leveraged to construct the graph to calculate different graph metrics
- A K-means algorithm has been trained with the data extracted from Gephi to cluster the ASes into one of three tiers. The model has been written in Python.

The rest of this paper is organized as follows; section II summarizes some of the related works that have been conducted in the area of AS modeling of the Internet. Section III overviews the theoretical background of graph theory and K-means algorithm. Section IV introduces the experiment and the results. Finally, we conclude this paper in section V.

2 Related Works

Studying the Internet and modeling it as AS relations have attracted the researchers over the years. In [9], the authors have shown a study of AS evolving over the time utilizing data

harvested over 17 years. Many insights have been shown of Internet ASes. However, the authors of this work did not generate graphs of their data and did not study the relations between the ASes. In [10], the author interested in studying the Latin America AS connections over public and private exchange points (IXPs). In [11] and [12], the author attempted to study Indonesian Internet at the AS level and their IXPs connections. In [14], the authors performed a study a historical study of the evolving of AS IXPs over the years. Moreover, they wrote a neural network model to predict the traffic volume of these IXPs. However, no graph theory has been utilized in these works. In [13], the author surveyed all the works that have been conducted in the area of studying the Internet at the AS level and they have shown the opportunities and the challenges that encounter this modeling of the Internet. However, none of these studies attempted to study the Internet as a mathematical graph.

In [15], the authors attempted to find the AS relations type utilizing graph theory and machine learning. A soft clustering algorithm has been utilized in this work. An accuracy exceeded 90% has been reported. However, the author utilized an uncompleted list of AS relation's for the AS relations calculations. Moreover, they have shown that betweenness and closeness centralities are good metrics in AS clustering process. Our work differs from this work in three main folds. First, the constructed graph is bigger and constructed from more sources, such as, PeeringDB that have been utilized in studying the AS peering relations [16], [17] and [18]. Second, the comparison list in this work is more compete. Finally, we have shown in this work that inly the Eigen value can be utilized for the clustering purpose of AS into their tiers.

3 Theoretical Background

In this section, the definition of graph and its metrics is shown. Moreover, K-means algorithm is introduced.

3.1 A Graph

A graph is a data structure that consists of nodes and links or relations that connects these nodes. Any two nodes are connected if a relation occurs between them. The graph is classified as directed and undirected. In the directed graph, the relations between the nodes are directed from one node as a parent to the second node as a child. In other words, the relation can be passed in only one direction. On the other hand, the relations in the undirected graph can be passed from both sides. The graph data structure is one of the most data structures in computer science. It is the main theory in the network science field [19]. Many mathematical properties have been proposed over the years to study this data structure. These properties are called the graph metrics. Each one of these properties has a physical meaning after calculation. In this work, we will focus in five graph metrics; degree, Eigen value, betweenness centrality, closeness centrality and triangles. The definition of these metrics is as follows. Table 1 shows definitions of the variables utilized in these metrics.

Table 1: Variables' definitions

Variable	Definition
n	Total number of nodes in the graph
$d(N,y)$	Shortest path between node N and node y
$\partial_{st}(N)$	Shortest path between node S and node T that pass through node N

$E(a,b)$	The distance between data point a with the features (x1,x2,x3…) and data point b with the features (x1,x2,x3….)

Node Degree: A measure of the number of links that start or end at the node. The node degree can be general as in the undirected graph or it can be divided into in-degree and out-degree. In this work, the constructed graph is undirected graph. This means that general degree is used.

Eigen Value: Eigen value of nodes in the graph is the Eigen values of the adjacency matrix. The adjacency matrix on the other hand is a square matrix that has zeros if no relation occurs between nodes an equal a value, called the weight if a relation occurs. The Eigen value and Eigen vector has an important role in understanding linear systems.

Triangles**:** is the number of triangles that any node in the graph participates in. A triangle means that the neighbors of a node are neighbors. The number of triangles is an important metric in calculation the cluster coefficient of a graph.

