University of Twente
Enschede - The Netherlands

EMANICS

Self-Management of Hybrid Networks: Can We Trust NetFlow Data?*

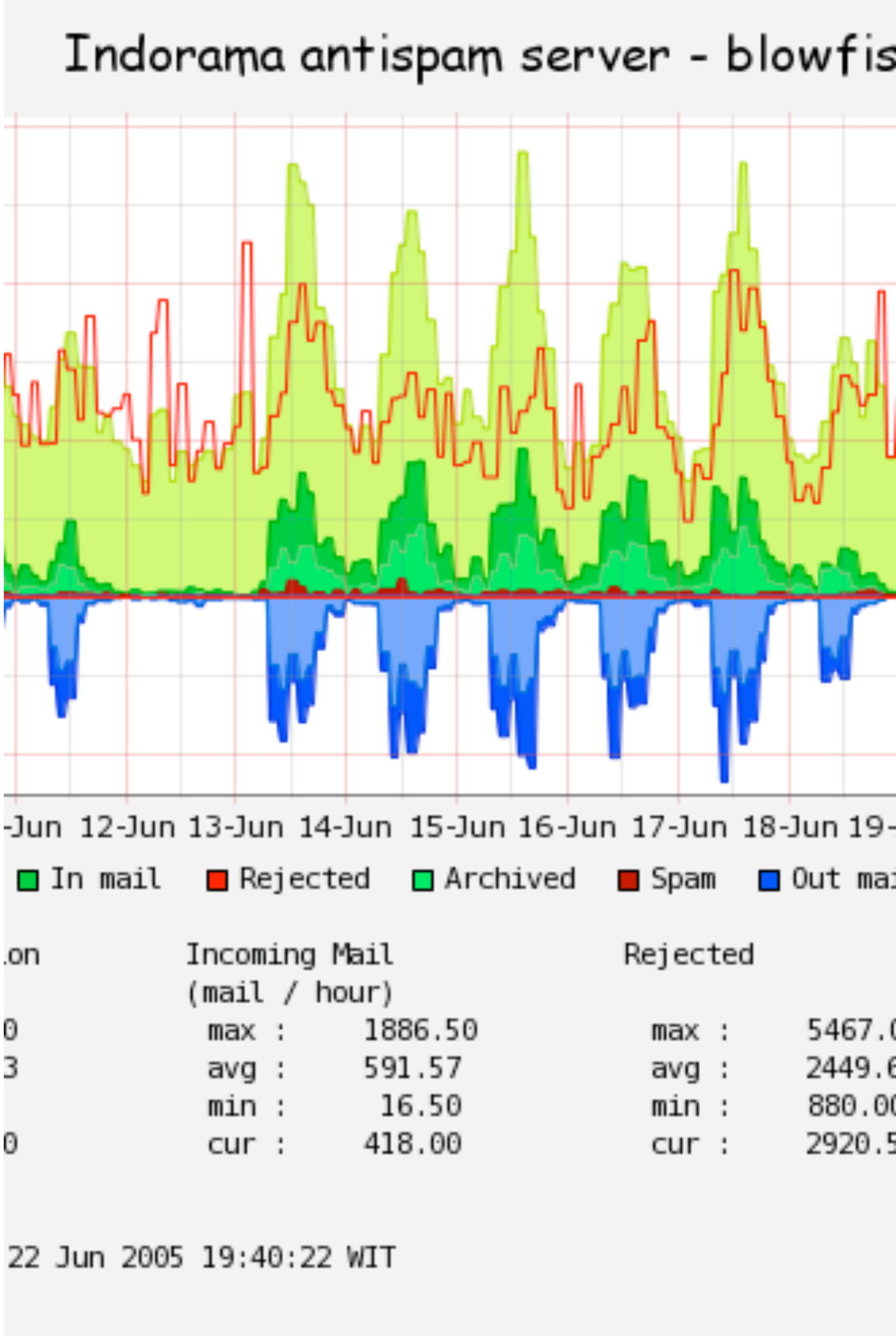
by Tiago Fioreze
Co-authors: Lisandro Zambenedetti Granville, Aiko Pras, Anna Sperotto, and Ramin Sadre

** Tiago Fioreze, Lisandro Zambenedetti Granville, Aiko Pras, Anna Sperotto, Ramin Sadre. "Self-management of Hybrid Networks: can we trust NetFlow data?" Mini-conference proceedings of the 11th IFIP/IEEE International Symposium on Integrated Network Management (IM 2009), 1-5 June 2009, New York, New York, USA (to appear)*

Joint EMANICS/IRTF-NMRG Workshop on Netflow/IPFIX Usage in Network Management - Oct 30th, 2008, Munich, Germany

