Connectivity Risk Analyzer: Development of an automated ETL Processs and investigation of AS-Level data

Carl Tramburg
M.Sc. Information Systems
Humboldt University Berlin
Berlin, Germany
carl.tramburg@gmail.com

David Berscheid
M.Sc. Business Administration
Humboldt University Berlin
Berlin, Germany
d.berscheid@outlook.com

***Abstract*—This paper deals with CORIA (Connetivity Risk Analyzer), a framework developed by Dr. Fabian and collegues to analyze multiple indicators explaining the risk of connections in a network like the internet. In addition to this already existing framework we aim to improve the ETL process by adding a automation to its functionality. We furthermore investigate characteristics of nodes and edges of the internet on the level of autonomous systems.**

***Index Terms*—CORIA, Connectivity Risk Analyzer, ETL, Automation, AS, Autonomous System**

I. Introduction

A. Motivation

The internet presents one of the most important networks for today‘s world. Consequently major financial and economical systems rely on its functionality and availability. In times of non-functioning of the internet, serious consequences for businesses and economies are the cosequence. There are many reasons for such a scenario, such as threats caused by nature, i.e. hurricanes or earthquakes, or deliberate hacking attacks trying to remove nodes from the network. In 2016 for example the Dyn cyberattack, which involved multiple distributed denial-of-service attacks (DDos attacks) was the reason for large unavailability of internet platforms and services in North America and Europe. It is known as the larges DDoS attack on record, involving tens of millions of IP adresses [1]. 900 000 users were infected by another attack in 2016, called Mirai Botnet, against the German company Deutsche Telekom, targeting routers and causing internet connectivity problems. [2]According to a study conducted by Ponemon Institute in 2016, the average cost of a data center outage has increased from $ 505,502 in 2010 to $ 740,357 in 2016 [3]. The implication on the importance of the internet becomes very clear, which gives high incentives to analyze the riskyness of networks, like the internet. CORIA's mission is exactely that: to help analyze connectivity risks [4].

B. Evolution of CORIA

The CORIA project started in ... with the scope of a (master) thesis by Mathias Ehlert [5]. With the objective of building a webframwork that is capable of analyzing connectivity risks of networks, CORIA 1.0 was developed. Through the access of large amount of network data, either publically available or individually provided, CORIA was able to calculate a variety of metrics, such as centrality measures or node degree measures, which then form a unified risk score for respective connectivity risks of respective nodes. In addition, CORIA offered a framework to investigate networks visually through graph visualisations. The webapplication CORIA 1.0 was written in python and ruby. It used NetworkX für its metrics and redis for database purposes [6]. After CORIA 1.5 presented a improved performance and the use of Graphtool instead of NetworkX für its metrics, Tom Kober developed CORIA 2.0 - now using Neo4j to store and manage data [7]. This version furthermore consisted of a native architecture of graph storage and processing [8]. CORIA 3.0 by Sebastian Gross benefits from a modular based improvements. It offers different levels of granularity of networks that can be investigated. Figure 1 shows parts of the CORIA dashboard and figure 2 presents an example of a graph visualisation. Version 3 allows usage of all features via one interface. Due to the modular approach, there is a clear separation of ETL process, graph analysis and exports. Amongst its

functionality is the usage of different data formats and the ability to calculate different metrics. The framework supports simultaneous execution and calculation of different metrics. Much attention was paid to a development without strong dependencies to specific technologies [9].

...

