Visual Analysis of the Migration of Domains Between Autonomous Systems

Eva van den Eijnden
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
e.k.vandeneijnden@student.utwente.nl

ABSTRACT

In situations where large amounts of data need to be analyzed, classical data representation quickly becomes overwhelming. This can lead to arduous analysis processes and overlooking interesting or important phenomena in the data. In the case of the OpenINTEL dataset, which monitors the DNS records of approximately 50% of the internet's domains, this is especially an issue, as the dataset grows every day. In order to maximize the amount of useful information retrieved from the data, visual analytics can be applied. This research investigates what can be learned about a specific aspect of the OpenINTEL dataset, namely the migration of domains between Autonomous Systems, through visual analysis.

Keywords

visual analytics, autonomous systems, migration, domains

1. INTRODUCTION

In a time when data is overabundant, it is beyond the human capacity to comprehend its meaning and it is becoming increasingly important to enable e.g. researchers, managers or incident response teams of any kind to oversee large amounts of data quickly and efficiently in order to come to the right decision. Computers alone are rarely enough for this; while they are incredibly good at performing computations quickly and efficiently, they are not so adept at analyzing data or finding relations/connections without extensive and time-consuming training. Humans, on the other hand, are quite proficient at this, provided the data is in an easy-to-understand format. This is the central idea of visual analytics: creating an optimal collaboration between man and machine in order to gain insight in massive amounts of data. According to Keim et al,*"Visual analytics is an iterative process that involves information gathering, data preprocessing, knowledge representation, interaction and decision making."* [6]

Since March 2015, researchers from the Design and Analysis of Communication Systems (DACS) group at the University of Twente have been measuring all DNS-records of domains registered in the `.com`, `.net` and `.org` zones [8] , [9]. The research leading to these results was made possible by OpenINTEL [1], a joint project of SURFnet, the University of Twente and SIDN. The data retrieval is done on a daily basis. In order to reduce the time researchers spend analysing data and increase the time they can spend actually learning from it, the goal of the research is to find out what can be learned from the gathered DNS data through visual analysis.

Because the dataset on which the research is based is so extensive, the focus will not be on the whole dataset at this time, but rather on creating a high quality visualization of one aspect of it. This research investigates the migration of domains between autonomous systems (AS's). An AS is defined by RFC 1930 [3] as *"a connected group of one or more IP prefixes run by one or more network operators which has a SINGLE and CLEARLY DEFINED routing policy".* The AS number for each domain in the dataset is retrieved by making use of CAIDA's IP-prefix to AS mapping dataset.

This research chooses to focus on the migration of domains between AS's because there are indications that the migration of domains between AS's might be related to the enabling of DDoS protection services for those domains [5]. It is therefore likely that the visualization of this data can provide valuable insight into the goings-on of the internet.

With the massive amount of data that is being added to the dataset every day, the amount of time to analyze the full dataset grows with every day of data added. To combat this development, the research therefore proposes the employ of visual analysis in order to boost researchers' ability to glean information from the pre-existing dataset.

2. RELATED WORK

As we saw in the introduction, Keim et al define visual analytics as *"an iterative process that involves information gathering, data preprocessing, knowledge representation, interaction and decision making."* [6] This is the definition we will use in this research, as it encompasses the most common features of visual analytics mentioned in papers on the subject.

In their paper *Visual Analytics: Scope and Challenges* [6], the researchers provide an introduction to the subject of visual analytics. They start by explaining why we are in need of such techniques in the first place, then go on to give a definition of the subject and explain how it is different from scientific visualization. Moreover, they also explain the visual analytics process and possible applications and technical challenges.

Ferreira de Oliveira and Levkowitz' paper, *From Visual Data Exploration to Visual Data Mining: A Survey* [2], starts off in the same vein as the previously mentioned paper, explaining why there is a need for visual analytics, as well as providing a definition of the subject itself as

24th Twente Student Conference on IT January 22nd, 2016, Enschede, The Netherlands.

well as of its basic concepts. Furthermore, it provides two taxonomies of the techniques for visual analytics, which provides valuable insight into the world of possibilities that exist within visual analytics. Lastly, it speaks of currently (at the time of writing) ongoing research.

