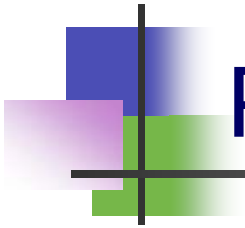
Algorithms and Tools for Anonymization of the Internet Traffic

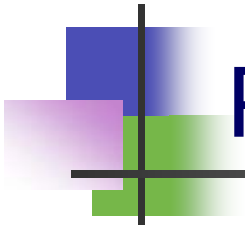
Tanjila Farah
tfarah@sfu.ca

Communication Networks Laboratory
School of Engineering Science
Simon Fraser University



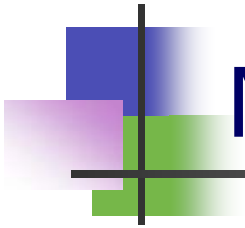
Roadmap

- Introduction
- Collection of network traffic
- Anonymization fields, algorithms, and tools
- Anonym tool
- Conclusion, future work, and references



Roadmap

- Introduction
- Collection of network traffic
- Anonymization fields, algorithms, and tools
- Anonym tool
- Conclusion, future work, and references



Motivation

- Internet is the easiest and the fastest medium for communication
- Measurement, characterization, and classification of Internet traces help network security
- Real-time network analysis relies on collection of trace logs
- Sharing traces may reveal the network architecture, user identity, and user information

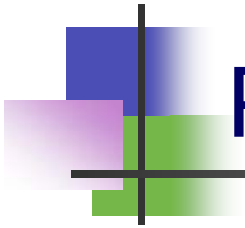
Anonymization

- Modifies network traces to protect user identity
- Removes the ability to identify the connection between two end-users
- Preserves the usefulness of the datasets
- Considers the type of analysis that may be performed
- Considers the requirements of the company sharing the datasets

Contributions

- Developed code in gawk to parse pcap and mrt input files
- Anonym tool:
  - developed the tool
  - developed the IPv6 address anonymization technique
  - implemented data analysis and visualization options
  - validated the tool performance

MRT : Multi-threaded Routing Toolkit
PCAP: Packet Capture
gawk: GNU AWK



Roadmap

- Introduction
- Collection of network traffic
- Anonymization fields, algorithms, and tools
- Anonym tool
- Conclusion, future work, and references

Collection of network traffic

- Internet is a collection of ASes exchanging information and delivering data
- Process of delivering data creates network traffic
- Network performance and QoS rely on network traffic characteristics
- Analyzing and understanding the network traffic helps ensure network security and QoS
- Network traffic collection helps:
  - traffic engineering
  - discovering the Internet topology
  - analyzing network security

AS : Autonomous System
QoS: Quality of Service

Role of traffic engineering

- Network troubleshooting:
  - deals with issues that disrupt or degrade the performance of a network: incorrect network address assignments and network anomalies
- Protocol debugging:
  - analyzes the existing and new protocols and performance of applications to determine required improvements
- Workload characterization:
  - examines the growth of network traffic volume due to new applications, protocols, and increasing number of users

Role of traffic engineering (cont.)

- Network performance evaluation:
  - estimates the network QoS by measuring traffic throughput and response time
- Capacity planning:
  - deals with network planning and managing by measuring bandwidth usage and availability

Discovering the Internet topology

- Discovering the Internet topology is important for:
  - simulating deployed networks
  - managing networks
  - mapping a network to determine location of the nearest servers and ISPs
  - designing and implementing new topology-aware protocols and algorithms

ISP: Internet Service Provider

Network security analysis

- Monitors policies adopted by network administrators to prevent the intruders from misusing the network
- It encompasses:
  - determining abnormal events: anomalies, attacks, and viruses
  - testing network firewalls
  - controlling access and network usage

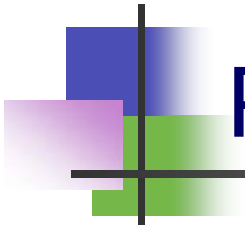
Network trace collection

- BCNET:
  - British Columbia's advance communication network
  - collected data are private and are only shared with the CNL
  - data are collected in the pcap format
- Cooperative Association for Internet Data Analysis (CAIDA):
  - collects, monitors, and visualizes various Internet data
  - collected data are public
  - data are collected in pcap and text formats

CNL: Communication Networks Laboratory

Network trace collection

- Route Views:
  - project at the University of Oregon
  - provides data and tools to the network administrators
  - collected data are public
  - data are collected in the mrt format
- Réseaux IP Européens (RIPE):
  - supports network operators in Europe, Middle East, Asia, and Africa
  - collected data are public
  - data are collected in the mrt format



Roadmap

- Introduction
- Collection of network traffic
- Anonymization fields, algorithms, and tools
- Anonym tool
- Conclusion, future work, and references

Anonymization fields

- Network traffic logs include data packet headers, which contain various fields:
  - time-stamp
  - IP addresses
  - MAC addresses
  - packet length
  - protocol

IP : Internet Protocol
MAC: Media Access Control

(2013) Summary of anonymization best practice techniques [Online]. Available: http://www.caida.org/projects/predict/anonymization/.

Anonymization algorithms

- Black marker:
  - deletes all the information or replaces the information by a fixed value

Time	IP	Length
0.0534	253.36.88.92	143

Time	IP	Length
0.0000	1.1.1.1	0

- Enumeration:
  - sorts the dataset, chooses a value higher then the first value, and adds the value to all data points

Length
143
60
1514

Length
203
120
1574

Anonymization algorithms (cont.)