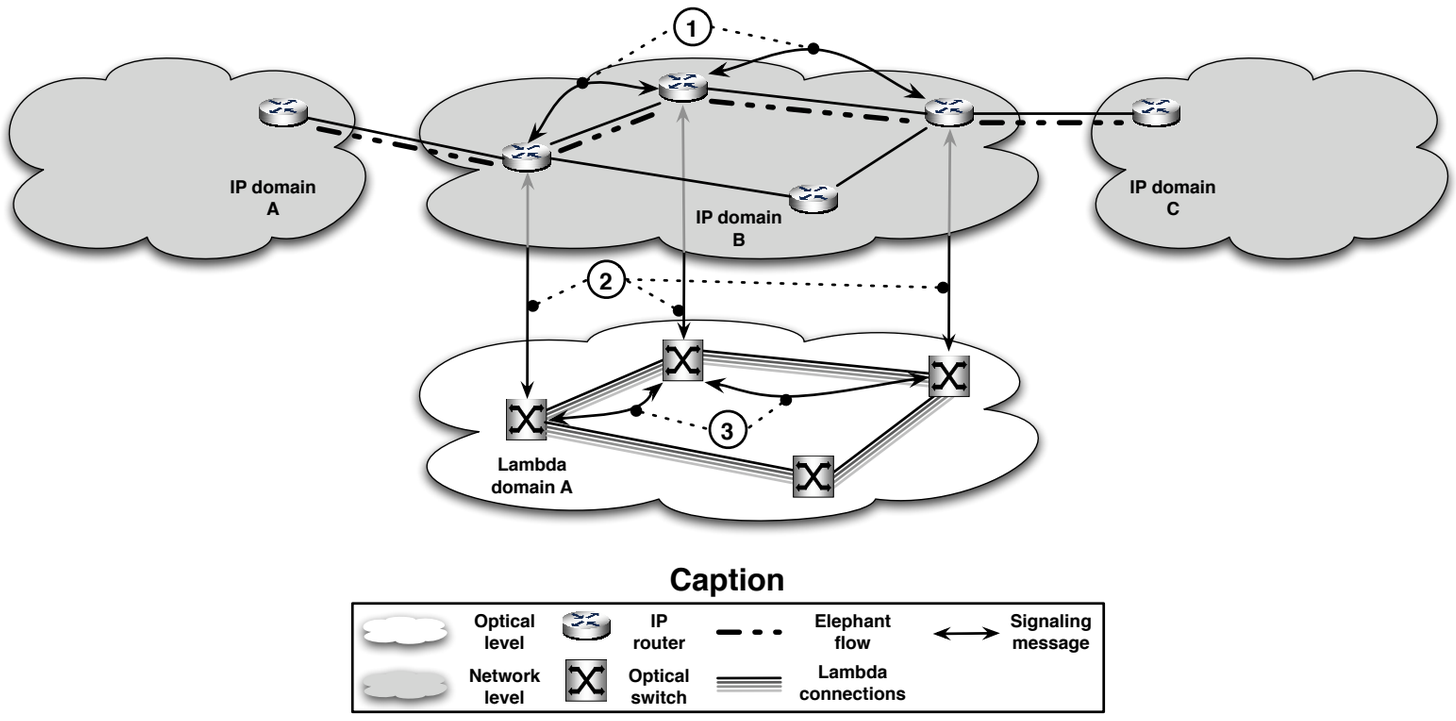
Introduction

- Network measurement provides vital information on the health of managed networks
- The collection of network information can be used for several purposes (e.g., security)
- There are some solutions to collect flow information out there: packet sniffers, SNMP, NetFlow, and maybe more!!!
- ***We focus on the use of Sampled NetFlow to collect real network information and how reliable such information is for our self-management of hybrid networks approach***



What is self-management of hybrid networks?