The proposed research is similar in intent to Lexis', as described in his paper *Identifying Patterns in DNS Traffic* [7]. While Lexis aims to find patterns in DNS traffic in order to mitigate denial-of-service attacks, and thus has a different subject to his research, the motivations are very much similar. In his paper, he explains the theoretical background to his work, as well as his exact methodology and, of course, his results.

The specifics of the dataset that is being accumulated by OpenINTEL can be found in *The Internet of names: A DNS big dataset* [8], and *A High-Performance, Scalable Infrastructure for Large-Scale Active DNS Measurement* [9]. The dataset described in this paper is the one on which this research is built.

3. RESEARCH QUESTION

The question this research will attempt to answer is: what can we learn about the movement of domains between AS's as seen in OpenINTEL's dataset by employing visual analytics? In order to answer the main research question, it was split up into three subquestions:

R.Q. 1 What are desirable properties of a well-functioning visualization of the migration of domains between AS's?

R.Q. 2 What is an adequate visualization of the raw data?

R.Q. 3 What information can be gathered from the aforementioned visualization?

4. APPROACH

In order to answer the research question (as posed in the previous section), a visualization tool was developed for the data. The tool is intended as a proof of concept, or a glimpse into what could be achieved when combining the OpenINTEL dataset with a tool for visual analysis. For that reason, the tool was built solely to visualize a subset of the OpenINTEL dataset, namely the `.com` zone, with a focus on the migration of domains between AS's.

To answer R.Q. 1, a set of requirements was set. They were developed in association with two DACS researchers who are very familiar with the dataset. As they are the intended users for the tool, it was important to take their needs and preferences into account. More on the requirements themselves can be found in the section *Requirements.*

Answering R.Q. 2 consisted of the actual design and implementation of the tool. A thorough explanation on the visualization system itself and the design choices that were made can be found in the section *Visualization System.*

R.Q. 3 is answered by the analysis in the section *Validation.* It is a determination of the level of fulfillment of the requirements and a performance analysis. This process did not include any user testing, as the intended product is meant to be a proof of concept, to see if such tools can provide meaningful insights, rather than a production-ready system. The sections *Conclusion* and *Future Work* predictably provide the final answer to the main research question and a number of things that could be improved in the future, respectively.

5. REQUIREMENTS

As mentioned in the previous section, the requirements for the developed tool were determined in association with two DACS researchers. The requirements are outlined in the list below. The visualization tool must be able to:

1. **Visualize the development of the data over time** Evidently, observing the migration of domains is only possible if one can see the changes over time, thereby making it a requirement for the visualization tool.

2. **Handle at least a years' worth of data** To optimize the timespans over which patterns can be observed, it was determined that the tool should be capable of handling at least a years' worth of data. This way, researchers can at the very least see if there are periods in a year in which there is a noticeably higher or lower migration rate than usual, rather than being restrained to mere weeks or months.

3. **Clearly visualize the change in the dataset** The tool should clearly visualize the change in the number of domains in each AS. It is not useful to merely visualize the number of domains that an AS contains, because this type of information could be much more easily gleaned from a table or a datasheet. However, visual representations are especially useful for detecting change, which a classic data representation would be much less adept at.

4. **Have easy-to-use controls** The tool must be quick and easy to use, so the users' focus may be on the data represented, rather than on working the tool itself.

5. **Be responsive** For the same reasons as requirement number 4, the tool must be responsive, meaning that users can change visualization settings in real time. This way, the tool can support the discussion in real time, rather than having users wait for the numbers to be crunched.

6. DATA RETRIEVAL

Since March 2015, OpenINTEL has been doing *"daily active measurements of all domains in the main top-level domains (TLDs) on the Internet (including `.com`, `.net` and `.org`, together comprising 50% of the global DNS name space)"* [9]. To see OpenINTEL's current coverage, refer to the *Coverage* tab on their website [1]. Due to these extensive measurements the dataset can be used to track change over time. In order to visualize these potential developments, the data present in the dataset needs to be extracted and manipulated into a workable format.