- Precision degradation:
  - removes the most precise components of a data field

1.017851	1.017000
1.017852	1.017000
1.017915	1.017000

- Prefix-preserving:
  - if two IP addresses share the first n bits then their anonymized IP addresses will also share the first n bits

IP un-anonymized		IP anonymized	
112.116.186.8	115.23.40.51	235.251.46.4	240.48.153.85
112.116.186.8	115.23.40.51	235.251.46.4	240.48.153.85

Anonymization algorithms (cont.)

- Random shift:
  - shifts each data point by adding a random number

Packet length un-anonymized	Packet length anonymized
143	150
60	230
1514	1674

- Truncation:
  - deletes the n least significant bits from an IP or MAC address

MAC address	Anonymized MAC address
Cisco_e7:a1:c0 (00:1b:0d:e7:a1:c0)	Cisco_0:0:0 (00:1b:0d:0:0:0)
JuniperN_3e:ba:bd(78:19:f7:3e:ba:bd)	JuniperN_0:0:0(78:19:f7:0:0:0)

Anonymization algorithms (cont.)

- Reverse truncation:
  - deletes the n most significant bits from an IP or MAC address

MAC address	Anonymized MAC address
Cisco_e7:a1:c0 (00:1b:0d:e7:a1:c0)	Cisco_e7:a1:c0 (0:0:0:e7:a1:c0)
JuniperN_3e:ba:bd (78:19:f7:3e:ba:bd)	JuniperN_3e:ba:bd (0:0:0:3e:ba:bd)

Anonymization tools

- Cryptography based Prefix-preserving Anonymization: Crypto-PAn
- Anontool
- Framework for Log Anonymization and Information Management: FLAIM

Crypto-PAn

- Properties of Crypto-PAn:
  - one-to-one mapping
  - prefix-preserving anonymization
  - consistent across traces
  - cryptography-based

Input		Output	
Time	IP address	Time	IP address
0.000010	10.1.3.143	0.000010	117.14.240.136
0.000015	10.1.3.156	0.000015	117.14.240.85

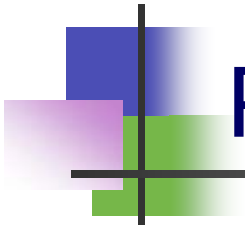
Anontool

- Anontool supports per-field anonymization
- Supports log files: pcap, netflow v5, and netflow v9
- Four-step anonymization process:
  - cooking function
    - assembles the flows according to protocols
  - filtering function
    - distinguishes the flows according to protocol and determine policy for anonymization
  - anonymization function
    - anonymizes the fields according to policy
  - un-cooking function
    - re-assembles the flows in the original format

FLAIM

- Supports an XML based policy
- Parsing modules are written based on the XML policy
- Supports log files: pcap, iptable, nfdump, and pacct
- FLAIM architecture consists of a module and a core:
  - the module provides policies to identify type of the log file
  - the core loads libraries responsible for anonymization

XML: Extensible Markup Language



Roadmap

- Introduction
- Collection of network traffic
- Anonymization fields, algorithms, and tools
- Anonym tool
- Conclusion, future work, and references

Anonym tool: functions

- Parses pcap and mrt files
- Anonymization options:
  - black marker, prefix-preserving, reverse-truncation, precision degradation, random shift, and truncation
- Data analysis options:
  - volume (bytes), volume (packets), volume curve fitting, throughput, empirical distribution, packet length distribution, protocol distribution, boxplot, and PDF and CDF curve fitting

Anonym tool: functions

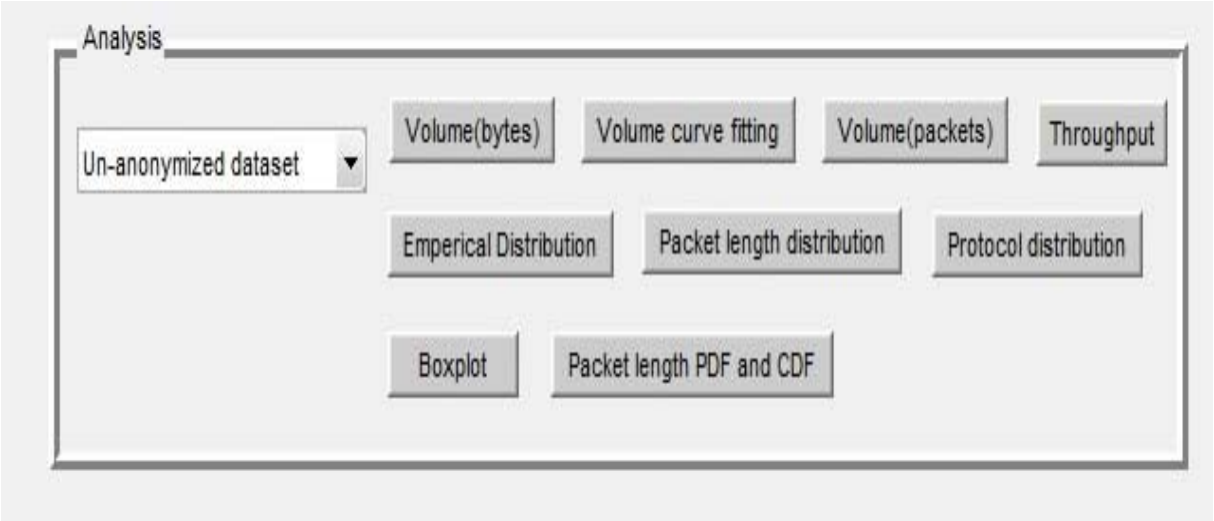
- Options for the K-S test:
  - determines if a dataset matches a tested distribution
  - provide options to test: normal, gamma, Weibull, exponential, Rayleigh, and lognormal distributions
- Additional options:
  - display anonymization results and analysis graphs
  - clear and upload new file
  - save figures and anonymization results

K-S: Kolmogorov-Smirnov

B. Vujicic, C. Hao, and Lj. Trajković, "Prediction of traffic in a public safety network," in Proc. IEEE International Symposium on Circuits and Systems (ISCAS' 06), Kos, Greece, May 2006, pp. 2637-2640.