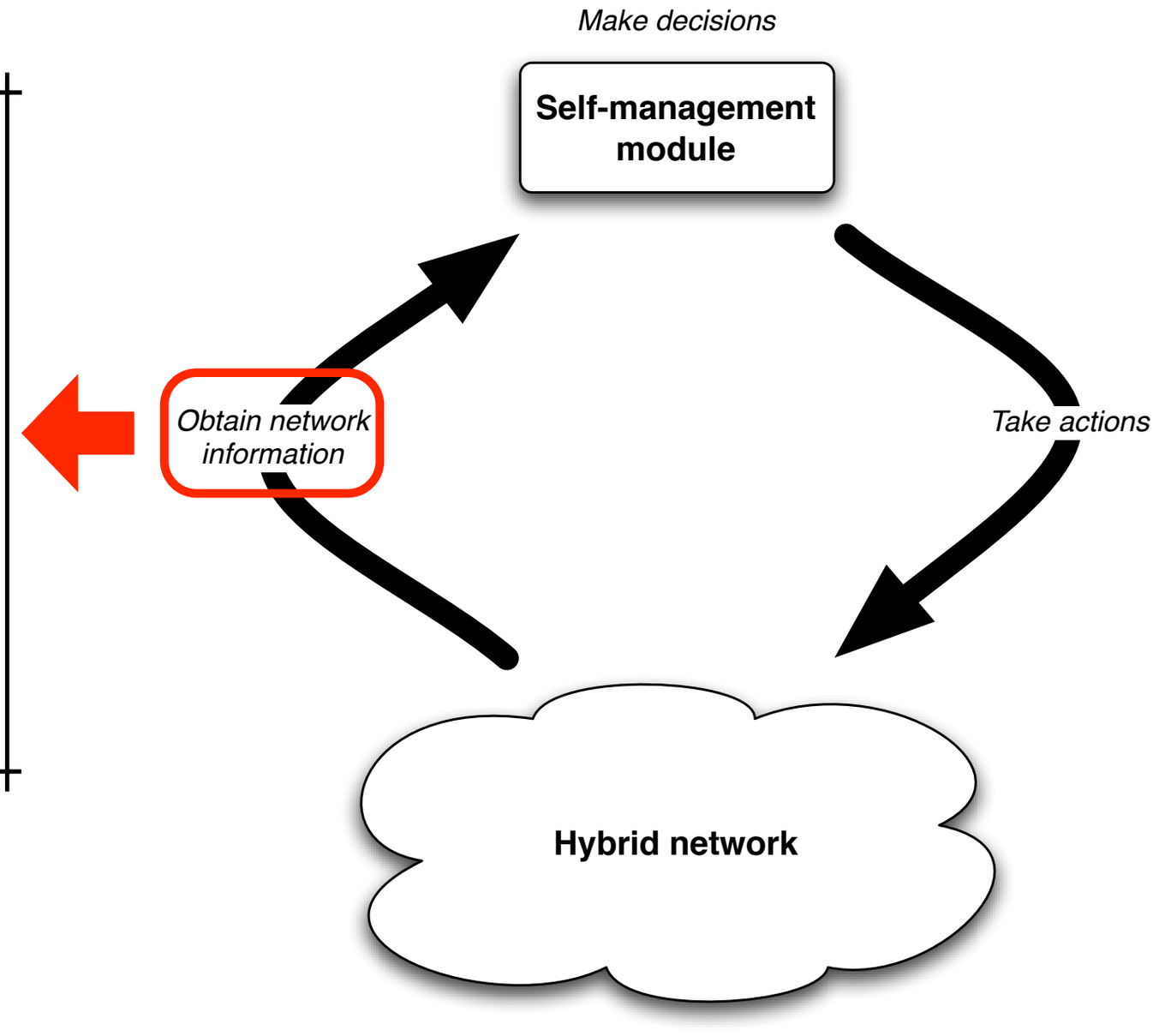
- It consists of a cooperation between the IP and optical network levels to automatically detect IP flows eligible to the optical level as well as establish/ release lambda-connections for them



- Objective: to move as much traffic as possible from the IP level to the optical level, that is, ***to offload elephant flows from the IP level onto lambda-connections at the optical level***

The self-management diagram

In order to make decisions our self-management approach relies on the information obtained from the managed network.



Pre-analysis decisions

- Why has NetFlow been chosen?
  - Because it is the most popular employed solution for flow measurement and it has been strongly influencing the definition of Internet Protocol Flow Information eXport (IPFIX)
- Why has Sampled NetFlow been considered?
  - Because high-speed networks (e.g., TBps/day) employ packet sampling in their measurement process in order to decrease the amount of processed data
- Why the collection of data from real networks?
  - Even though simulation tools and controlled lab networks could be employed to reproduce a network being measured using packet sampling, none of them can 100% capture the real behavior of sampling on actual networks

How we performed our analysis

1) Collection of network flows from 3 different networks using NetFlow with different sampling ratios

Network	Sampling ratio	Inactive timeout	Active timeout
UT	1:1	15s	60s
SURFnet	1:100	30s	300s
GÉANT	1:1000	60s	300s

