Internet measurement data management challenges**

Amolishwar Pidulkar [16104009], Koushik Balasubramanium [16104044], Vikram Singh [16104100]

Abstract: By this report we try to critically analyze the major issues of internet measurement and data management. The researchers in the field are required with the data to test their models and verify their experiments for improved understanding of the structure and dynamics of Internet topology, routing, workload, performance. But this data is collected by the agencies for the user networks which is considered as a threat to their security. This lead to the use of some unrealitistic data which hinders the progress in the field of internet measurement and network security. As the internet users and data is increasing with time exponentially is stimulating to think about some effective data management techniques. The major topics of focus in the paper were sustainable collection, curation, and storage of large volumes of data; privacy-respecting sharing; and long-term archiving for reproducibility and we tried to find an extent up to which these issues has been solved and what are the new issues in the field. We collected the new data to see where this paper stands according to today's scenario. Internet is characterized at three topics relevant to research as well as operations of network infrastructure: IP address (interface), router, and Autonomous System (AS).

2. OVERVIEW OF CAIDA DATA

CAIDA: Centre for applied internet data analysis: CAIDA mainly collects data from both active and passive measurement.

Active includes macroscopic topology data using ARK infrastructure.

Passive includes data collected through collaboration with various organisations that monitor passively on selected links and may even anonymous IP address.

2.1 Active Measurement data: As of September 2016 ARK has 132 PC's located across 49 countries. The ARK(archipelago) is a globally distributed measurement platform whose primary goals are to Reduce effort for large scale measurement and to allow the contributors to run their measurement task in high security platform .The Data sets generated by ARK are as follows:

2.1.1 IPv4 traceroute data: In this data is collected by sending scampers probes to destination IP address selected from randomly from routed /24 .Each random address is probed every 48 hours. Many destinations are probed by different ARK monitors. As of now prefix list includes 10.13 million prefix.

2.1.2 DNS Lookups: After the collection of traceroute data CAIDA custom built DNS lookup service. This is done to gain additional knowledge on the topology of the routers and the hosts.

Instead of giving the query, response traffic, from 2014 onwards DNS traffic is available in two forms the most recent 30 days and quarterly data of past data.

2.1.3 Router level and AS level internet graph:(ITDK) Router level internet graphs from the BGP data is derived by using the technique of alias resolution. The AS- level information are done by mapping IP address to the AS's according to the presence of the hops in the trace. These are also done using publically available BGP data. These AS;s are then labelled according to their magnitude(large or small).

2.1.4 IPV6 traceroute data: Used for studying topology of AS and IP in IPv6.This is also done using scampers which perform ICMP based traceroutes. Till April 2014 36 ARK servers were probing 10269 IP v6 route prefixes.

406 million /48's prefixes were probed by a globally distributed set of 31 Ark monitors between 2014 and 15 to find the amount of subnetting existing in IPv6 prefixes announced in the BGP

2.2 PASSIVE MEASUREMENT TRAFFIC:

2.2.1 Core Link Data Traffic: Collecting and sharing traffic data is often limited by cost of equipment and up gradation. As of now there are four passive real time monitoring system,2 each in Chicago and San Jose

2.2.2 Unsolicited traffic: To observe the traffic sent to unassigned address space Network telescopes are employed .The UCSD Network telescope releases these anonymous traces. Major drawback is the telescope does not send any packet in response and hence this limits it capability of what it sees.

3. Sharing CAIDA data

Introduction: In this section we will discuss about various policies possessed by CAIDA, tools and other infrastructure that is provided and various fields of research enables and supported by CAIDA. In this section we will also try to see the various challenge t hat are faced by CAIDA and its involvement in the supporting the research in past few years.

3.1 Access policies

CAIDA mainly possesses two policies which are stated below:

a. Type I: This policy was majorly introduced to help and motivate the researchers by providing them with the data suitable for their research by understanding and respecting the kind of inputs required as in lots of cases the data is sensitive to the environment and tools used. As the appropriate and delicate data used under given conditions gives the most adequate results for the verification of the experiments and scientific models. In type I policy the tools and data are freely available but the user has to abide by the Acceptable Use Policy (AUP).