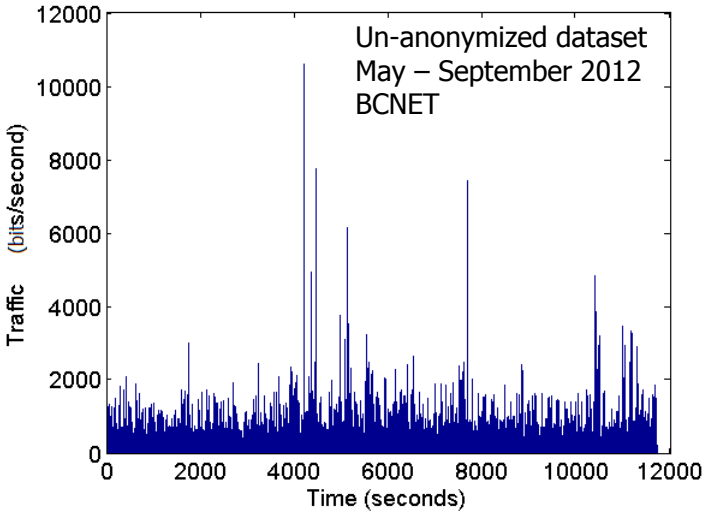
Data analysis options

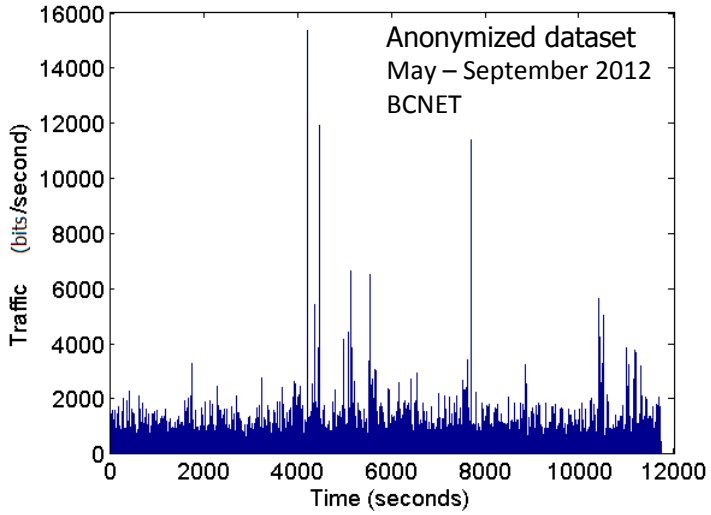
- We analyzed the effect of anonymization on the dataset by using the analysis options implemented in the Anonym tool



Data analysis option: volume

- Number of bits or packets per second
- Identifies the pattern of traffic flow through a network
- Shown are BGP, TCP, and UDP traffic volume:





TCP: Transmission Control Protocol
BGP: Border Gateway Protocol
UDP: User Datagram Protocol

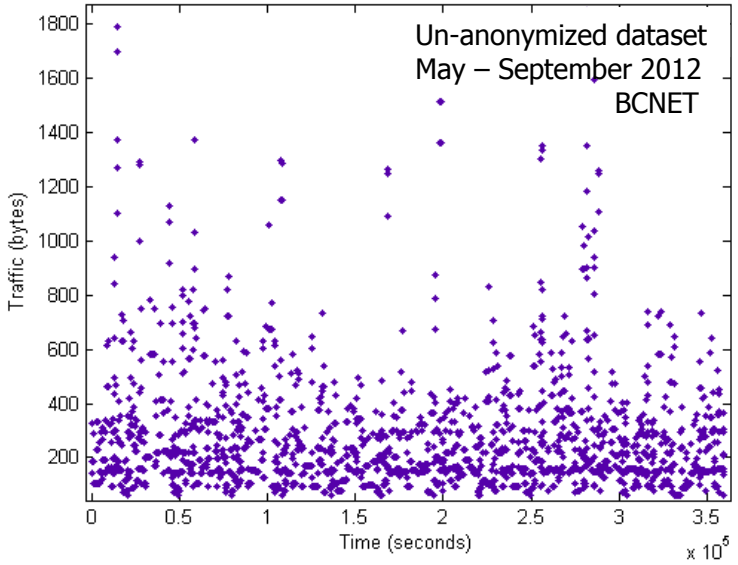
Data analysis option: volume

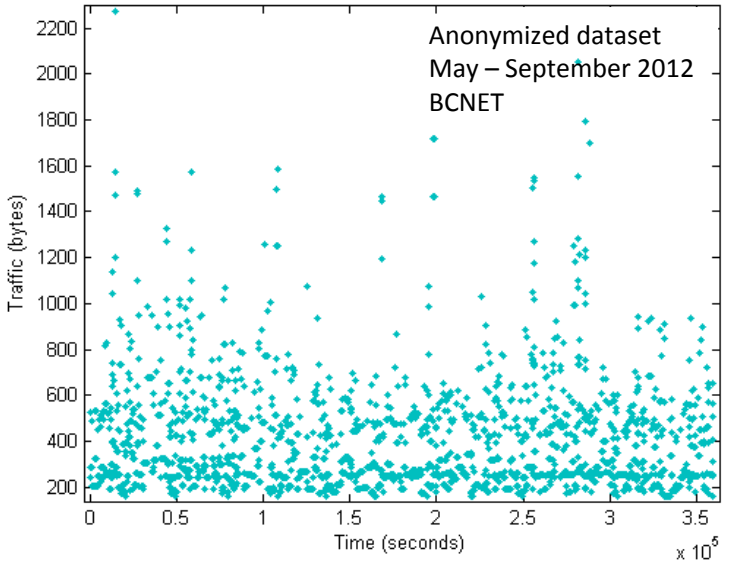
- Statistics of packet length:
  - an enumeration algorithm is applied to the dataset