Centrality metric: To measure the centrality of a node in the graph, many metrics have been proposed. The following subsection show two of them; closeness and betweenness

- Closeness: It is defined as the sum of all the shortest paths between any node and all other nodes in the graph. If this number is small, the node is more central than other nodes. Eq.1 shows how to calculate this metric.

$$C(N) = \frac{n}{\sum_y d(N,y)} \qquad (1)$$

- Betweenness: It is defined as the number of shortest paths in the graph that passes through the node. Eq.2 shows how to calculate the betweenness.

$$b(N) = \sum_{s \neq N \neq t} \frac{\partial_{st}(N)}{\partial_{st}} \qquad (2)$$

3.2 K-means

One of the oldest well-known unsupervised machine learning algorithms utilized for clustering purposes [20]. It has been adopted by different researchers as a clustering method like the work presented in [21], [22] and [23]. This algorithm works on data features as input parameters only without the target labels as in the supervised algorithms. The algorithm requires a second input with the data features, the number of required clusters ‘K’. After entering the number of clusters and the data to the algorithm, the algorithm creates ‘K’ random values called centroids. Each one of these centroids will be a central of one of the ‘K’ clusters. As any machine learning algorithm, the K-means algorithm has a training step that consists of a number of iterations. The training step differs from other supervised algorithms in the calculation process and what to calculate. In this algorithm is training process is utilized to optimize the location of the random centroid by calculating the distance between these centroids and all other data nodes in the dataset. After calculating theses distances, the dataset rows are classified to be in the cluster of the nearest centroid. Subsequently, the data nodes in each cluster are averaged to construct a new centroid rather than the random one. This process is iterated until no new updates occur on the calculated centroids. The Euclidean distance is

the most popular method for distance calculations in K-means algorithm. Eq.3 shows this metric [24].

$$E(a,b) = \sqrt{(a_{x1} - b_{x1})^2 + (a_{x2} - b_{x2})^2 + \cdots} \qquad (3)$$

After optimizing the centroids, the centroid locations are used to find to which cluster any new data points belongs by calculating the distance to each one of these centroids and selecting the nearest 'shortest distance'.

4 Experiment and Results

Our experiment in this work consists of two parts. In the first part the AS graph has to be constructed from dump data. The data has to be harvested from the Internet. The data should contain the AS and their connections with each other. In the second part, the constructed AS graph has to be analyzed to extract nodes' metrics, such as, closeness, betweenness, Eigen value, degrees and the number of triangles. Finally, the nodes' calculated metrics have to be utilized as features for K-mean algorithm to cluster the nodes into three clusters; tier 1, tier 2 and tier 3.

To construct the AS graph, first, BGP tables dumps have been collected from 32 different locking glass servers that can be found in [25]. A python code has been written to create a telnet connection to these servers and execute 'show BGP tables' and save the telnet output into files. Subsequently, the saved outputs have to be analyzed to extract the last field in the BGP table 'AS path'. Finally, the extracted 'AS paths' have been written to find the connections between different AS numbers. This operation has been repeated 32 times to reach all server. The output of these 32 operations has been merged and duplicated connections have been deleted. Second, to enhance the accuracy of the constructed AS graph, the AS peering connections have been harvested from [26]. The harvested connections have been added to the constructed list with removing of the duplicated links. Finally, A list of the AS connection types that has been reported by CAIDA and can be found in [27] has been downloaded. The connections in this list has been added to our constructed AS graph. Moreover, the list extracted from PeeringDB has been added to the CAIDA list for the type of relations between ASes. This new list has been constructed to compare our K-mean algorithm clustering output in the last step to calculate its accuracy.

The constructed AS graph has 62051 nodes and 189980 connections. Only 42070 ASes have been found in CAIDA list. However, with the list extracted from PeeringDB, the number increased to approximately 50,000 AS. The constructed list has been fed as undirected graph into Gephi, the graph analyzer tool, to find its features, metrics and different statistics. Gephi [28], has been utilized widely in studying direct and undirected graphs in different areas, such as, WSN [29], Internet graph [14], [30], Wikipedia and social media [31]. Table 2 shows the characteristics of the constructed graph. Figure 1 shows a visualization of the constructed graph.