   Popular datasets: a). CodeRed Worm Dataset and AS relationship in 2010.

   b). Anonymized internet trace and UCSD Telescope in 2016

b. Type II: This type of dataset and tools are limited in access to academic researcher, US government agencies and CAIDA members. This type is majorly employed to protect the privacies of the users and data providers. The most popular data sets are shown above.
The statistics of utility of the data sets by the users is as shown in the figure 1

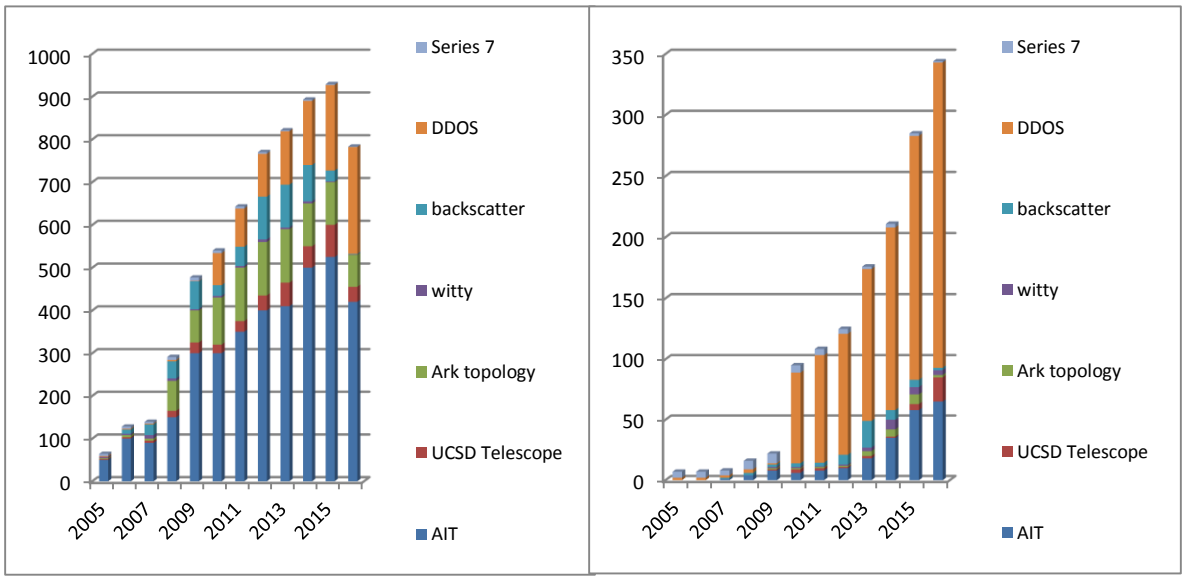

Figure 3.1 graph representing data/no of users requests with time[8]

3.2 User Support Infrastructure

User support infrastructure is built in order to help the people who are using the data or tools. Users can request CAIDA by various means famously by the emails and they are helped within 3 days at worst and 50% of the people are acknowledged within 1 day span. Various surveys and feedbacks are conducted periodically from the users to improve the system and usability.

3.3 Research Enabled by CAIDA data

In this section we would mention various communities which are helped by CAIDA and various new fields created or benefited directly or indirectly by CAIDA.

a) A lots of fields like Routing on overlay networks, BGP behavior, router level topology, improving anycast implementation etc. are required with the data of topology for pursuing their research is downloaded from CAIDA.

b) For classification and modeling of internet traffic, performance, modeling, monitoring and filtering techniques, intrusion detection can be conducted using backbone traffic data.

c) Denial-of-service attack and various internet worms can be tested and studied using UCSD Networks. It has also supported the researches on worst case delays, path analysation and speed of flash worms.