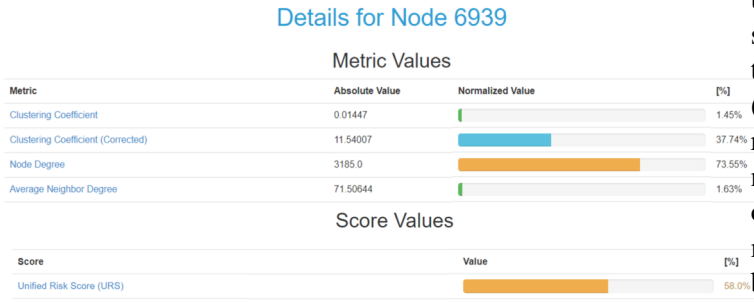

Fig. 1. Extract of CORIA Dashboard

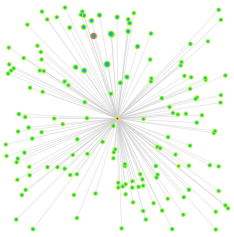

Fig. 2. Graph Presentation of Nodes in a Network

C. Objective

The CORIA framework is a project on which multiple developers already contributed to, in order to create a platform, which is able to analyze and visualize connectivity risks of graph data [10]. Within this paper and project we would like to further contribute and improve specific aspects of CORIA. In the following we present related work regarding network risk analysis especially on AS level. We investigate characteristics of these networks through some descriptive statistical methods and take a look at the time development of networks and its components. Then we present an approach for an automated ETL process, which shall improve and simplify the usage of CORIA and ensure most recent data being available for analysis. Herby we focus on AS-level data. Our emphasis lies on flawless usage of datasets coming from Caida that can be downloaded and imported automatically on a regular basis

II. Theoretical Background

A. Autonomous Systems

CORIA is capable of working with data on many levels of granularity. As Figure 3 models it, on the lowest level of the internet topology are Internet Protocol network interfaces (IPs). Multiple Ips can access the internet through one router (R). Then, on the next level, Autonomous Systems (AS) represent a set of routers under a single administration. The top of this hierarchy is formed by Internet Service Providers (ISP), with selfdescribing funcionality. Note that this represents a simplified model of internet topology, i.e. not specifying Points of Presence [11]. In the scope of this paper the level of Autonomous Systems is our main object of concern. Two types of relationships between ASes and ISPs are of interest in that regard: Provider-to-Customer (P2C) and Peer-to-Peer (P2P).

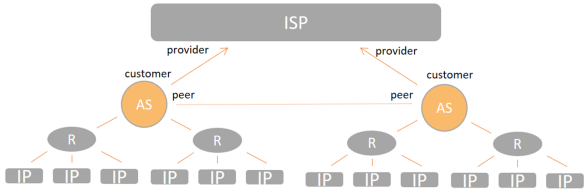

Fig. 3. Hierarchy of the Internet

Figure 4 provides an example of an AS and its characteristics. Here Kabel Deutschland, a German network operator, represents an AS. Its customer cone specifies "a set of ASes it can reach using customer links". As it is two in this case, there are only two nodes reached through customer links, meaning Kabel Deutschland itself and its provider Vodafone [23]. AS rank defines the importance of a node in its global routing system, often using customer cone information as a measure [23]. AS degree refers to the number of neighbors that a node has in a graph - here 20 [12].

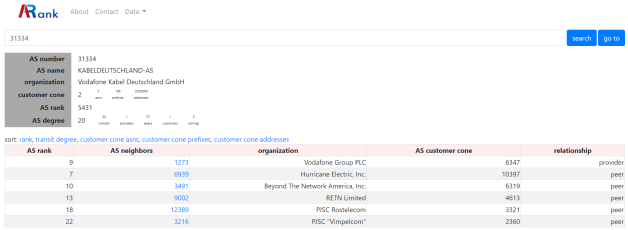

Fig. 4. Example of an AS and its characteristics

B. ETL Process

Carls part, blablabal

Extraction: from data source(s), varying data formats Transform: convert different data into proper format, that is storable Load: load into target database

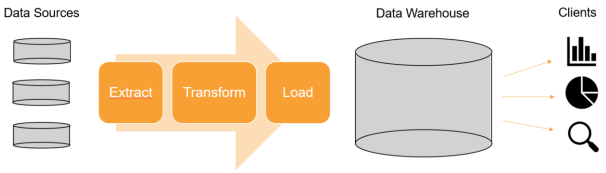

Fig. 5. Theoretical ETL Process

III. Data

As a data source for this work, we exclusively focus on Caida. Caida is an appreviation and stands for Center for Applied Internet Data Analysis. Located in the San Diego, CA in the United States, the center studies networks, its infrastructure up to a large scale. For their investigation on theoretical and practical aspects of the internet they monitor, collect and provide network data [13].

Regarding the data granularity our primary focus lies on AS level data. Within this domain, we investigate five (?) data sets. AS-Rank offers a data base giving information about specific Autonomous Systems, like its rank, relationship to other ASes or customer cone [23] . AS Classification represents a dataset including information on the business types of Autonomous Systems [21]. Trough machine-learning inference, Caida is able to offer this knowledge [21]. A dataset named Pv4 Routed 24 AS Links Dataset is used in detail to investigate the topology on the internet, its structure regarding flows of traffic, and ratios of sending and receivin autonomous systems [15]We investigate the dataset AS Relationships in order to find out about Provider to Customer relationships as well as Peer to Peer relationships [17]. Lastly, we take into account geographic information of the Autonomous Systems to draw conlclusions on regions with high imapact on the internet network and less involved ones [**?**](http://www.caida.org/data/as-relationships-geo/)

For the purpose of investigating the characteristics of Ases itself and their network, we analyzed recent data of December 2017. As part of our time series analysis, we used data from 2007 until 2017.