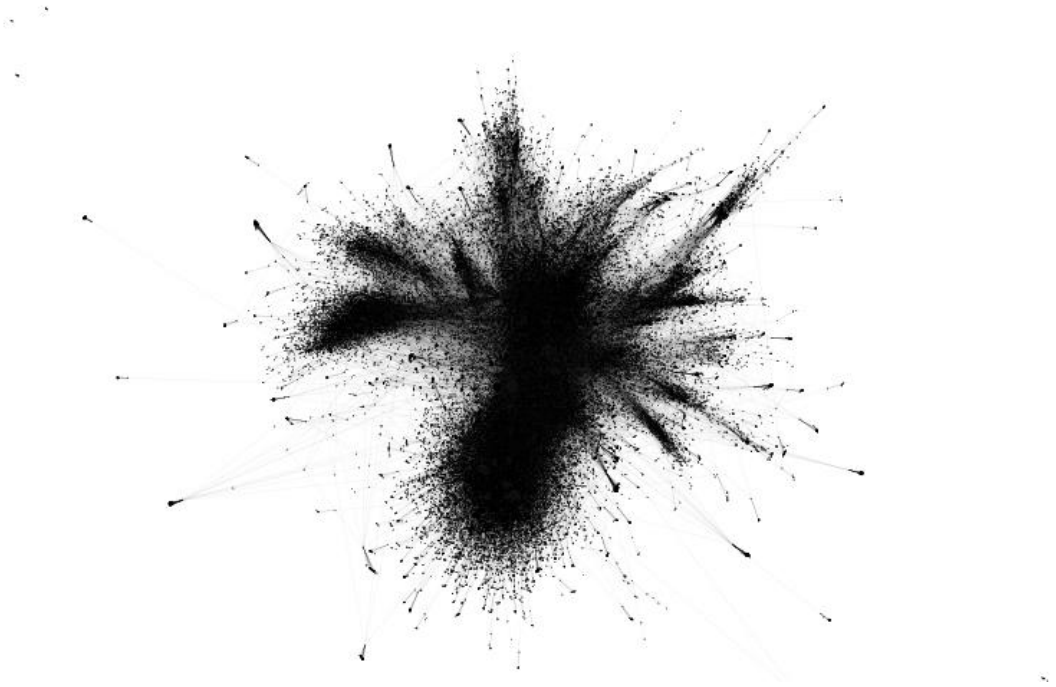

Fig. 1. The Constructed Graph in Gephi.

Table 2: Constructed Graph Properties

Graph Type	Undirected
Number of Nodes	62051
Number of Links	189980
Average degree	4.622
Graph Diameter	9
Average Path Length	5.733
Global Cluster Coefficient	0.333

Figure 2 shows the nodes degree of the ASes in the graph. We can observe from the log scale figure that few nodes have a massive nodes degree while the rest of the nodes have a small node degree value. This shows that the graph follows the small world phenomenon. The second prove on this is the value of the global clustering coefficient and the average path length calculated from the graph as shown in table 2.