Statistics	Un-anonymized dataset (bits)	Anonymized dataset (bits)
Minimum	60	160
Maximum	1,514	1,614
Mean	246.2475	346.2475
Median	157	257
Standard deviation	259.4509	259.4509

Data analysis option: volume curve fitting

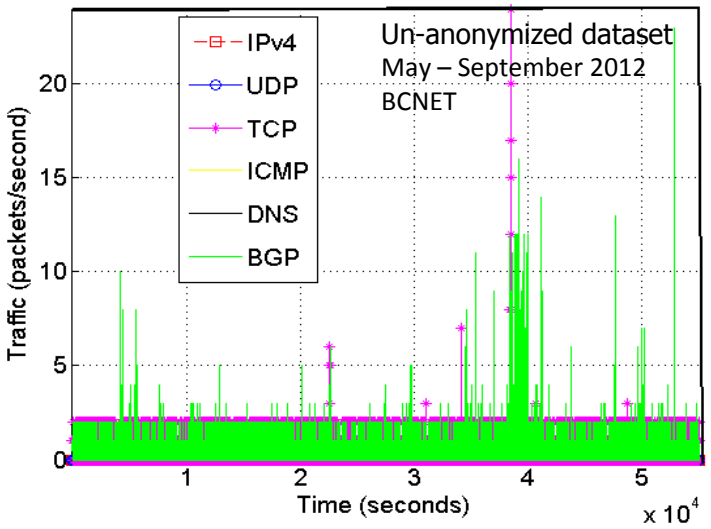
- Run-sequence: displays a graphical representation of a dataset
- Fitting curves to a dataset: Fourier, Gaussian, Weibull, exponential, polynomial, and sum of sine distributions

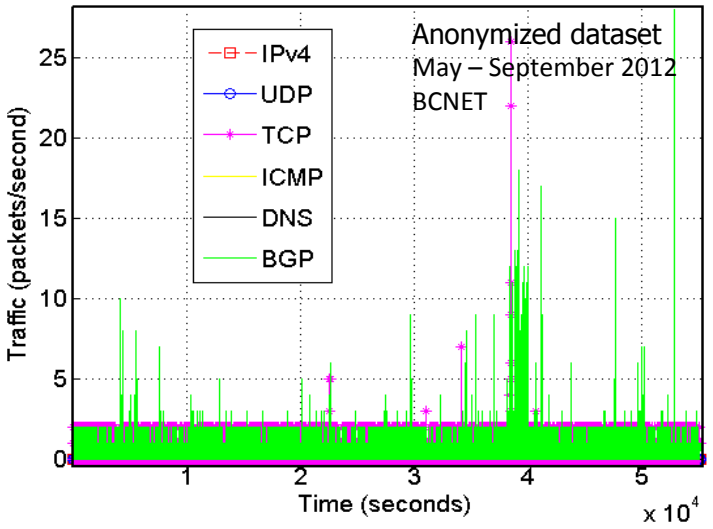




Data analysis option: protocol distribution

- Provides an overview of various protocols occupancy in the network
- Classifies IP, UDP, TCP, ICMP, DNS, and BGP traffic

