Classifying Internet One-way Traffic

Eduard Glatz,
Xenofontas Dimitropoulos

ETH Zurich

May 15, 2012

Overview

- Classification scheme for dissecting one-way traffic that relies solely on flow-level data
- Observation on one-way traffic based on a massive dataset of 457 billion flows
- Show how one-way flows are useful for service availability monitoring

Preliminaries

- Study **incoming** one-way traffic at the network level: connections that do not receive a reply.
- Example causes of one-way traffic:
  - Failures & Policies
  - Attacks
  - Special application behavior

Preliminaries

- Study **incoming** one-way traffic at the network level: connections that do not receive a reply.
- Example causes of one-way traffic:
  - Failures & Policies
  - Attacks
  - Special application behavior
- Sampling and asymmetric routing can result in artificial one-way traffic
- One-way traffic can be measured in edge networks

Classification Scheme

- Associate each one-way flow with a number of **signs**
- Introduce 18 signs exploiting in 4 cases techniques from the literature
- Classify flows based on their signs
- Classes:
  - Unreachable services
  - P2P applications
  - Scanning
  - Backscatter
  - Suspected Benign
  - Bogon

Signs: Host pair behavior

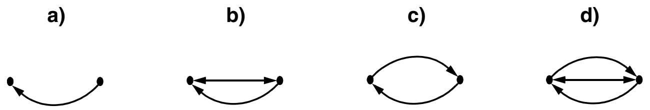


Figure: Mixture of incoming one- and two-way flows exchanged between a host pair. Hosts are represented by nodes and the presence of inflow/outflow/biflows by arrows.

Signs: Host pair behavior

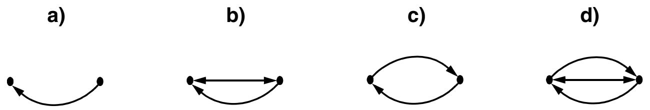


Figure: Mixture of incoming one- and two-way flows exchanged between a host pair. Hosts are represented by nodes and the presence of inflow/outflow/biflows by arrows.

- End-hosts-communicating: One-way flow between productive host pair
- Limited dialog: One-way flows between unproductive host pair

Signs: Local host behavior

- Unused local address: Unpopulated local IP address
- Service unreachable: Unanswered request to local service
- Peer-to-peer[1]: Flow towards local P2P host

[1] W. John and S. Tafvelin. Heuristics to classify internet backbone traffic based on connection patterns. International Conference on Information Networking (ICOIN), 2008

Signs: Remote host behavior

- Service sole reply: no biflow on srcIP $\wedge$ dstPort$\geq$1024 $\wedge$ srcPort $<$ 1024
- Remote scanner 1[2]: TRW algorithm (suspected scanner)
- Remote scanner 2[3]: Host classification (suspected scanner)
- Remote non-scanner: TRW algorithm (suspected regular host)

[2] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan. Fast portscan detection using sequential hypothesis testing. In Proceedings of the IEEE Symposium on Security and Privacy, 2004

[3] M. Allman, V. Paxson, and J. Terrell. A brief history of scanning. In Proceedings of the 7th ACM SIGCOMM IMC, 2007

Signs: Flow feature

- Artifact: UDP/TCP flow with both port numbers=0
- Single packet: Flow contains one packet only
- Large flow: Flow carries $\geq$ 10 packets or $\geq$ 10240 bytes
- Bogon: Source IP belongs to bogon space
- Protocol: IP protocol type of flow