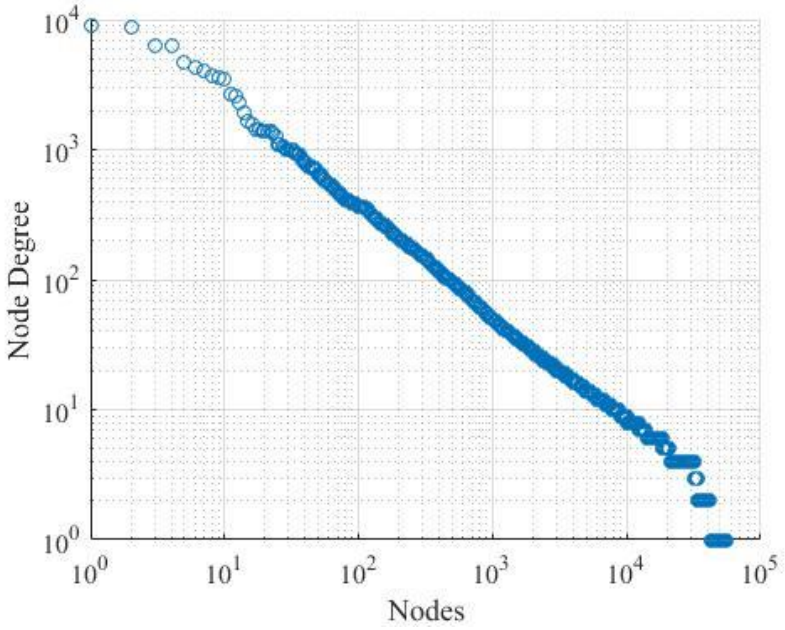

Fig. 2. Node Degree Distribution

According to [15], the most important metric in categorizing ASes is the nodes' Eigen value. We have studied the Eigen value of the nodes in this work to show the relation between this value and other graph metrics. Figure 3 shows the relation between Eigen value and node degree. We can observe from the figure that nodes with high degree have a high Eigen value. However, some nodes with a high Eigen value have a small node degree as we observe in the left of the figure.

Figure 4 shows the relation between the betweenness centrality and Eigen value. We also can observe from this figure that nodes with a higher betweenness values than other nodes have a higher Eigen value. However, it is easy to find nodes with a high Eigen value with a low betweenness value. Finally, figure 5 shows the relation between the closeness centrality and Eigen value. We can observer from this figure that with small closeness value a high Eigen value occurs. This means that the relation between closeness and Eigen value is the opposite of the other metrics.

To cluster the ASes into tiers, the constructed AS graph with all of its metrics have been fed into a K-mean algorithm. The K-mean algorithm has been written in Python. Five graph metrics have been utilized as features for the clustering purpose. These metrics are; closeness, betweenness, Eigen Value, node degree and number of triangles.