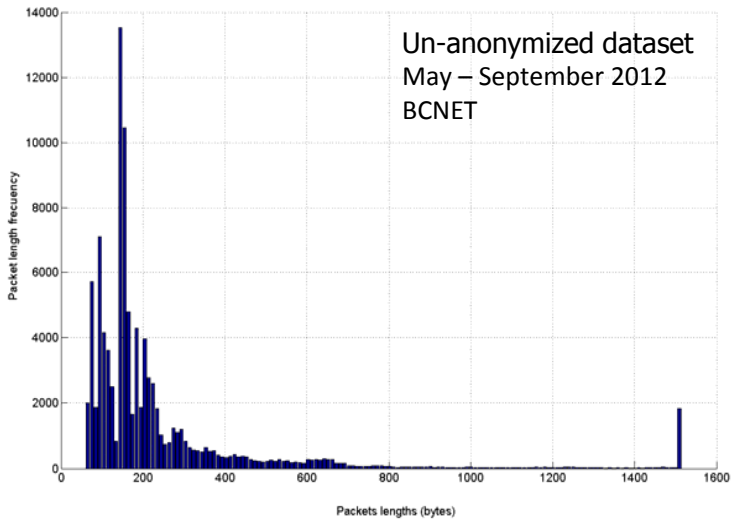


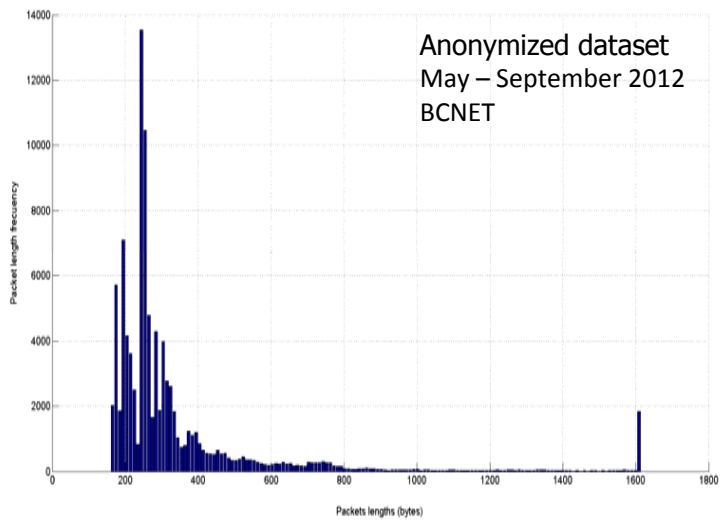


ICMP: Internet Control Message Protocol
DNS : Domain Name Service

Data analysis option: packet length distribution

- Displays the histogram plot of a dataset
- Indicates appropriate distribution model of a dataset
- Significant percentage of packets are 150 bytes for un-anonymized dataset and 250 bytes for anonymized dataset as shown:

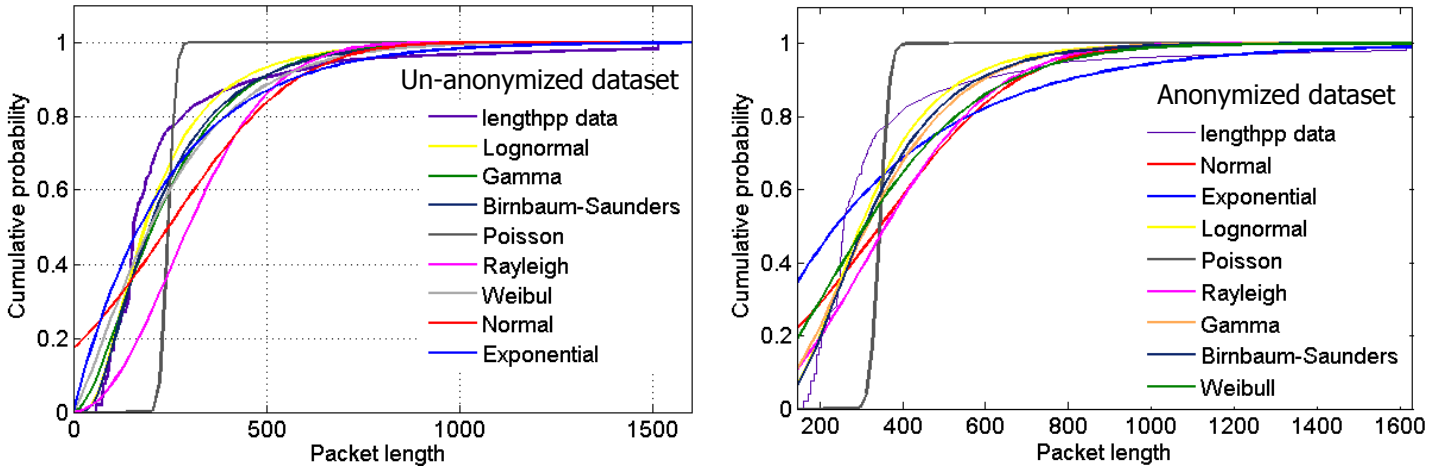




M. Fras, J. Mohorko, and Z. Cucej, "A new goodness of fit test for histograms regarding network traffic packet size process," in *Proc. International Conference on Advanced Technologies for Communications (ATC' 2008)*, Hanoi, Vietnam, Oct. 2008, pp.345-348.

Data analysis option: fitting PDFs and CDFs

- PDF and CDF indicate the probability that the structure of a dataset follow certain distribution
- Provides options to fit thirteen distributions to PDF and CDF distribution curves of a dataset