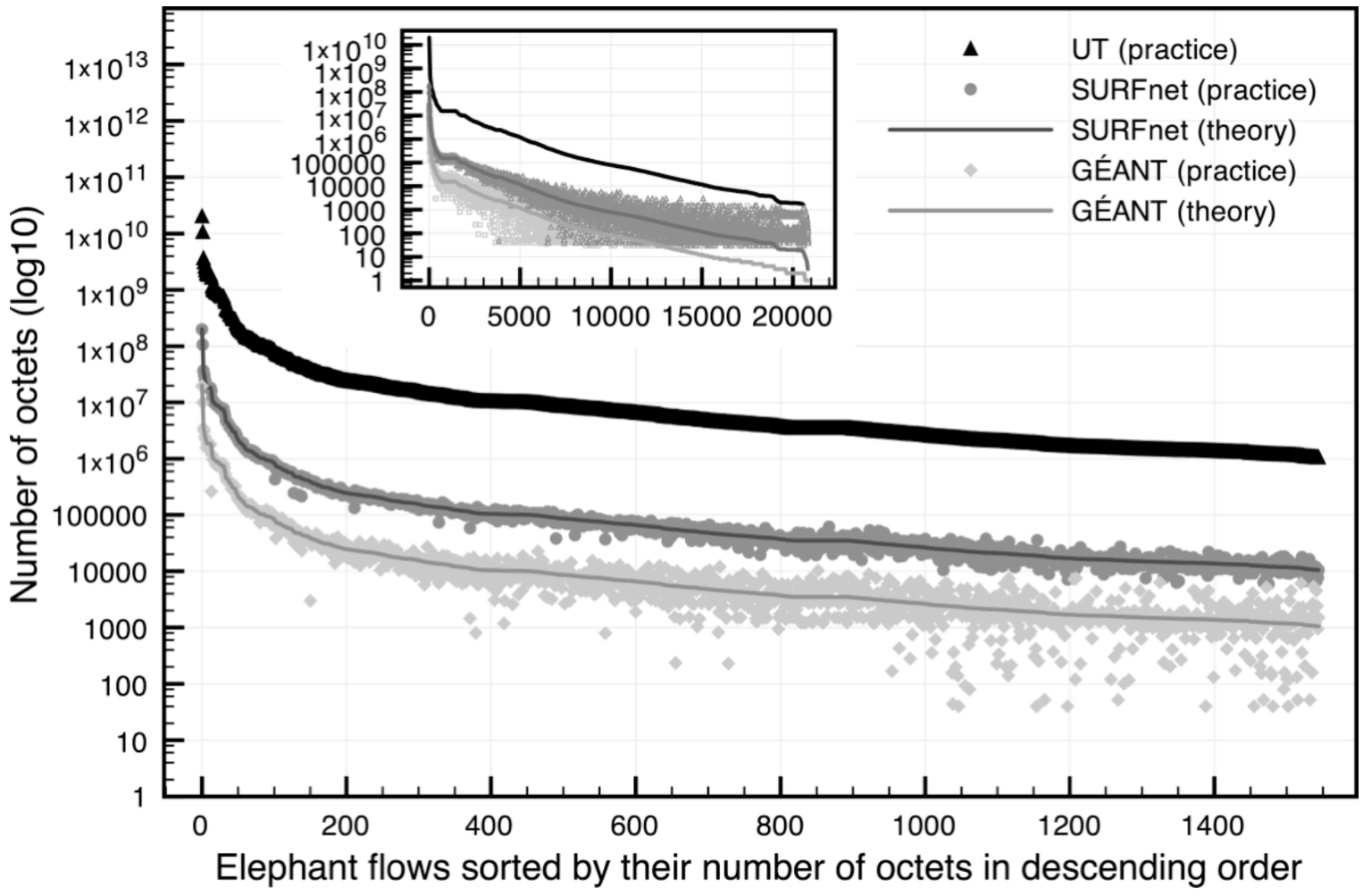
2) Combining of NetFlow records into flows: consecutive NetFlow records with gaps smaller or equal to 30 seconds are grouped into the same flow

3) Selection of long-living flows that transited in the 3 considered networks and that generated most of the observed traffic (i.e., focus on elephant flows)

4) Consideration of 3 flow metrics: octets, packets, and duration, by observing their expected value in theory and their obtained value in practice

Network	Octets	Packets	Duration
UT	O	P	D
SURFnet	O/100	P/100	D
GÉANT	O/1000	P/1000	D