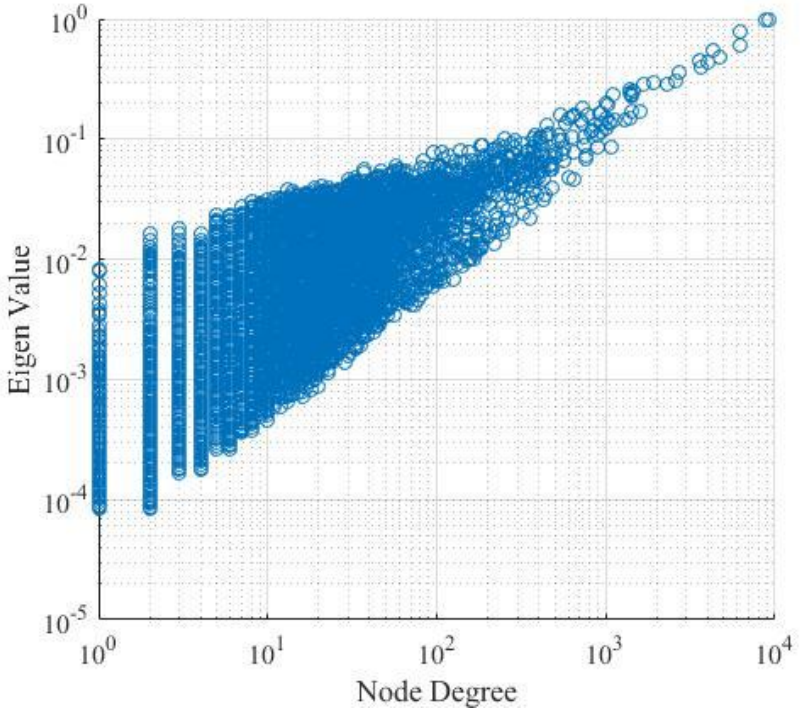

Fig. 3. Node Degree vs Eigen Value

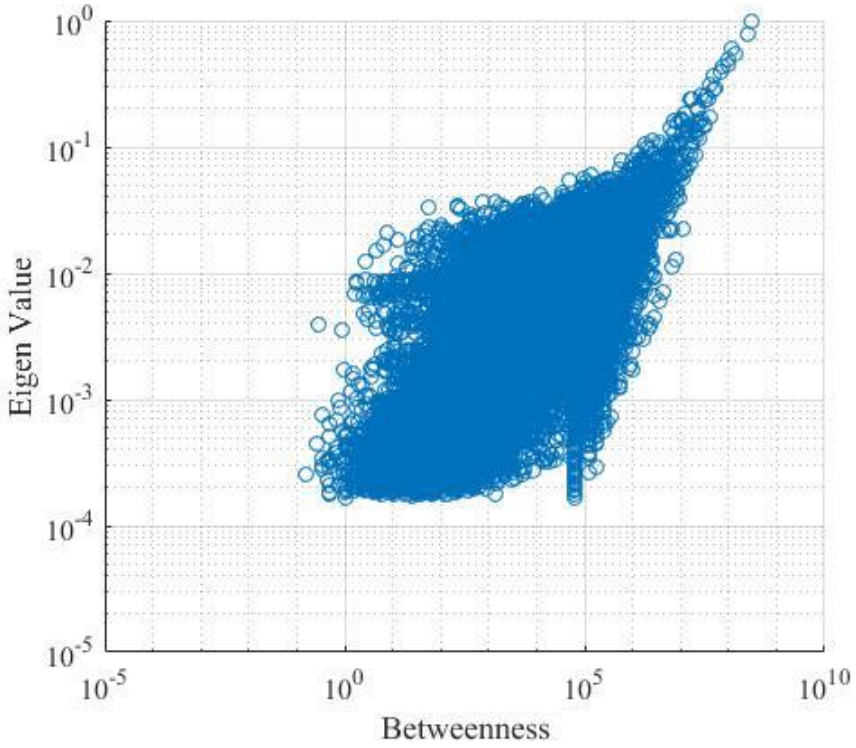

Fig. 4. Betweenness vs Eigen Value

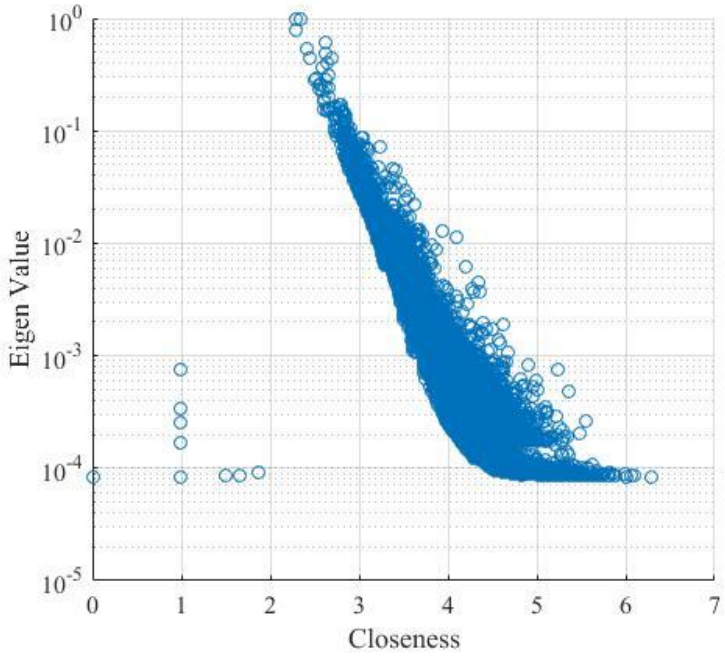

Fig. 5. Closeness vs Eigen Value

Each one of these metrics has been leveraged as a feature to the K-mean algorithm. This means that the algorithm has clustered the AS according to these features five times. In the last time, these metrics have been fed to the algorithm as 5 features. The output clusters of each run of the algorithm has been compared to the second list that has been constructed from PeeringDB and CAIDA as mentioned. Finally, the comparison has been utilized as the accuracy measure of the clustering process. To enhance the accuracy of the K-mean clustering process, 30 runs have been used for each metric and an averaging of all the 30 runs has been used.

Figure 6 shows the number of ASes in each clustering process according to the five metrics that have been leveraged. We can observe from the figure that the number of ASes in each of the tiers vary according to the metric. The figure is logged scale to show the small numbers in the other tiers. For example, when degree metric is used, the number of ASes in Tier 1 is 11295, Tier 2 is 27996 and Tier 3 is 20527. These numbers are far from the real numbers since the list of Tier 1 ASes that has been constructed consists of only 17 AS. However, when Eigen value metric is utilized, the number of ASes in these tiers are 54545, 5257 and 15 respectively, which is very close to the real numbers.