IV. Results

A. Data Analysis

First aspect of the data analysis process is the type of autonomous system. Caida distinguished here between three types: The first type are ASes that provide internets access or function as a transit. They make out 42.2% of all nodes in the network. Second type are ASes providing content hosting and distribution systems, like Dropbox or Google, with only 4,5% of all overall nodes. Third category are enterprises meaning organizations, universities or companies that are mostly users. They account for 53.3%. Insights we are gaining from this aspect is that most of the nodes are representing users of the internet, which makes intuetively sense. More interesting is the large amount of transits and access points needed to provide the respective infrastructure. This hints to the internet's high complexity.

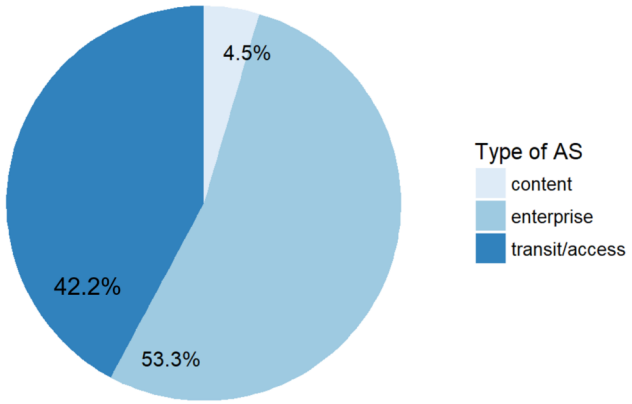

Fig. 6. Ratio of Autonomous Systems regarding their Type

As depicted in the theoretical part, autonomous systems have relationships. We took a look at this characteristic and noticed a quite balanced ratio of Provider-to-Customer relations and Peer-to-Peer relations. If we translate this into a graph, we imagine a balanced graph, which in terms of network riskiness, is rather roboust.

We furthermore built a graphs describing the distributional characteristics of autonomous systems. Figure ... shows the distribution of autonomous systems and how many outgoing connections the single nodes have. Approximately 5.500 sending nodes are containted within that data set. We ranked them on the x-axis. Wit the number of connections an AS is sending to on the y-axis, we obtrain a very left centered distribution. A very small number of nodes in the network are sending traffic to a high number of other nodes. It follows, what is called

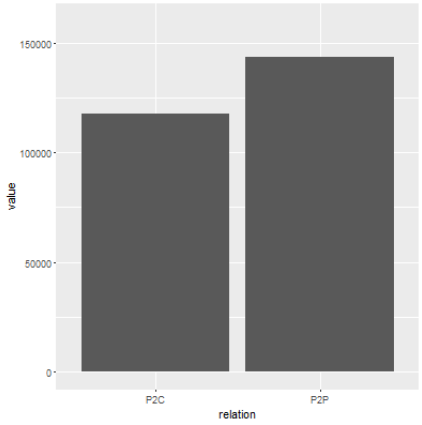

Fig. 7. Ratio of AS-Relationships

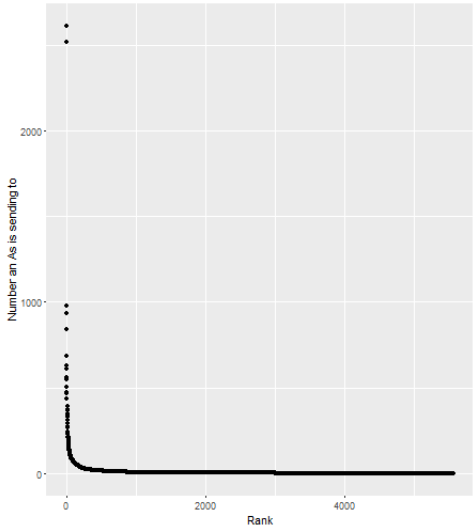

Fig. 8. Distribution of Outgoing Connections per AS

a power-law function [16]. Take the first 30 (?): ... This top ranked 30 nodes send ... % of overall trafic of the whole network. Note that when we make statements of the whole network, we only refer to the data at hand as the whole network. That is the data collected by CAIDA. Information on CAIDA's data collection process can be found in the Appendix .. (?).