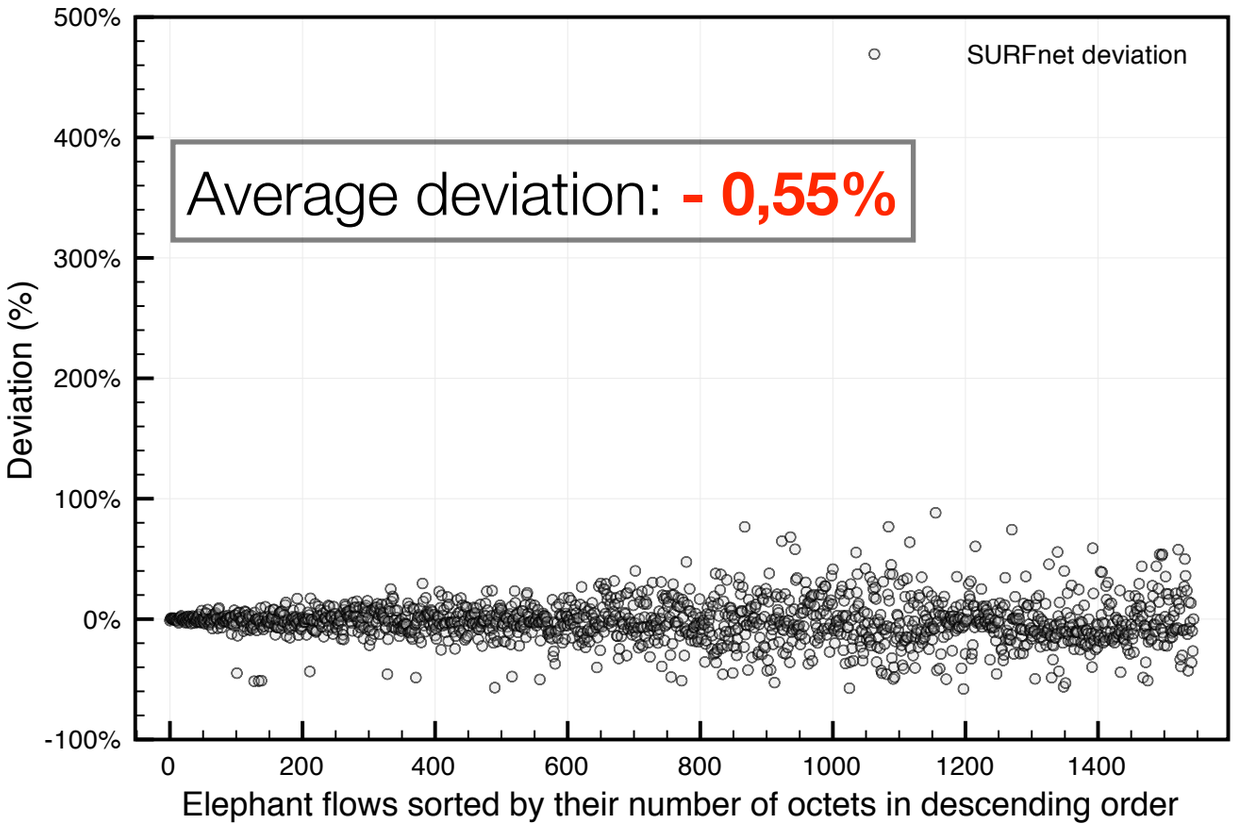
Octets



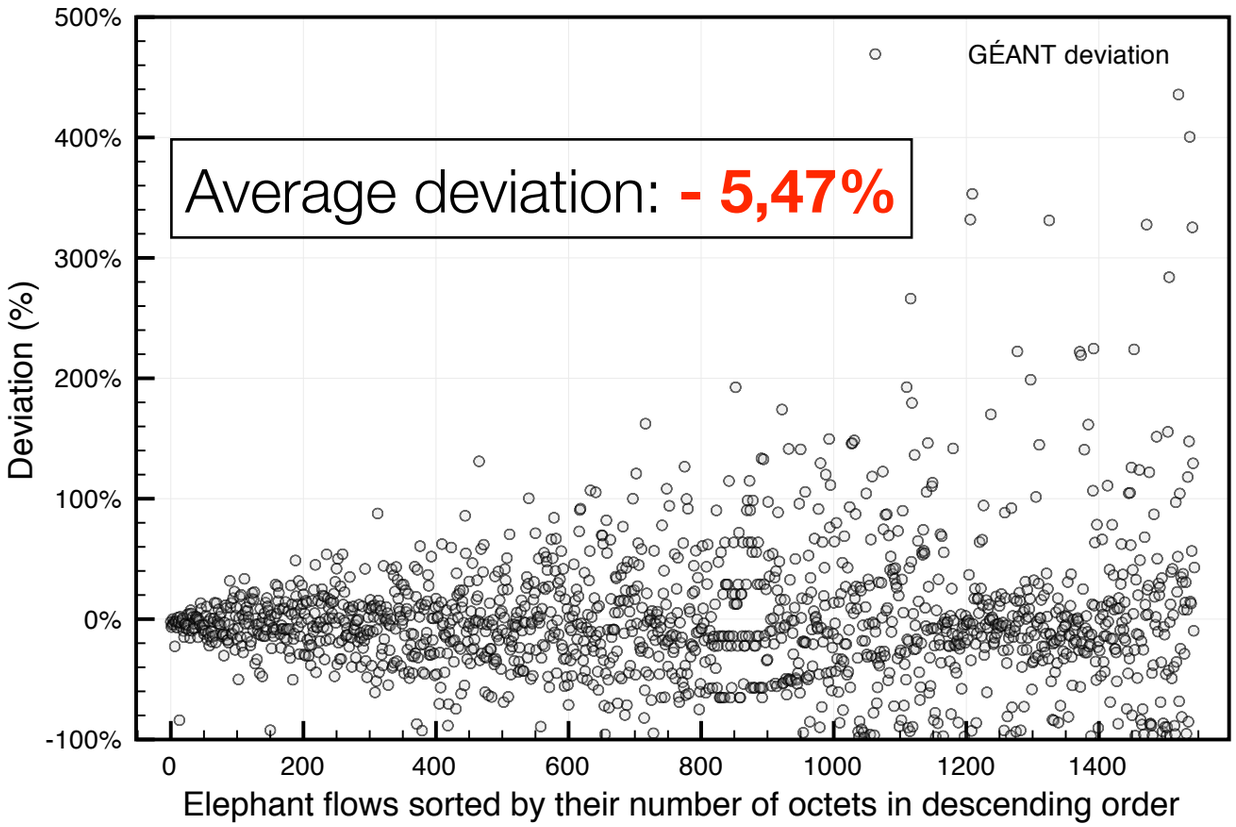
The bigger the flow volume is, the closer to the reality its behavior is reported. Therefore, elephant flows are more precisely reported than mice flows