To calculate the accuracy of our clustering process, a simple mean square error (MSE) has been calculated for each clustering process. The calculated accuracies have been given in table 3. From table 2 we can observe that Eigen value is the best feature to cluster the ASes since the accuracy is 88% which is higher than other metrics.

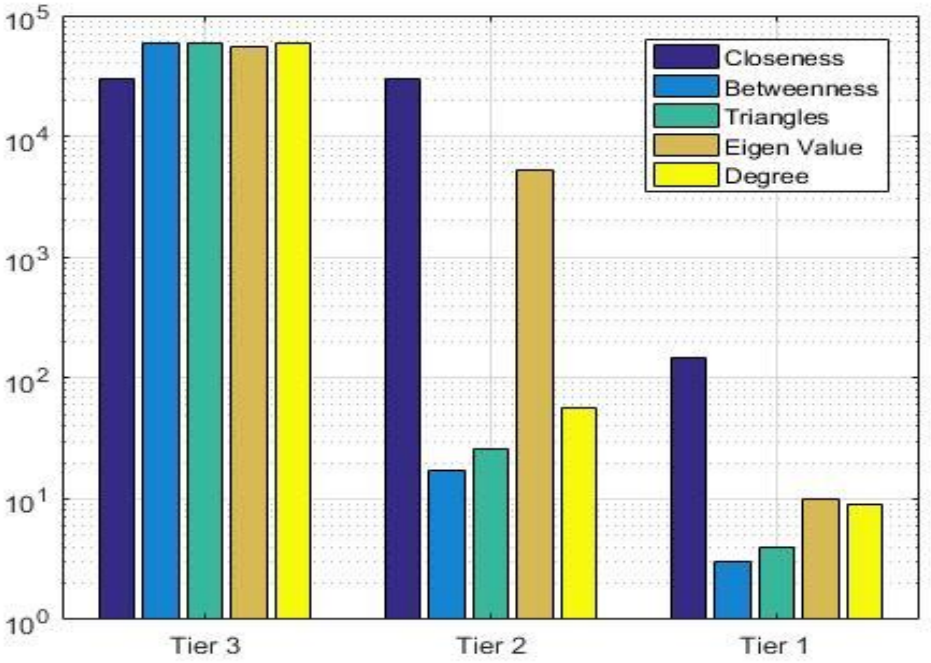

Fig. 6. K-mean Clustering Output

Table 3: The Accuracy of the Graph Metrics Clustering Process

Metric	Accuracy
Degree	55%
Betweenness	69%
Closeness	62%
Eigen Value	88%
Triangles	63%
All Metrics	69%

Finally, figure 7 shows the convergence process of the K-mean with the Eigen value feature. The algorithm has executed 50 iterations. However, the convergence occurred at 10 iterations.