Simultaneously, we drew the distribution of incomming connections per Autonomous System (Appendix ...) and received a similar result. With approximately 250.000 nodes, receiving traffic, only very few nodes are receiving traffic from a high number of nodes, again following the properties of a power-law distribution. The top ... ranked nodes are receiving ... % of the overall trafic in this network.

An important finding regarding connectivity risks of the network is, how vulnerable this network is. An attack targeting the most active and most influential nodes in the network, achieving a non-functioning of those quickly leads to wide shot-downs of the network. This again underscores the demand for CORIA - a tool to analyze connectivity risks.

Next, we analyzed the geographic locations of autonomous systems and their flow to traffic. Figure ... symbolizes autonomous systems as organge points on a word map. We can see a high concentration of ASes in North America and Europe. The concentration of ASes is less dense in areas like South America, Africa or Asia. Insights that we can draw from that representaion are that more autonomous systems are situated in highly developed economies than they are in less developed regions. This result corresponds to previous research conducted by ... [20]. While we need to keep in mind that we are only analyzing one data set - with challenging data collection on top of that, we can assume this result still to be a representative sample of the overall network.(Remark: Difficult collection of AS-data and gathering through inference [**?**]) The results confirm the assumption that most traffic is taking place between parties of richer and more developed economies. (auf punkte ohne verbindung hinweisen?)

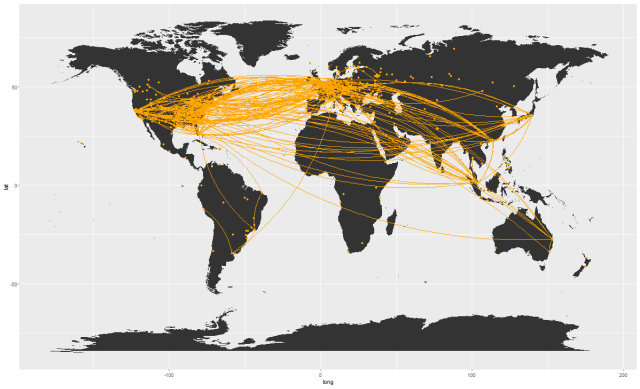

Fig. 9. Location of ASes and respective Streams of Traffic

In addition to the analysis of a single point in time, we investigated the timely development of the network. Figure .. shows the number of traffic-sending autonomous systems from the beginning of 2007 until the end of 2017. Within these 10 years, there is a clear upward trend. In 2015 this upward trend vanishes and and reaches a constant level of approximately 5.500 sending AS nodes within the network. (Note the consistency of results with respect to Figure .. 'Distribution of number

of sending ASes‘.) Striking in that graph are the multiple outliers appearing throught the time series. A qualitative research about exectional events happening on these dates that might have been the reason for a downtime of multiple nodes did not lead to reasonable results. As a consequence, we assume the outliers to be caused by monitoring and data collection problems of CAIDA. (hint to appendix with data collection problem).

With respect to Appendix ..... (other time series graph) you can see a simultaneous development for the amount of nodes, receiving traffic - on a level of approx. 250.000 nodes of course.

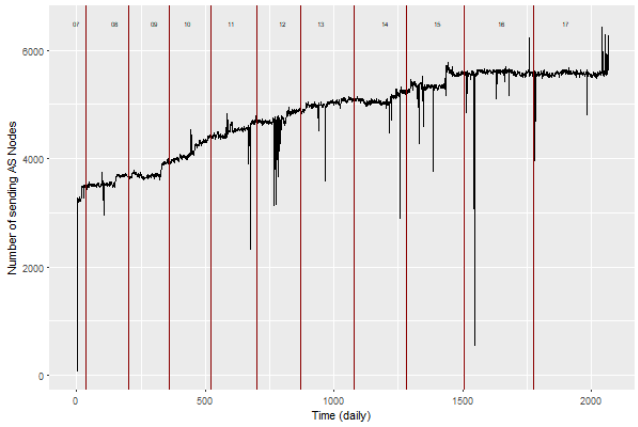

Fig. 10. Time development from 2007 - 2017: Number of sending AS nodes

B. Development of an automated ETL Process

Carls part blablablabalbal

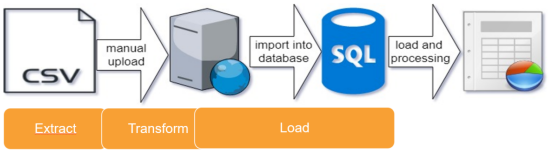

Fig. 11. A Theoretical ETL Process

V. Conclusion

A. Summary

The webapplication CORIA allows the analysis of connectivity risks of various network graphs. Through visualizations and a unified risk score it offers comprehensive results. Fokussing on the level of Autonomous Systems, our data analysis revealed a power-law distribution when it comes to the influence of ASes within the internet network - a few highly important nodes and