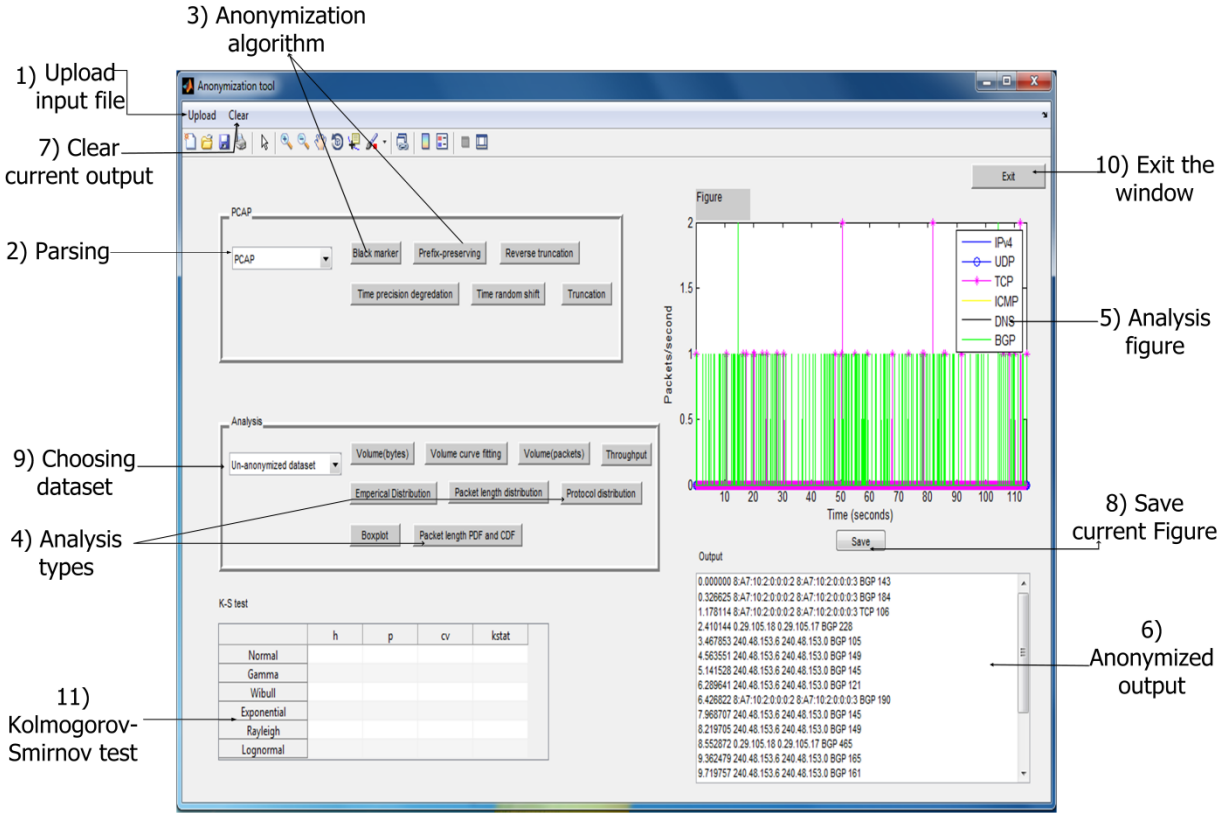


M. Fras, J. Mohorko, and Z. Cucej, "Packet size process modeling of measured self-similar network traffic with defragmentation method," in *Proc. 15th International Conference on Systems, Signals and Image Processing (IWSSIP' 08)*, Bratislava, Slovakia, June 2008, pp. 253-256.

PDF: Probability Density Function
CDF: Cumulative Distribution Function

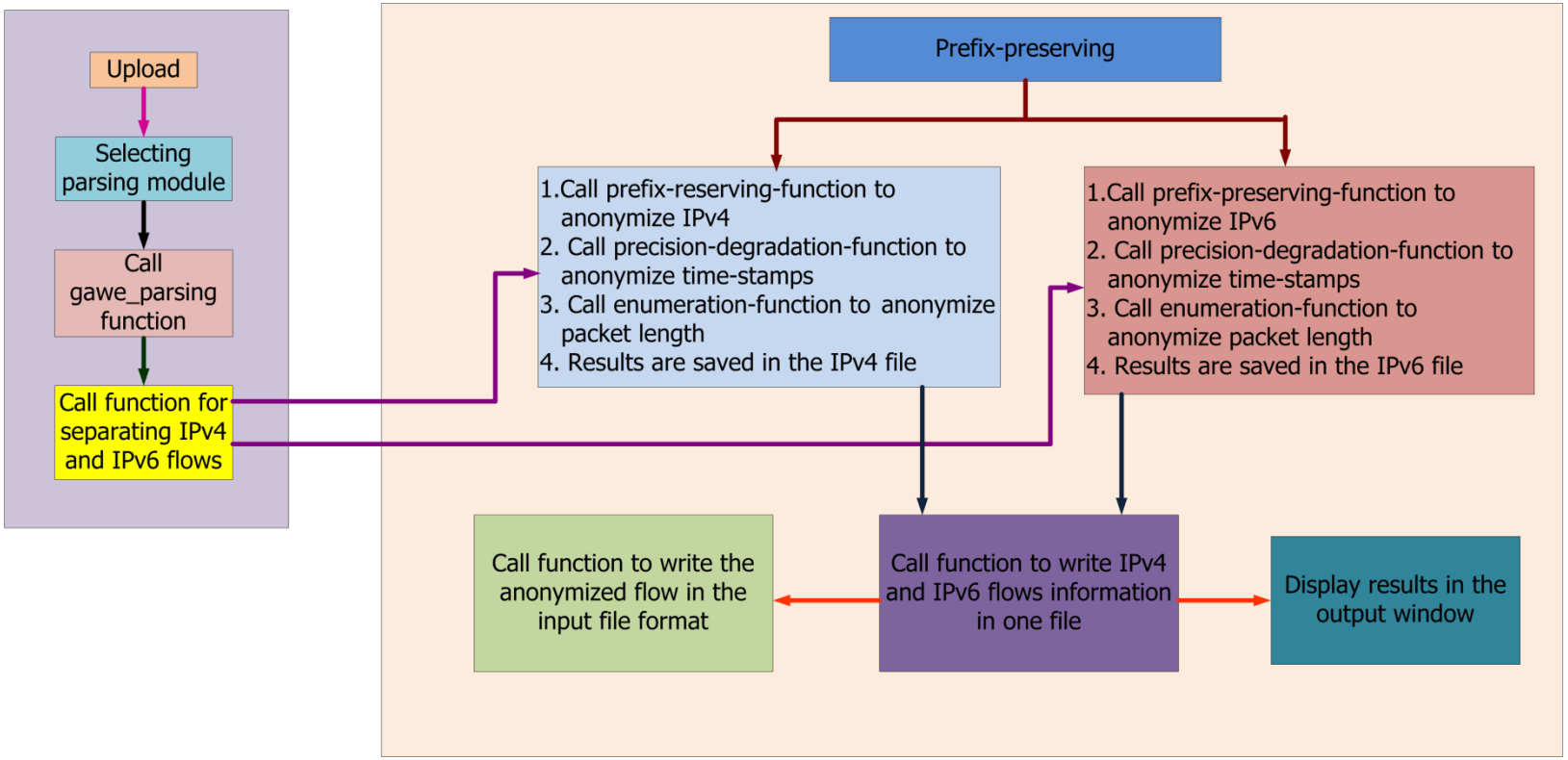
Anonym tool: GUI

- Graphical user interface



Operational diagram

- Prefix-preserving option



Functions: code

Call function for separating IPv4 and IPv6 flows

```
iporder=[];
f=fopen(name);
f4=fopen(name4,'w');
f6=fopen(name6,'w');
while 1 % For each line

    line = fgetl(f);
    if (strfind(line, ':'))
        iporder=[iporder 1];
        fprintf(f6, '%s\n',line);
    else
        iporder=[iporder 0];
        fprintf(f4, '%s\n',line);
    end
    if (feof(f))
        break;
    end
end
fclose(f);
fclose(f4);
fclose(f6);
[time4,IPv4source,IPv4destination,protocol4,
ipv4pktlength] = ipv4decode(name4);
[time6,IPv6s,IPv6d,protocol6,pktlength]
 = ipv6decode(name6);
```