Octets

(a) SURFnet (1:100)

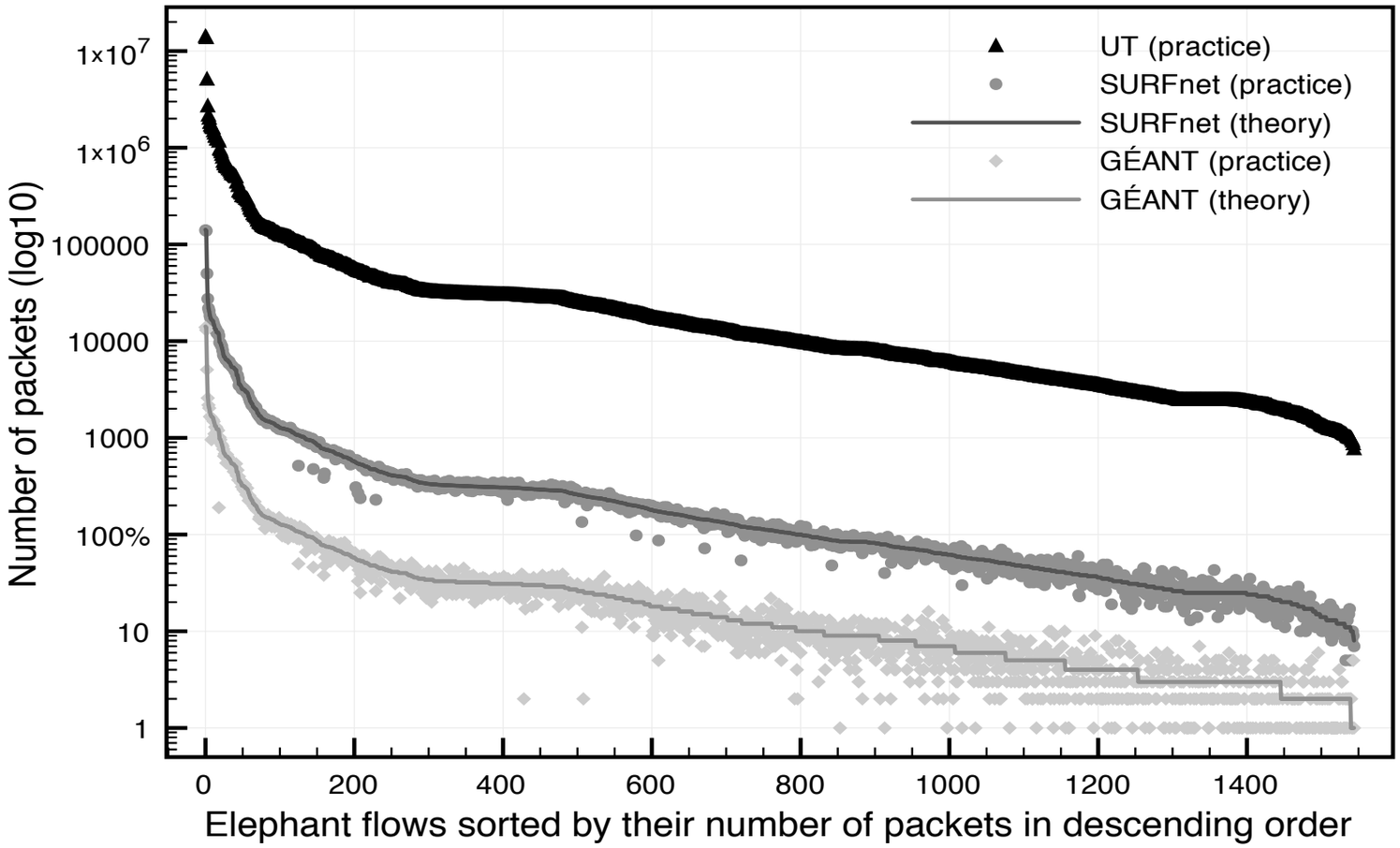


(b) GÉANT (1:1000)



The deviation from the expected value increases when: 1) smaller flows are sampled and 2) bigger sampling ratios are employed