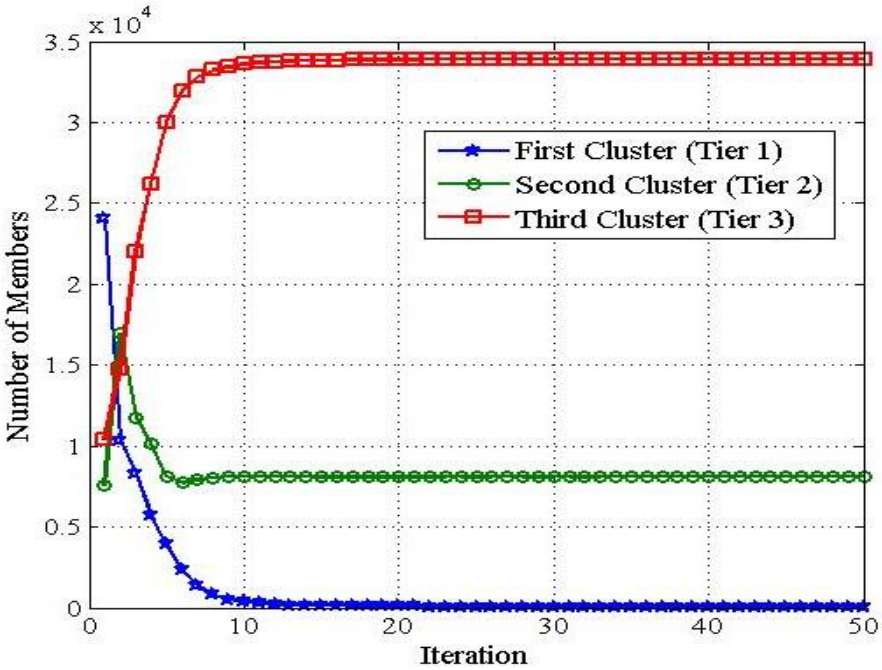

Fig. 7: K-means Convergence

5 Conclusion

Studying the AS relations has attracted the researchers over the years. In this work, the AS graph has been constructed from locking glass servers and PeeringDB harvested data. The constructed graph has been analyzed to find five important metrics; degree, triangles count, Eigen value, betweenness and closeness. Subsequently, the extracted metrics have been fed as features to K-means algorithm to cluster the Internet ASes into one of three tiers. This classification is an important process in studying the properties of the internet. Our results show that the Eigen value is the only graph metric that can be leveraged to classify the ASes into their tiers. An accuracy of 88% has been recorded for the Eigen value feature in K-means clustering algorithm.

References

[1] Masoud, M., Jaradat, Y., Manasrah, A and Jannoud,I. (2019). Sensors of smart devices in the Internet of Everything (IoE) era: big opportunities and massive doubts. *Journal of Sensors* 2019.

[2] Baldi, P, Frasconi,P and Smyth,B.(2003). Modeling the Internet and the Web. *Chichester, UK: John Wiley.*

[3] Liu, J. (2021). Case study on PPTV. *The Frontiers of Society, Science and Technology* 3(1).

[4] Delgado-Segura, S, Pérez-Solà, C, Herrera-Joancomartí, J, Navarro-Arribas, G and Borrell,J. (2018). Cryptocurrency networks: A new P2P paradigm. *Mobile Information Systems* 2018.

[5] Kara, M, Şanlıöz, S,J Merzeh,H, Aydın, M and Balık,H. (2021). Blockchain Based Mutual Authentication for VoIP Applications with Biometric Signatures. In *2021 6th International Conference on Computer Science and Engineering (UBMK), IEEE,* 133-138.

[6] Green, Th, Lambert,A, Pelsser, C and Rossi, D.(2018). Leveraging inter-domain stability for BGP dynamics analysis. In *International Conference on Passive and Active Network Measurement, Springer, Cham*, 203-215.

[7] Jaiswal, Sh, Rosenberg,A and Towsley, D. (2004). Comparing the structure of power-law graphs and the Internet AS graph. In *Proceedings of the 12th IEEE International Conference on Network Protocols*, IEEE, 294-303.

[8] Masoud, M. Z., Hei,X and Cheng,W. (2013). A graph-theoretic study of the flattening internet as topology. In *2013 19th IEEE International Conference on Networks (ICON)*, 1-6.

[9] Nemmi, E.N, Sassi,F, Morgia,M, Testart,C, Mei,A and Dainotti. A. (2021). The parallel lives of autonomous systems: ASN allocations vs. BGP. In *Proceedings of the 21st ACM Internet Measurement Conference*, 593-611.

[10] Berenguer, S and Pintor, F.V.(2018). Radiography of internet autonomous systems interconnection in Latin America and the Caribbean. *Computer Communications* 119, 15-28.

[11] Witono, T and Yazid,S.(2020). Portrait of Indonesia's Internet Topology at the Autonomous System Level." In *Computational Science and Technology, Springer, Singapore*, 237-246.

[12] Witono, T, and Yazid,S. (2020). Analysis of Indonesia's Internet Topology Borders at the Autonomous System Level. In *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), IEEE,* 545-550.