The tool that was developed handles data retrieval through the following pre-existing infrastructure; a Jupyter notebook that connects to an also pre-existing Hadoop Cluster on which the dataset resides. The script executes an SQL query that extracts, for each AS in the database, the number of second-level domains in it. This means that for example, if an AS contains both `mail.example.com` and `www.example.com`, this is only counted once. Once the SQL query had been executed, the data is written to a JSON file. This format was chosen because it was convenient for the actual visualization, which will be discussed in the next section. Furthermore, the AS's are order by descending size in this file, which is favorable for the visualization system. The tool was designed to have a clear division between data retrieval and data visualization, however the visualization techniques applied uses the order of the source file to order the objects to visualize. This results in a descending AS order being more aesthetically pleasing in the visualization.

The tool creates a separate file for each day of data that the user wishes to process. This helps satisfy requirement 2, as the script simply loads the appropriate file whenever it is visualizing a certain day. This means that it is mostly irrelevant to the script how many days of data are present in the tooling.

The resulting files can then be downloaded from the Jupyter server onto the local host. This host need not be special in any way: all it requires is a web browser and an internet connection. Once the data is present on the host, the visualization system covers the rest of the process.

It is important to note that as of now, the tool is restricted to a subset of the OpenINTEL dataset, namely the .com zone data. This was deemed sufficient functionality for a prototype application, however it is desirable to expand this to the other DNS zones that are being monitored by OpenINTEL. As of now, one zone could be easily switched out for another in the SQL query. Allowing the user to alternate between two zones in the visualization system would require slightly more effort, but could still be set up quickly.

7. VISUALISATION SYSTEM

The visualization system uses the data that was retrieved in the first part of the process and displays it in a web browser. To do this, the tool uses a combination of HTML, CSS, JavaScript and D3, a JavaScript library used for data visualization. This set of techniques was chosen so that the users could rely on the familiarity of a web page, allowing them to focus on the data instead of the tool, as requirement 4 dictates. Furthermore, the employ of the D3 library means that the result of the visualization is a Scalable Vector Graphic (SVG), meaning a user can use the browser's zoom functions if they wish to see a certain part of the visualization in more detail, or if they would prefer to have more of a general overview of the situation.

The concept behind the visualization itself is fairly simple; as seen in figure 1, each AS in the dataset is represented by a bubble on the screen. Each bubble contains the number of the AS it represents as a label, and the size of the bubble is proportional to the number of domains that reside within it. This means that the bubble changes size whenever the number of domains in an AS fluctuates, making it easy for users to observe change, fulfilling requirement 3. The color of the bubble is dependent on the AS number, ensuring that it will stay the same color no matter how the AS develops. Colors were chosen according to the SRON color scheme [11]. In order to avoid overwhelming users, it is important not to show too many data points. A single day of data contains just under 300,000 data points, visualizing which would crowd the screen so much that the tool would become utterly useless. In order to prevent this, the tool has an automatic cutoff point; the tool draws only as many data points as is necessary to ensure that 80% of the data is covered. This cutoff percentage is based on the Pareto principle [10], which states than in general, 80% of data is covered by 20% of the data. By applying this principle, the number of data points shown on the screen is reduced dramatically, varying between 100 and 200 points drawn, depending on the day picked.

The controls are located in the top left corner of the page. The topmost control is a slider that the user can employ to select the date they wish to review. The exact date is shown just below the slider. This is an easy way for the user to select a date (requirement 1), without having to do a lot of manual typing or clicking through menu's, which helps accomplish requirement 4.

Below these inputs, the user can override the automatic cutoff of the number of AS's shown that was mentioned before. In theory, the user could choose to show anywhere from 0 to all 300,000 data points. However, in reality, requirement 5 is only satisfied if the user keeps the limit under 1000 AS's shown. If the limit is set any higher, drawing the bubbles requires so much time that the delay starts being a hindrance to the user.