Classification Rules

Final classifier includes 17 classification rules

Class Name	Rule #	Flow Membership Rules
Malicious Scanning	1	$\{TRWscan, \overline{HCscan}, \overline{PotOk}\} \Rightarrow Scanner$
	2	$\{HCscan, \overline{TRWscan}, \overline{TRWnom}, \overline{PotOk}\} \Rightarrow Scanner$
	3	$\{TRWscan, HCscan, \overline{PotOk}\} \Rightarrow Scanner$
	4	$\{TRWnom, HCscan\} \Rightarrow Scanner$
	5	$\{GreyIP, Onepkt, \overline{TRWscan}, \overline{HCscan}, \overline{Backsc}, \overline{ICMP}, \overline{UDP}, \overline{bogon}\} \Rightarrow Scanner$
	6	$\{GreyIP, \overline{TRWscan}, \overline{HCscan}, \overline{Onepkt}, \overline{ICMP}, \overline{Backsc}, \overline{bogon}\} \Rightarrow Scanner$
	7	$\{Onepkt, \overline{GreyIP}, \overline{ICMP}, \overline{TRWscan}, \overline{HCscan}, \overline{TRWnom}, \overline{bogon}, \overline{P2P}, \overline{Unreach}, \overline{PotOk}, \overline{Backsc}, \overline{Large}\} \Rightarrow Scanner$
	8	$\{GreyIP, Onepkt, \overline{TRWscan}, \overline{HCscan}, \overline{Backsc}, \overline{ICMP}, \overline{TCP}, \overline{bogon}\} \Rightarrow Scanner$
	9	$\{ICMP, \overline{TRWscan}, \overline{TRWnom}, \overline{HCscan}, \overline{InOut}, \overline{bogon}, \overline{PotOk}\} \Rightarrow Scanner$
Backscatter	10	$\{Backsc, \overline{TRWscan}, \overline{HCscan}, \overline{P2P}, \overline{InOut}, \overline{PotOk}\} \Rightarrow Backscatter$
Service Unreachable	11	$\{Unreach, \overline{TRWscan}, \overline{HCscan}, \overline{bogon}, \overline{P2P}\} \Rightarrow Unreachable$
Benign P2P Scanning	12	$\{P2P, \overline{TRWscan}, \overline{HCscan}, \overline{bogon}\} \Rightarrow P2P$
Suspected Benign	13	$\{PotOk, \overline{Unreach}, \overline{P2P}, \overline{TRWnom}, \overline{bogon}\} \Rightarrow Benign$
	14	$\{Large, \overline{GreyIP}, \overline{TRWscan}, \overline{HCscan}, \overline{P2P}, \overline{Unreach}, \overline{PotOk}, \overline{ICMP}, \overline{Backsc}, \overline{bogon}, \overline{TRWnom}\} \Rightarrow Benign$
	15	$\{TRWnom, \overline{GreyIP}, \overline{HCscan}, \overline{P2P}, \overline{Unreach}, \overline{bogon}, \overline{Backsc}\} \Rightarrow Benign$
	16	$\{ICMP, InOut, \overline{TRWscan}, \overline{HCscan}, \overline{TRWnom}, \overline{bogon}, \overline{PotOk}\} \Rightarrow Benign$
Bogon	17	$\{bogon, \overline{TRWscan}, \overline{HCscan}, \overline{Backsc}\} \Rightarrow Bogon$

Data-Sets

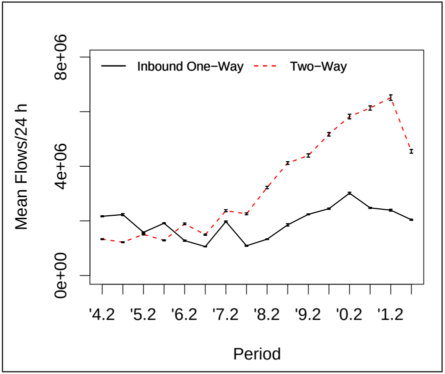
- Use data from the Swiss academic backbone network (SWITCH)
- Analyze the first 400 hours of each Feb and Aug between 2004 and 2011
- The studied traffic data correspond to:
  - 457 billion flows
  - 7.41 petabytes
  - cover 9% of the total number of flows

Data Sanitization

- Double-counting elimination reduces total traffic volume by 32.3%
- Defragmentation reduces the number of flows by a fraction ranging between 20.6% and 39.6% for different years
- Bi-flow Pairing:
  - For TCP and UDP based on standard 5-tuple
  - For other protocols based on 3-tuple