[13] Witono,T and Yazid,S. (2022). A Review of Internet Topology Research at the Autonomous System Level. In *Proceedings of Sixth International Congress on Information and Communication Technology, Springer, Singapor,* 581-598.

[14] Böttger, T, Cuadrado,F and Uhlig,S.(2018). Looking for hypergiants in peeringDB. *ACM SIGCOMM Computer Communication Review,* 48(3),13-19.

[15] Lv, T, Qin,D and Ge,L. (2018). Research and Analysis of Statistical Characteristics of Internet Exchange Points. In *2018 Sixth International Symposium on Computing and Networking Workshops (CANDARW),IEEE,* 558-560.

[16] Dey, P.K, Mustafa, S, and Yuksel, M. (2021). Meta-peering: towards automated ISP peer selection. In *Proceedings of the Applied Networking Research Workshop*, 8-14.

[17] Barabási, A.L. (2013). Network science." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371(1987), 20120375.

[18] Likas, A, Vlassis, N and Verbeek,J.(2003). The global k-means clustering algorithm." *Pattern recognition* 36(2), 451-461.

[19] Bora, D.J and Gupta, A.K. (2014). Effect of different distance measures on the performance of K-means algorithm: an experimental study in Matlab. *arXiv preprint arXiv*,1405,7471.

[20] [a]Kernen ,TH. (2011). BGP dumps. [online]: available at: http://traceroute.org/.

[21] Joldasbayev, S., Sapakova, S., Zhaksylyk, A., Kulambayev, B., Armankyzy, R., & Bolysbek, A. (2023). Development of an Intelligent Service Delivery System to

Increase Efficiency of Software Defined Networks. *International Journal of Advanced Computer Science and Applications,* 14(12), 644-656.

[22] Vikhyath K. B., & Achyutha Prasad N. (2023). Optimal Cluster Head Selection in Wireless Sensor Network via Combined Osprey-Chimp Optimization Algorithm: CIOO. *International Journal of Advanced Computer Science and Applications*, 14(12), 401-407.

[23] Qi, X., & Chen, H. (2023). Recurrence Prediction and Risk Classification of COPD Patients Based on Machine Learning. *International Journal of Advanced Computer Science and Applications*, 14(12), 840-849.

[24] PeeringDB. (2021). The interconnection Database. [Online]:available at: www.peeringdb.com

[25] Cadia. (2023). The caida as relationships dataset. [Online]. Available: http://data.caida.org/datasets/as-relationships/serial-1/20160901.asrel.txt.bz

[26] Bastian, M, Heymann,S and Jacomy, M.(2009). Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

[27] Masoud, M. Z., Jaradat, Y, Jannoud, I and Al Sibahee,M.(2019). A hybrid clustering routing protocol based on machine learning and graph theory for energy conservation and hole detection in wireless sensor network. *International Journal of Distributed Sensor Networks* 15(6)6, 1550147719858231.

[28] Masoud, M. Z., Jaradat, Y and Jannoud, I. (2017). A measurement study of internet exchange points (IXPs): history and future prediction." *Turkish Journal of Electrical Engineering & Computer Sciences* 25(1), 376-389.

[29] Masoud, M.Z., Hei, X, and Cheng, W. (2012). A measurement study of AS paths: Methods and tools." In *2012 18th Asia-Pacific Conference on Communications (APCC), IEEE*, 738-743.

[30] Geenen, D.(2020). Critical Affordance Analysis for Digital Methods: The Case of Gephi, 1-21.

[31] Masoud, M., Y. Jaradat, and Ahmad, A. (2017). Machine learning approach for categorizing internet autonomous systems' links. *ICGHIT, Hanzhou, China Google Scholar*.

Notes on contributors

Sajidah Shahadha Mahmood is an instructor at the Department of Radio and Television Journalism, Collage of Mass Media, University of Al Iraqia, Baghdad, Iraq. She was born in Baghdad, Iraq in 1977. She received her B.SC. Degree in Control and Systems Engineering\ Control Engineering in 2002 from University of Technology in Baghdad, Iraq and she received her M.SC. Degree in Control and Systems Engineering\Computer Engineering in 2020.