4.Challanges and open issues

The major issues in data intensive internet research are sustainable collection, curation, and storage of large volumes of data; privacy-respecting sharing; and long-term archiving for reproducibility. Most of the issues in this field are interrelated

4.1 Sustainable collection, curation & storage

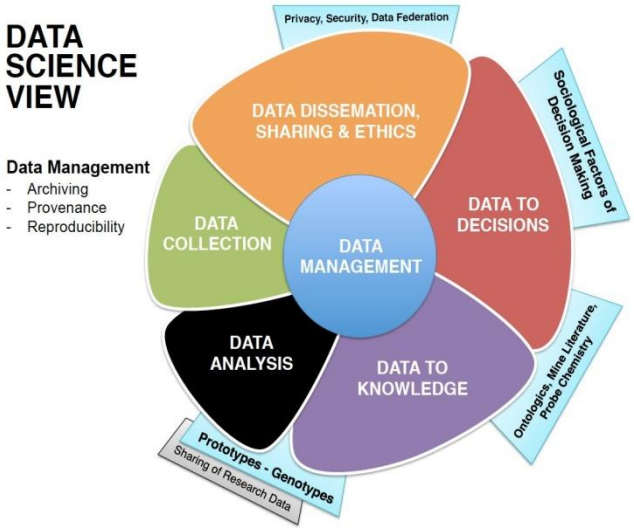

Figure 4.1 Flow of data, privacy in internet [7]

Now days the amount of data is increasing exponentially and it became very expensive to store them which can limit the number of users which access it. Such kind of efforts has a great tendency to reduce the research output so some algorithm has to be designed which can remove the data which is nearly useless as we knew that in research as a thing keep on becoming old researcher losses interest in it. Automated malicious softwares are a large threat as small things possesses a great damage causing ability. Large database are not only thing which can solve this problem because good processing speeds are also required so that data can be supplied without hampering the performance.

4.2 Privacy-sensitivity sharing Major issue is to develop a data sharing framework which takes care of privacy which incorporate the security and integrate the data without losing its utility which is a major objective in order to validate the research output of scientific society. Hence the major criteria in data sharing are:

4.2.1 How to ensure privacy

4.2.2 How to handle the dependence of these factors over each other.

4.2.3 How to achieve utility objective.

It's important to learn what actually meant by privacy and people need to understand research which doesn't directly involve humans.

4.3 Archive for reproducibility A new NSF policy was established according to which a committee was constituted. The job of the society is give the report and publish it immediately on the contribution of various people and major breakthroughs. This report must be authorized and guarantees that researcher finishes the work in time with proper utilization of resources.

4.4 Perspectives on the information stored in large and heterogeneous data sets This is relatively newer topic as the internet networks will certain newer challenges with the pioneering of fields like internet of things(IOT) [4] because nearly everything will be packet based and data access would become much more random then now. The other major breakthroughs in the field of internet and data management are HDOOP and BIGDATA.

Conclusions

The Active ARK infrastructure is increasing across many other countries in the days to come.
In the coming months ITDK is planning to add additional data that will come combine both router-level and AS-level internet topology
Data is learned over the networks hence it can be a threat to individual security. It's good that authorities and proper law enforcement prevents the unauthorized people to access the public network but on the other hand it makes it nearly impossible to provide academic researcher with data needed to scientifically study the internet. Our critical dependence on the internet has rapidly grown much stronger than our comprehension of structure, performance limits, dynamics, and evaluation, and unfortunately current law the had played the major boon in the progress of internet measurement and network security.

References

[1] Cooperative Association for Internet Data Analysis. CAIDA Data-Overview, April 2011. http://www.caida.org/data/overview/.

[2] Cooperative Association for Internet Data Analysis. Non-CAIDA Publications using CAIDA Data, June 2011. http://www.caida.org/data/publications/bydate/index.xml.

[3] Erin Kenneally and Kimberly Claffy. Dialing Privacy and Utility: A Proposed Data-sharing Framework to Advance Internet Research. IEEE Security and Privacy (S&P), July 2010. http://www.caida.org/publications/papers/2009/dialing privacy utility/.

[4] National Science Foundation. NSF Data Management Plan Requirements, 2010. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg 2.jsp#dmp.

[5] http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirA

[6] http://link.springer.com/article/10.1007/s11227-016-1677-z

[7] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[8] https://www.caida.org/data/about/downloads/tables.xml

[9] https://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/aag_6.jsp#VID4

**nothing in this paper has been copied from anywhere as per my knowledge except the figure2 from reference[6]