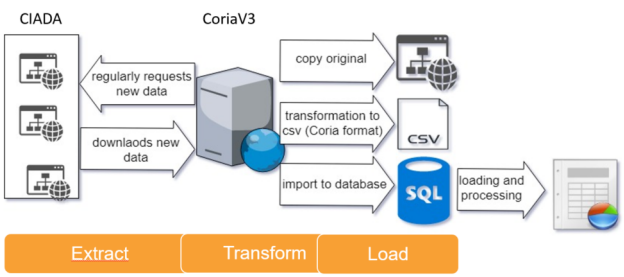

Fig. 12. ETL Process in the CORIA framework

many less important ones. Moreover the geographical investigation of its locations showed the high concentration and flow of traffic of ASes in highly developed economies, with only sparse density in less developed economies. With regard to the trend of the internet network, we exhibited a linearly growing trend of AS nodes until 2015, when the size of the network stayed constant until today.

Regarding the software-development aspect, we developed an ETL tool in Python that provides automation of the validation of latest data files, copying respective files, and transforming them into csv format. (haben wir noch mehr gemacht?)

B. Limitations and Further Work

One limitation of this work is caused by our data source Caida. The data that we used, i.e. types of Autonomous Systems or their relationship was partly gathered through statistical inference. Its machine learning classifier provided a positive predicted value of 70% [21], which still leaves misslassified observations within the data and limits our results to some extent. Nevertheless Caida is a serious and trusted source, which is why we chose it in the first place.

other limitations

Further work that can follow could be to apply the automated ETL process to further data sources, than only Caida. Also, deploying the the CORIA framework on a server would be a neccessary improvement. (CARL? Was hat es mit diesem Punkt auf sich? Läuft es nicht schon auf einem Server?) For improvements of internal development activities the creation of an extensive code and installation documentation is helpful.

REFERENCES

[1] K. York. (2016, Oct.) Dyn's Statement on the 10/21/2016 DNS DDoS Attack, [Online]. Available: https://dyn.com/blog/dyn-statement-on-10212016-ddos-attack/

[2] E.Auchard. (2016, Nov.)Deutsche Telekom attack part of global campaign on routers, [Online]. Available: https://www.reuters.com/article/us-deutsche-telekom-outages/deutsche-telekom-attack-part-of-global-campaign-on-routers-idUSKBN13O0X4

[3] Ponemon Institute LLC, "Cost of Data Center Outages," Data Center Performance Benchmark Series, 2016

[4] B. Fabian et al., " CORIA – Analyzing Internet Connectivity Risks Using Network Graphs," IEEE International Conference on Communications Paris (IEEE ICC 2017), May 2017. 10.1109/ICC.2017.7996828.

[5]

[6]

[7]

[8] D.Kiene-Maksimovic, E. Zinovyeva, J. Park, "CORIA 2.0: Augmenting a Universal Framework for Connectivity Risk Analysis," Jahr???

[9] S.Gross, "Entwicklung eines modularen Frameworks für die Analyse von Verbindungsrisiken von Netzwerken basierend auf Netzwerkgraphen," November 2017.

[10]

[11]

[12] M. Luckie et al., "AS Relationships, Customer Cones, and Validation", Internet Measurement Conference (IMC), Oct 2013, pp. 243–256.

[13] Caida, [Online] Available: https://www.caida.org/home.

[14]

[15]

[16] Yaneer Bar-Yam. "Concepts: Power Law," New England Complex Systems Institute, August 2015.

[17] X. Dimitropoulos, G. Riley, "Modeling Autonomous System Relationships," Principles of Advanced and Distributed Simulation (PADS), May 2006, pp. 143–149.

[18] X. Dimitropoulos et al., "Revealing the Autonomous System Taxonomy: The Machine Learning Approach," Passive and Active Network Measurement Workshop (PAM), Mar 2006.