Call prefix-reserving-function to anonymize IPv4

```
v4source= 'v4SColumn.txt';
% Output of decoded IPv4 destination column
v4destination='v4DColumn.txt';
% This gives the size of the input file
inputsize= size(IPv4source,1);
% Time in defind zero because at this point we are not decoding the
time data. This need to be fixed.
% time=0;
% Writing the output of IPv4 decode in a files v4SColumn.txt and
v4DColumn.txt.
% Source and destination is writen in two files because Crypto-PAN
takes
% input as it this formate (time length, address)
ipv4writefile (v4source,v4destination, time4,
ipv4pktlength,IPv4source,IPv4destination,inputsize);
v4Sanonymized='v4sourceanonymized.txt';  % IPv4 source address
anonymized
v4Danonymized='v4destinationanonymized.txt';   % IPv4
sdestination address anonymized
[s,s]=dos(['cryto_run.exe ' v4source ' > ' v4Sanonymized]);
[s,s]=dos(['cryto_run.exe ' v4destination ' > ' v4Danonymized]);
```

Functions: code

Call prefix-preserving-function to anonymize IPv6

```
IPv4s=IPv6s(:,1:4);
IPv4d=IPv6d(:,1:4);
lines = size(IPv6s,1);
namev4s='10outv4s.txt';
namev4d='10outv4d.txt';
time=0;
ipv6toipv4writefile(namev4s,namev4d, time6, pktlength, IPv4s,
IPv4d,lines );
namev4sout='10outv4sout.txt';
namev4dout='10outv4dout.txt';
[s,s]=dos(['cryto_run.exe ' namev4s ' > ' namev4sout]);
[s,s]=dos(['cryto_run.exe ' namev4d ' > ' namev4dout]);
[IPv4sout,IPv4dout, timeout,
pktlengthout]=ipv4toipv6readfile(namev4sout, namev4dout, lines );
Anonymized IPv6 address output
nameano='10outv6ano.txt';
writeipv6anon(time6,IPv6s,IPv6d,IPv4sout,IPv4dout,protocol6,pktle
ngthout,nameano);
```

Call precision-degradation-function to anonymize IPv4 flow time-stamps

```
[IPv4source,IPv4destination, time4,
ipv4pktlength]=ipv4readfile(v4Sanonymized, v4Danonymized,
inputsize );
 precision degradation timeanony4=fix(time4*100)/100;
nameano='10outtimev4ano.txt';
writeipv4anon(timeanony4,IPv4source,IPv4destination,protocol4,ipv
4pktlength,nameano)
[IPv4sout,IPv4dout, timeout,
pktlengthout]=ipv4toipv6readfile(namev4sout, namev4dout, lines);
precision degradation
```

Call precision-degradation-function to anonymize IPv6 flow time-stamps

```
[IPv4sout,IPv4dout, timeout,
pktlengthout]=ipv4toipv6readfile(namev4sout, namev4dout, lines);
timeanony6=fix(time6*100)/100;
Anonymized IPv6 address output
nameano='10outv6ano.txt';
writeipv6anon(timeanony6,IPv6s,IPv6d,IPv4sout,IPv4dout,protocol6
,pktlengthout,nameano);
```

Validation tests

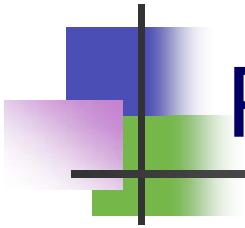
- Implementation of the Anonym tool was validated using various tests:

Fields	Anonym	Anontool	FLAIM
Source 64.251.87.209	0.29.105.18	110.13.240.136	103.51.250.0
Destination 64.251.87.210	0.29.105.17	110.13.246.137	103.51.250.28

The Anonym tool: results

- Per-field anonymizatoin results:

Time-stamp	IPv4 and IPv6		Packet length
	Un-anonymized dataset		
0.000000	2001:4958:10:2::2	2001:4958:10:2::3	143
1.178114	2001:4958:10:2::2	2001:4958:10:2::3	106
2.410144	64.251.87.209	64.251.87.210	228
4.563551	206.47.102.206	206.47.102.201	149
	Anonymized dataset		
0.000000	8:A7:10:2:0:0:0:2	8:A7:10:2:0:0:0:3	243
1.170000	8:A7:10:2:0:0:0:2	8:A7:10:2:0:0:0:3	206
2.410000	0.29.105.18	0.29.105.17	328
4.560000	240.48.153.6	240.48.153.0	249



Roadmap

- Introduction
- Collection of network traffic
- Anonymization fields, algorithms, and tools
- Anonym tool
- Conclusion, future work, and references

Conclusions

- The Anonym tool provides options to anonymize time, IPv4 and IPv6 addresses, MAC addresses, and packet length data
- Supports log files in mrt and pcap formats
- Provides options to analyze the datasets
- Provides options to apply the K-S test on the datasets
- Analysis of un-anonymized and anonymized datasets indicates insignificant variations