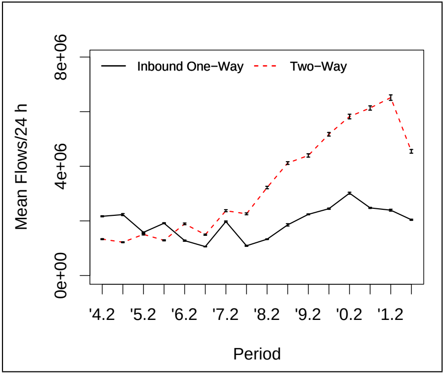
Evolution of One- and Two-way Traffic

- One-way flows are a large fraction of all flows:
  - In 2004, 2 out of every 3 flows were one-way
  - From 2007 to 2010, 1 out of every 3 flows were one-way



Evolution of One- and Two-way Traffic

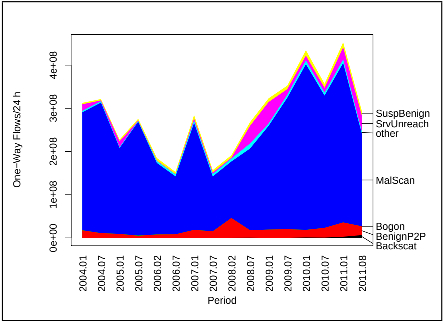
- One-way flows are a large fraction of all flows:
  - In 2004, 2 out of every 3 flows were one-way
  - From 2007 to 2010, 1 out of every 3 flows were one-way
- The number of one-way flows in 2011 is almost equal to 2004
- The fraction of one-way flows has declined



Composition of One-way Traffic

Class	% of flows	% of pkts	pkts/flow
Scanning	83.5%	62.6%	1.6
P2P applications	6.7%	13.0%	6.8
Unreach services	4.8%	10.1%	4.1
Suspected Benign	2.6%	9.1%	12.1
Other	2.2%	4.7%	4.6
Backscatter	0.3%	0.5%	3.3

- The top sources of one-way traffic are scanning, P2P protocols, and unreachable services



Service Availability Monitoring

- One-way flows are very useful for service availability monitoring
- Traditional service availability monitoring is based on active probing

Service Availability Monitoring

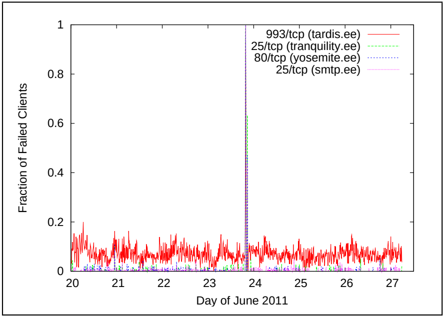
- One-way flows are very useful for service availability monitoring
- Traditional service availability monitoring is based on active probing
- Advantages of flow-based approach:
  - Provides a tangible assessment of the impact of disruptions
  - Discovers running services without requiring manual configuration
  - Exploits passive measurements

Outages and Misconfigurations in ETH Zurich

- Examine a week of NetFlow data from the EE Department of ETH Zurch
- Found 32 main services ($>$ 99% availability) and 11 transient services

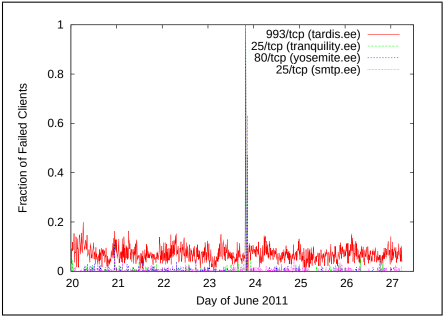
Outages and Misconfigurations in ETH Zurich

- Examine a week of NetFlow data from the EE Department of ETH Zurch
- Found 32 main services ($>$ 99% availability) and 11 transient services
- Identified a coinciding global outage



Outages and Misconfigurations in ETH Zurich

- Examine a week of NetFlow data from the EE Department of ETH Zurch
- Found 32 main services ($>$ 99% availability) and 11 transient services
- Identified a coinciding global outage
- During the identified interval **287,583 unique IP addresses failed** to access target services!