[19] X. Dimitropoulos et al., "Classifying the Types of Autonomous Systems in the Internet", SIGCOMM, Aug 2005.

[20]

[21] Caida., "AS Classification," [Online]. Available: http://www.caida.org/data/as- classification.

[22] Caida., "IPv4 Routed /24 AS Links Dataset," [Online]. Available: http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml.

[23] Caida., "AS Rank," [Online]. Available: http://as-rank.caida.org/.

[24] Caida., "AS Relationships," [Online]. Available:http://www.caida.org/data/as-relationships/.

[25] Caida., "AS Relationships – with geographic annotations," [Online]. Available: http://www.caida.org/data/as-relationships-geo/.

[26] M. Oehlers, B. Fabian, "Graph Metrics for Internet Robustness: A Survey," ACM Comput. Surv. 1,1, Article 1, January 2018.

[27] J.Dümig, "Modelling of an Extraction Transformation Loading (ETL) system for the connectivity risk analyzing framework CoRiA, December 2016.

[28] T.Kober, "Business Intelligence in der Telekommunikation: Konzeption und Umsetzung einer Graphendatenbank mittels Neo4j," December 2016.

[29] A. Baumann, B. Fabian. "Towards Measuring the Geographic and Political Resilience of the Internet," International Journal of Networking and Virtual Organisations, 13(4):365–384, 2013. abstand

APPENDIX

Nummerierung vom Appendix stimmt noch nicht

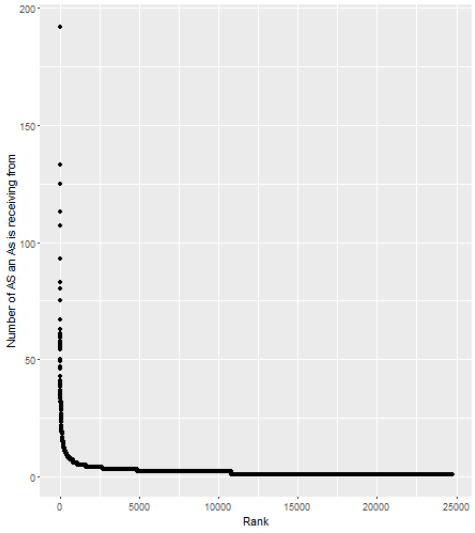

Fig. 13. Distribution of Incoming Connections per AS

Geographical Locations of ASes

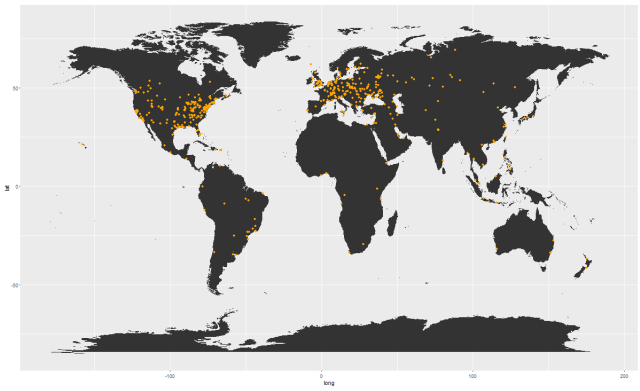

Fig. 14. Geographical Locations of ASes

Time development from 2007 - 2017: Number of receiving AS nodes

Data Collection Process of CAIDA

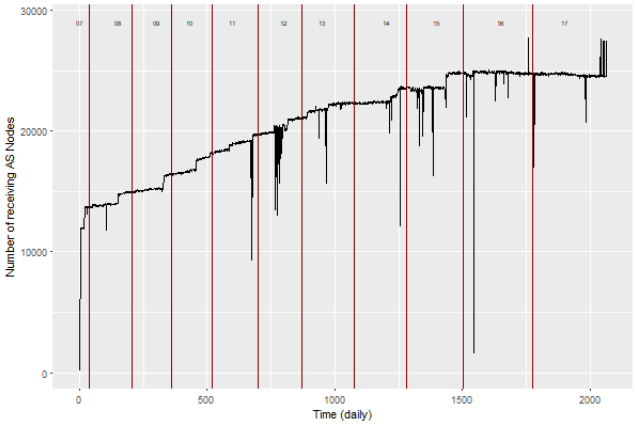

Fig. 15. Time development from 2007 - 2017: Number of receiving AS nodes