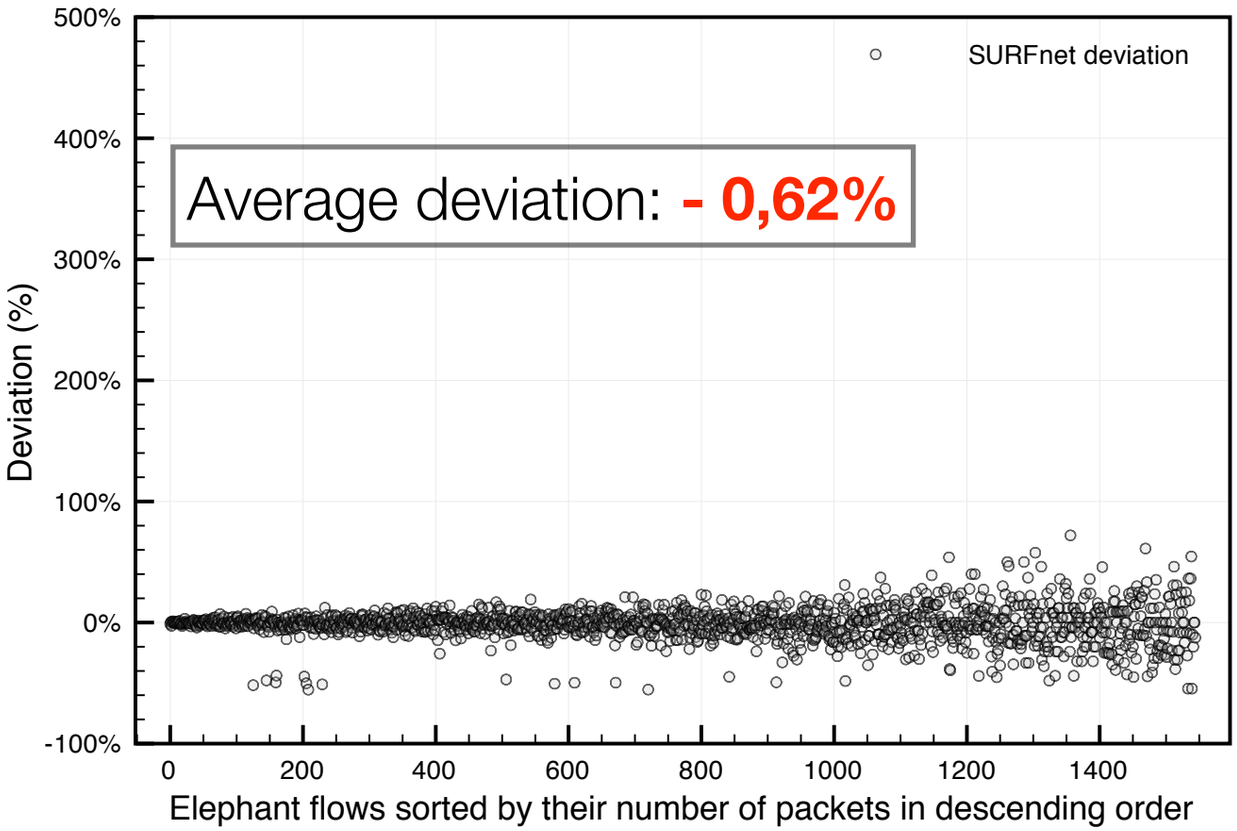
Packets



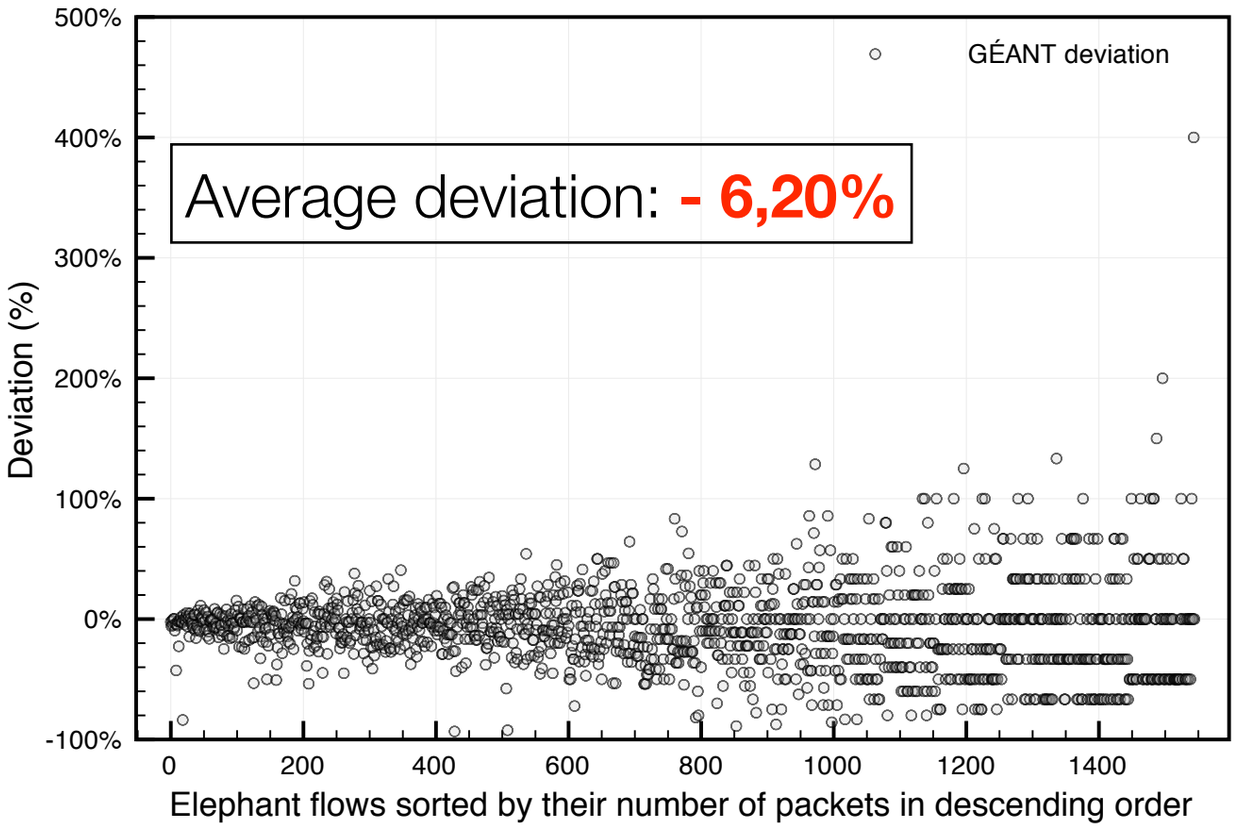
The larger the amount of packets generated, the higher the chances are these packets are sampled