Lastly, the user can choose to track a single AS. By entering the appropriate identifier in the last input field, the focus is put on the matching bubble, or none if the AS number provided is not currently in the visualization. An example can be found in figure 2. This feature allows the user to watch how a specific AS develops over time, even if it does not stay in the same place on the screen. This could happen if an AS were to shrink or grow, causing it to be either set back or put more toward the front of the ordering.

The list below shows an overview of how each requirement is satisfied.

1. Date slider control

2. Separate file for each day of data (section *Data Retrieval*)

3. Bubble size dependent on number of domains in AS

4. Familiar (web) techniques (section *Data Retrieval*), easy date selection through slider control

5. Limiting the number of data points shown

8. VALIDATION

Looking at the chosen visualization, there are roughly three types of phenomena that can be distinguished through the use of this tool that would be much more arduous using the 'raw' data.

Firstly, the tracking function in the tool makes it possible to observe an AS steadily growing/shrinking over time. For example, as is depicted in figure 3, the OpenINTEL dataset shows that AS number 13335, Cloudflare, has been steadily growing since the start of data capture. This is indicative of the growth of Cloudflare's business, ı.e. more people are using Cloudflare's DDOS protection services.

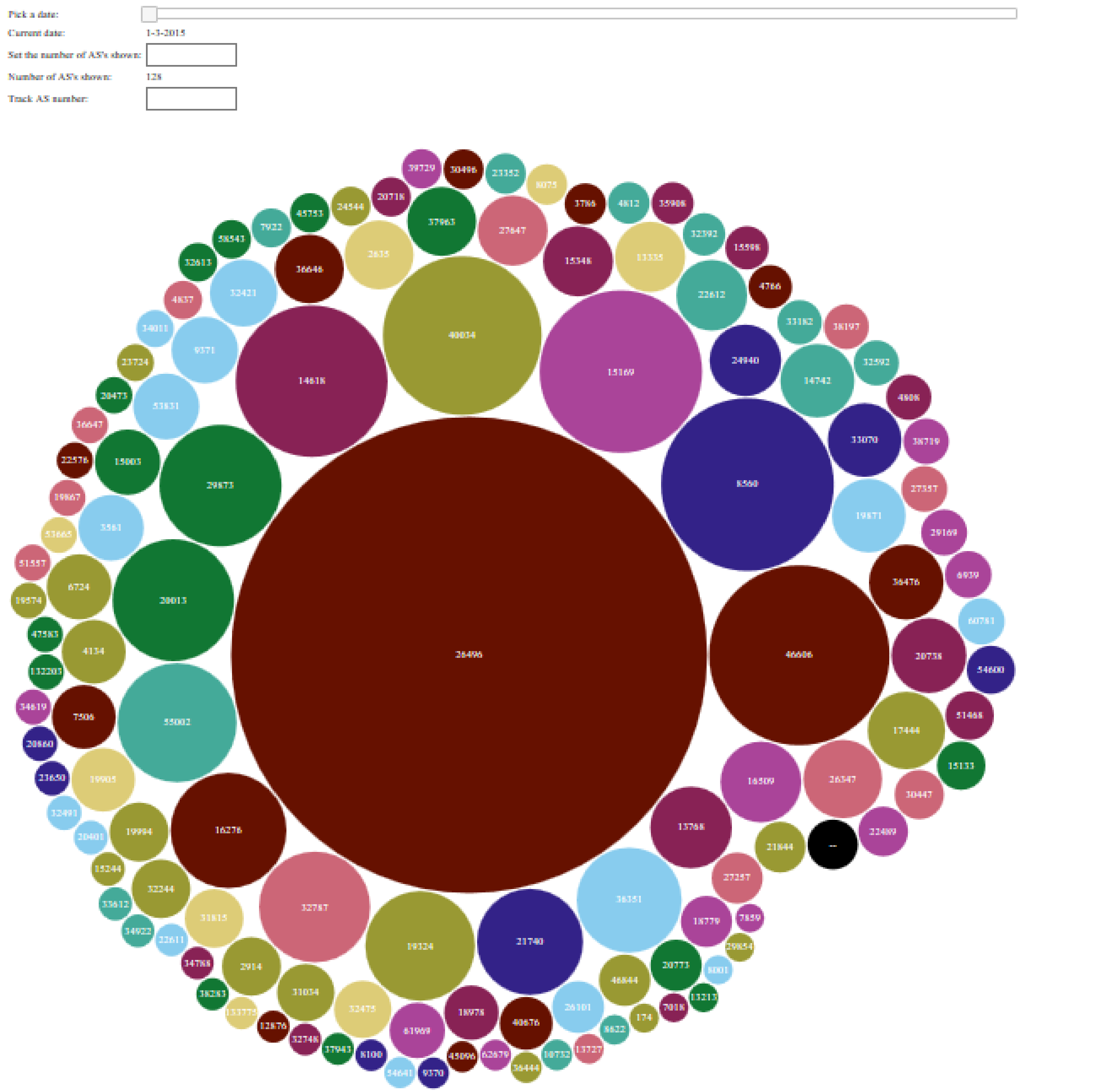
Secondly, the visualization can reveal an event that takes place in a short timeframe. Figure 4 shows such an event, which is described in [5] where a large number of domains moved from AS number 14618, Amazon, to AS number 19551, Incapsula.