Future work

- The Anonym tool may be enhanced to support other log formats: netflow, iptable, and pccat
- Additional anonymization algorithms may be implemented: binning, hash, partitioning, permutation, and random noise addition
- Anonymization of additional fields may be implemented: port numbers, TCP window size, and IP ID number

References

1. P. Porras and V. Shmatikov, "Large-scale collection and sanitization of network security data: risks and challenges," in *Proc. Workshop on New Security Paradigms (NSPW '06)*, Germany, Oct. 2007, pp. 57-64.
2. J. Xu, J. Fan, M. Ammar, and S. B. Moon, "On the design and performance of prefix-preserving IP traffic trace anonymization," in *Proc. 1st ACM SIGCOMM Workshop on Internet Measurement (IMW' 01)*, San Francisco, CA, USA, Nov. 2001, pp. 263-266.
3. A. Slagell, J.Wang, and W. Yurcik, "Network log anonymization: application of Crypto-PAn to Cisco netflows," in *Proc. NSF/AFRL Workshop on Secure Knowledge Management (SKM '04)*, Buffalo, NY, USA, Sept. 2004, pp. 223-228.
4. M. Foukarakis, D. Antoniades, S. Antonatos, and E. P. Markatos, "Flexible and high-performance anonymization of netflow records using anontool," in *Proc. Third International Workshop on the Value of Security through Collaboration (SECOVAL' 07)*, Nice, France, Sept. 2007, pp. 33-38.
5. D. Koukis, S. Antonatos, D. Antoniades, E. Markatos, and P. Trimintzios, "A generic anonymization framework for network traffic," in *Proc. IEEE International Conference on Communications (ICC' 06)*, Istanbul, Turkey, June 2006, vol. 5, pp. 2302-2309.
6. A. Slagell, K. Lakkaraju, and K. Luo, "FLAIM: a multi-level anonymization framework for computer and network logs," in *Proc. 20th conference on Large Installation System Administration (LISA' 06)*, Washington, DC, USA, July 2006, pp. 101-115.
7. M. Bishop, B. Bhumiratana, R. Crawford, and K. Levitt, "How to sanitize data," in *Proc. 13th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE '04)*, Washington, DC, USA, June 2004, pp. 217-222.
8. R. Crawford, M. Bishop, B. Bhumiratana, L. Clark, and K. Levitt, "Sanitization models and their limitations," in *Proc. Workshop on New Security Paradigms (NSPW '06)*, Germany, Mar. 2006, pp. 41-56.
9. R. Pang, M. , Allman, V. Paxson, and J. Lee, "The devil and packet trace anonymization," *ACM SIGCOMM*, Aug. 2006, vol. 36, pp. 29-38.
10. PRISM state of the art on data protection algorithms for monitoring systems. [Online]. Available: http://fp7-prism.eu/images/upload/Deliverables/fp7-prism-wp3.1-d3.1.1-nal.pdf.

References

11. T. Farah, S. Lally, R. Gill, N. Al-Rousan, R. Paul, D. Xu, and Lj. Trajković, "Collection of BCNET BGP traffic," in *Proc. 23rd International Teletraffic Congress (ITC' 11)*, San Francisco, CA, Sept. 2011, pp. 322-323.
12. S. Lally, T. Farah, R. Gill, R. Paul, N. Al-Rousan, and Lj. Trajković, "Collection and characterization of BCNET BGP traffic," in Proc. *2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM' 11)*, Victoria, BC, Canada, Aug. 2011, pp. 830-835.
13. S. Lau and Lj. Trajković, "Analysis of traffic data from a hybrid satellite-terrestrial network," in *Proc. Fourth Int. Conf. on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine' 2007)*, Vancouver, BC, Canada, Aug. 2007, pp. 9:1-9:7.
14. B. Vujicic, C. Hao, and Lj. Trajković, "Prediction of traffic in a public safety network," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS' 06)*, Kos, Greece, May 2006, pp. 2637-2640.
15. N. Al-Rousan, S. Haeri, and Lj. Trajkovic, "Feature selection for classification of BGP anomalies using Bayesian models," in *Proc. ICMLC 2012*, Xi'an, China, July 2012, pp. 140-147.
16. Lj. Trajković, "Analysis of Internet topologies," *Circuits and Systems Magazine*, Sept. 2010, vol. 10, no. 3, pp. 48-54.
17. M. Najiminaini, L. Subedi, and Lj. Trajković, "Analysis of Internet topologies: a historical view," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS' 09)*, Taipei, Taiwan, May 2009, pp. 1697-1700.
18. J. Chen and Lj. Trajković, "Analysis of Internet topology data," in *Proc. International Symposium on Circuits and Systems (ISCAS' 04)*, Vancouver, British Columbia, Canada, May 2004, vol. 4, pp. 629-632.

19. (2013) BCNET. [Online]. Available: https://wiki.bc.net.
20. (2013) University of Oregon Route Views Project. [Online]. Available: http://www.routeviews.org.
21. (2013) RIPE (Reseaux IP Europeens). [Online]. Available: http://www.ripe.net.
22. (2013) The Corporative Association for Internet Data Analysis. [Online]. Available: http://www.caida.org/data.

Acknowledgements

- Prof. Veselin Jungic, Chair
- Prof. Parvaneh Saeedi, Supervisor
- Prof. Emeritus Stephen Hardy, Examiner
- Prof. Ljiljana Trajković, Senior Supervisor

Acknowledgements

- Toby Wong, BCNET
- CNL mates
- Eva María Cavero Racaj, Universidad de Zaragoza

Thank You!