Conclusions

- Classification scheme for one-way traffic that relies on 18 signs derived from flow data
- Observations based on a very large data-set:
  - One-way flows are a large fraction of all flows
  - In terms of flows, the share of one-way traffic has declined since 2004
  - The top sources of one-way traffic are scanning, P2P protocols, and unreachable services
- One-way traffic is very useful for assessing the impact of failures

Questions?

Contact: fontas@gmail.com

E. Glatz and X. Dimitropoulos. Classifying Internet One-way Traffic. TIK-Report 336, ETH Zurich, May 2012

Validation

- Collect packet traces from a small campus network
- Exploit additional information:
  - Extended host profiles
  - ICMP types and codes
  - TCP flags (Check protocol state machine)
  - DPI-based application identification[4]
  - Precise timestamps

[4] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee. Internet traffic classification demystified: myths, caveats, and the best practices. ACM CoNEXT, 2008

Validation

- Collect packet traces from a small campus network
- Exploit additional information:
  - Extended host profiles
  - ICMP types and codes
  - TCP flags (Check protocol state machine)
  - DPI-based application identification[4]
  - Precise timestamps

Class Name	Recall [%]	Precision [%]
Malicious Scanning	99.9	99.8
Service Unreachable	99.6	96.1
Benign P2P Scanning	95.3	95.5
Backscatter	62.4	88.4
Suspected Benign	85.1	75.0
Bogon	40.4	100.0

[4] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee. Internet traffic classification demystified: myths, caveats, and the best practices. ACM CoNEXT, 2008

Outages and Misconfigurations in ETH Zurich

- Found server that was not reachable during the studied week in total by 2.2 million unique clients!
- What was this server? Hint: Switzerland is famous for chocolate, banking, swiss army knifes, and watches

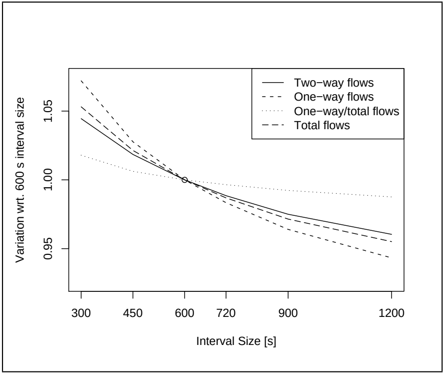
Outages and Misconfigurations in ETH Zurich

- Found server that was not reachable during the studied week in total by 2.2 million unique clients!
- What was this server? Hint: Switzerland is famous for chocolate, banking, swiss army knifes, and watches
- Popular NTP server `swisstime.ee.ethz.ch` preconfigured in NTP clients and used in NTP "hello world" examples
- It was not reachable to 12.9% of its clients cause by invalid CRC checksums and a filtering policy

Impact of the Interval Size

Doubling the interval size:

- decreases absolute count metrics by 3-5%.
- decreases relative volume metrics by 1.2% and does not
- decrease further with an increasing interval size.



Signs

Sign Type	Sign Name	Detection Criterion/Algorithm
Host pair behavior	End-hosts-communicating	One-way flow between productive host pair
	Limited dialog	One-way flows between unproductive host pair
Remote host behavior	Service sole reply	no biflow on srcIP $\wedge$ dstPort$\geq$1024 $\wedge$ srcPort $<$ 1024
	Remote scanner 1	TRW algorithm (suspected scanner)
	Remote scanner 2	Host classification (suspected scanner)
	Remote non-scanner	TRW algorithm (suspected regular host)
Local host behavior	Unused local address	Unpopulated local IP address
	Service unreachable	Unanswered request to local service
	Peer-to-peer	Flow towards local P2P host
Flow feature	Artifact	UDP/TCP flow with both port numbers=0
	Single packet	Flow contains one packet only
	Large flow	Flow carries $\geq$ 10 packets or $\geq$ 10240 bytes
	Bogon	Source IP belongs to bogon space
	Protocol	IP protocol type of flow