Lastly, the tool makes it easy to see that some AS's, such as AS 26496, GoDaddy, are relatively constant. As shown in figure 5, it is easy for users to see the lack of change in position and/or size here.

9. FUTURE WORK

It is likely that there is more to the migration of domains between AS's than has been observed at the moment, even through the use of the tool that was developed during this research. To make optimal

Figure 1. Visualization screenshot



Pick a date:

Current date: 1-3-2015

Set the number of AS's shown:

Number of AS's shown: 128

Track AS number:

Figure 2. Tracking an AS

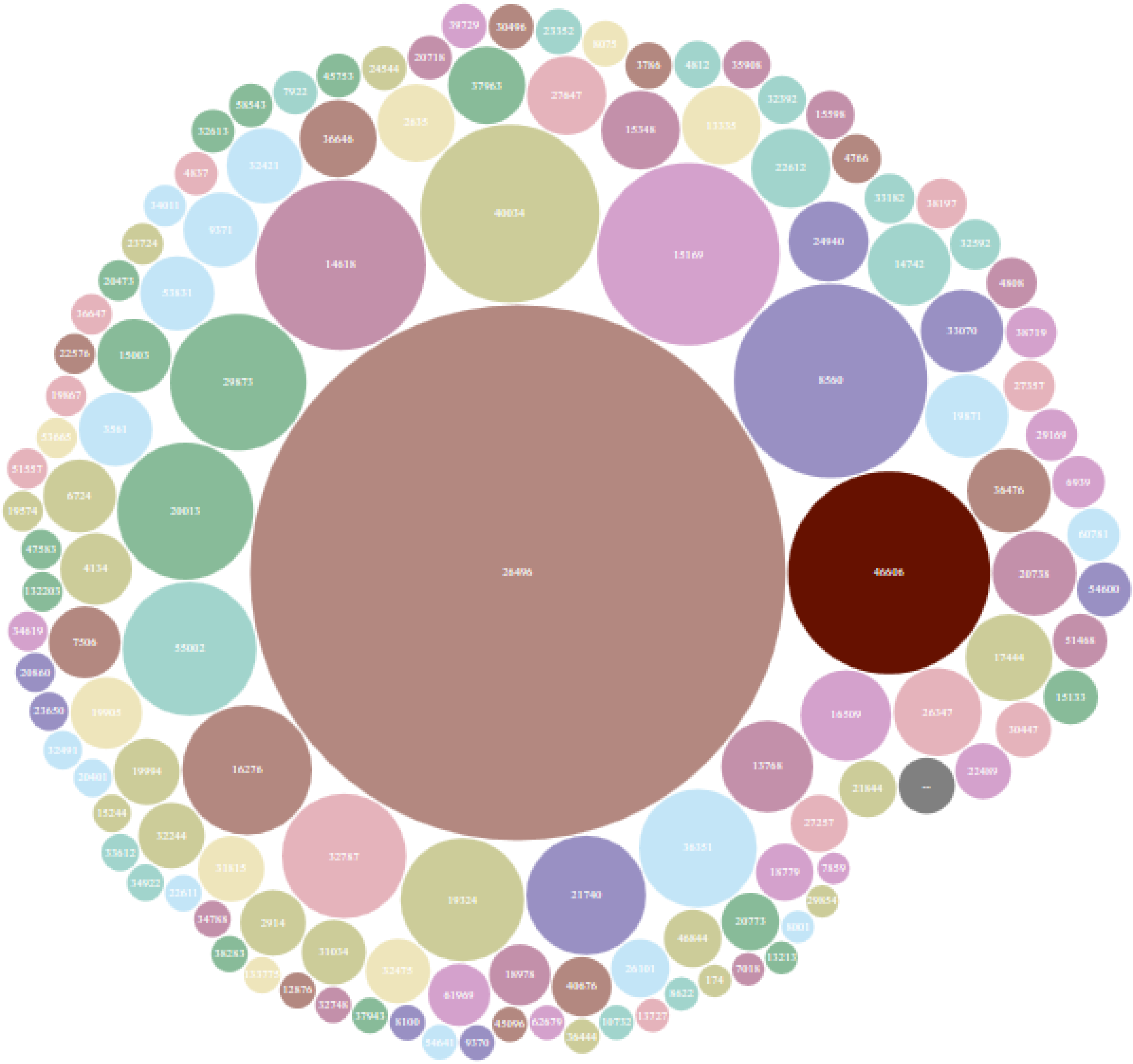


Figure 3. Steady development of an AS

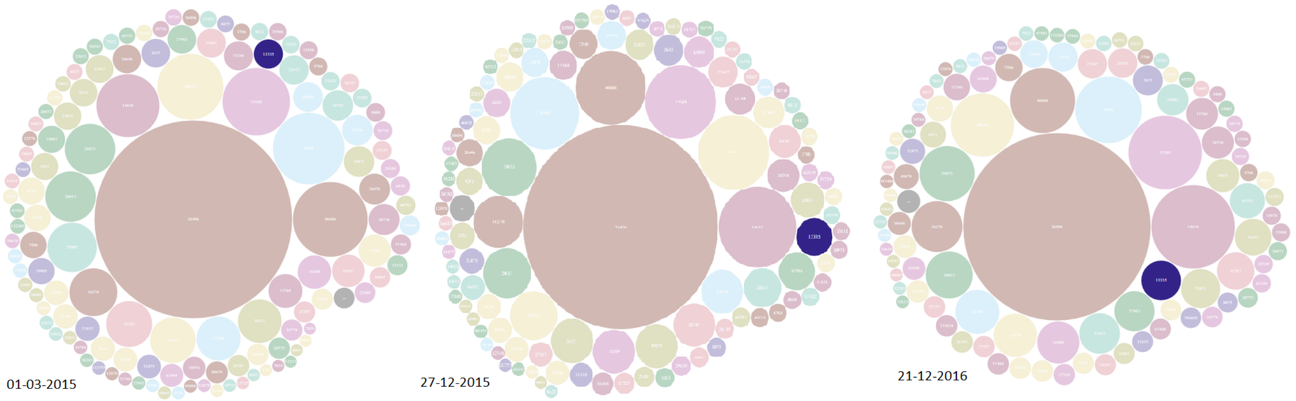


Figure 4. Sudden event regarding an AS

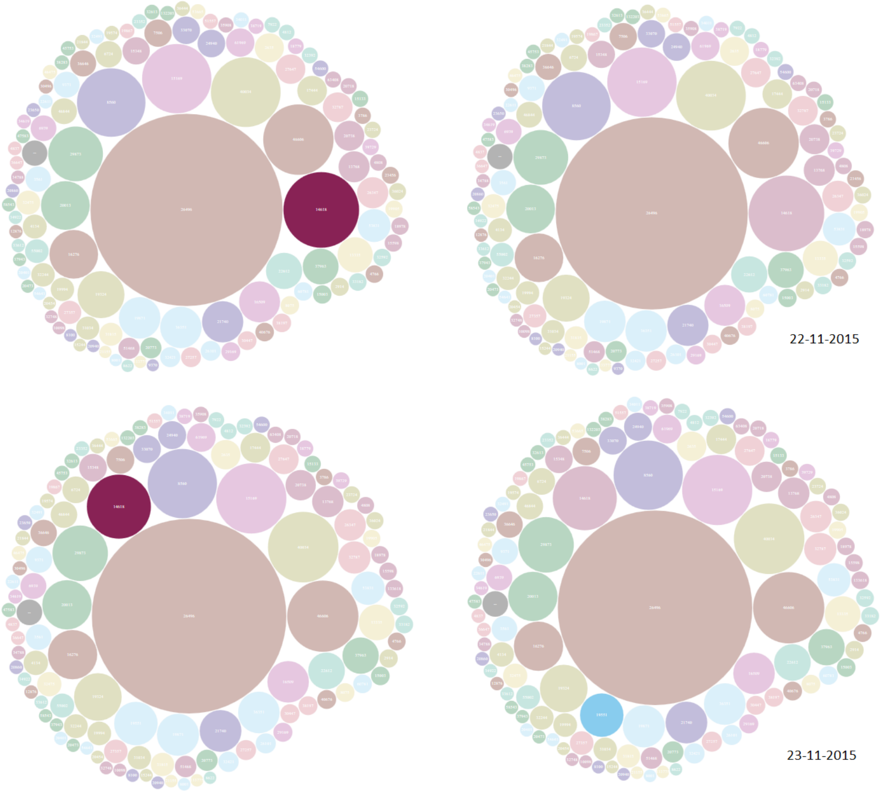
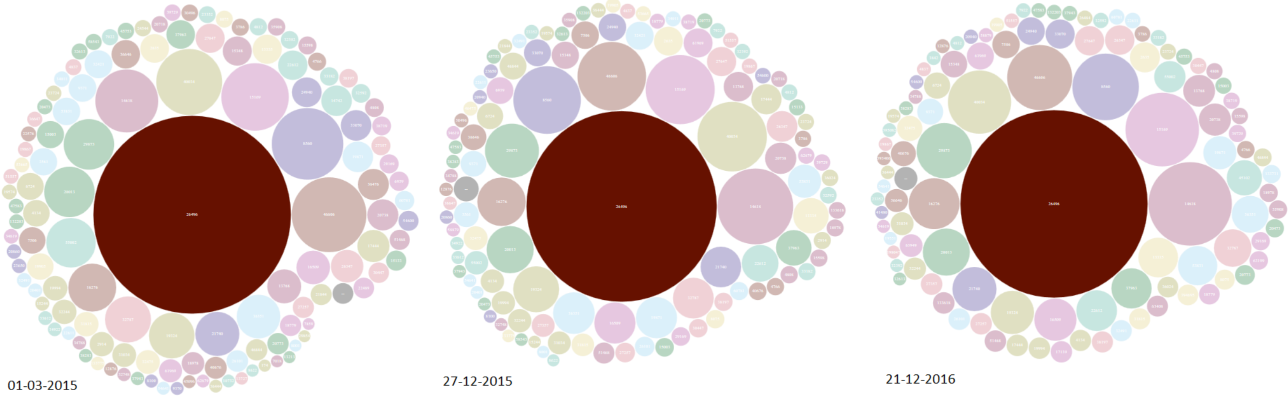


Figure 5. AS remaining constant



use of the dataset, there are a number of features that could be amended or added to the tool. A few options are:

- Allowing the user to switch between visualizations of the different DNS zones, rather than allowing only the visualization of the `.com`-zone.
- Giving users the option to show only the difference between the data of two days, either consecutive ones or dates that are further apart.
- Giving users the option to color the bubbles for growth/shrinkage compared to a certain moment in time.
- Allowing the user to switch between a linear scale for the size of each bubble and an exponential one, i.e. a scale that would emphasize the difference in size between the very small and very large AS's.

10. CONCLUSION

This research has investigated what can be learned about the migration of domains between AS's from the OpenINTEL dataset. In order to do this, a tool for the visualization of this data was designed and developed. The visualization can lead to the observation of three types of phenomena; slow developments, sudden events and constant factors. Using only the 'classic' approach to data exploration, these types of events would likely be overlooked.

Furthermore, the research has revealed that the number of domains per AS has a very long tailed distribution, as a small number of AS's (between 100 and 200) contains at least 80% of all domains in the `.com` DNS zone. This is in line with the distribution of many other variables related to the internet, such as website popularity or the pages requested from a cache [4].

While it seems that making a visualization of the data is indeed beneficial to the exploration of a dataset, there are also limits to what it can be used for. For example, in this proof of concept it was possible to cover enough of the data in one image that the user was given a balanced impression, while not being overwhelming in the sheer number of data points for the user to process. However, in a dataset whose distribution is less long tailed, several hundred data points would need to be in a single image. Putting so many data points in front of the user at one time would likely hinder the user more than it would help them create insight into the data, as it could become very overwhelming. On the other hand, restricting the number of data points shown to the user too much could lead to the oversight of interesting phenomena, as it could trick users into either not picking up on the more subtle changes in the data.

This is not a problem that can be solved by simply applying a different visualization to the data; rather it is a limitation of data visualization itself that while visualizations can provide insight into data much faster than the 'classical' representation techniques, this also means that users tend to lose the overall picture more easily when too many data points are added into the mix.

In a data visualization environment, it is also very important to not only have a tool that functions well, but also a user knows the subject matter well. If an expert were to analyze the data as visualized in the tool, they would undoubtedly observe a large number of phenomena and events that have not previously been uncovered.

11. REFERENCES

[1] Openintel active dns measurement project, 2017.

[2] M C F de Oliveira and H Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE transactions on visualization and computer graphics*, 9(3):378–394, 2003.

[3] J. Hawkinson and T. Bates. Guidelines for creation, selection, and registration of an Autonomous System (AS). pages 1–11, 1996.

[4] Bernardo A Huberman. Zipf ' s law and the Internet. pages 143–150, 2002.

[5] Mattijs Jonker, Anna Sperotto, Roland van Rijswijk-Deij, Ramin Sadre, and Aiko Pras. Measuring the adoption of ddos protection services. In *Proceedings of the 2016 ACM on Internet Measurement Conference*, IMC '16, pages 279–285, New York, NY, USA, 2016. ACM.

[6] Daniel Keim and Jim Thomas. Scope and Challenges of Visual Analytics. *IEEE Visualization Conference 2007*, pages 1–58, 2008.

[7] Pieter Lexis. Identifying Patterns in DNS Traffic.

[8] Roland Van Rijswijk-deij, Mattijs Jonker, Anna Sperotto, and Aiko Pras. The Internet of Names : A DNS Big Dataset Actively Measuring 50 % of the Entire DNS Name Space , Every Day. pages 91–92, 2015.

[9] Roland Van Rijswijk-deij, Mattijs Jonker, Anna Sperotto, and Aiko Pras. A High-Performance , Scalable Infrastructure for Large-Scale Active DNS Measurements. (c):1–11, 2016.

[10] Robert Sanders. The pareto principle: Its use and abuse. *Journal of Services Marketing*, 1(2):37–40, 02 1987.

[11] Paul Tol. Colour Schemes. (December), 2012.