Packets

(a) SURFnet (1:100)

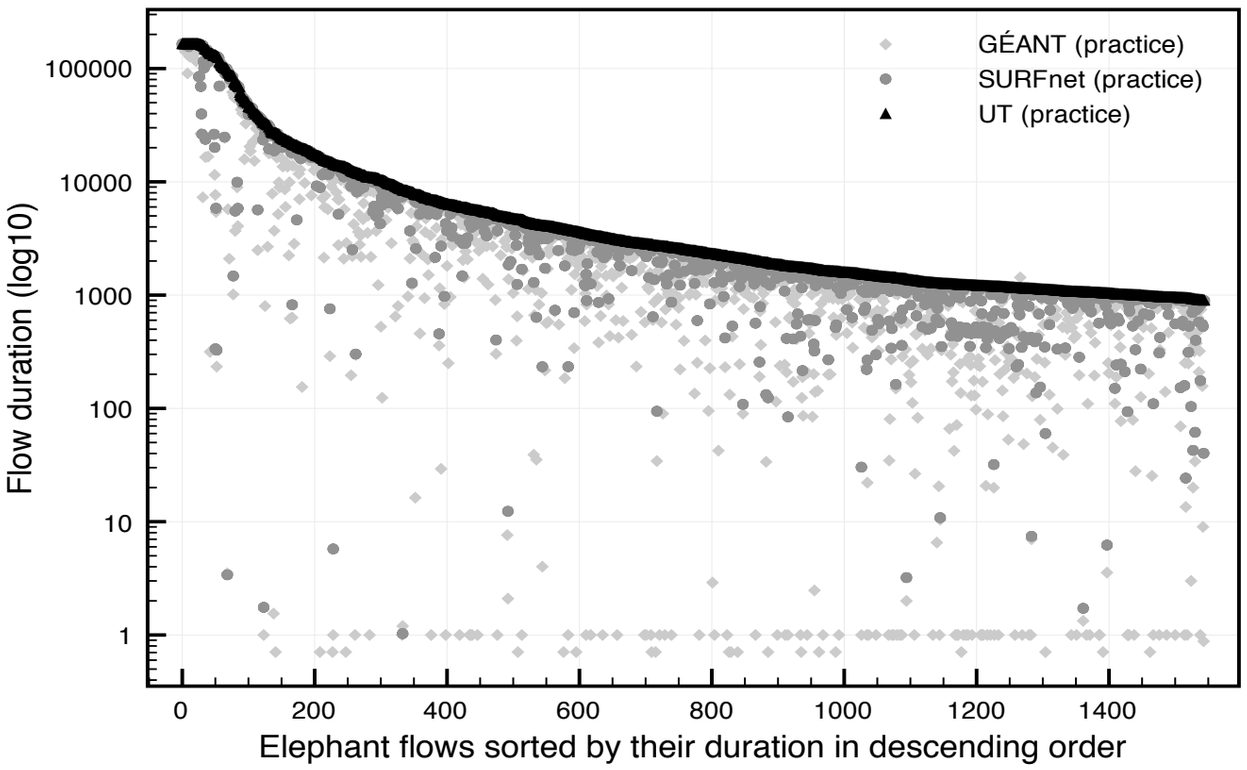


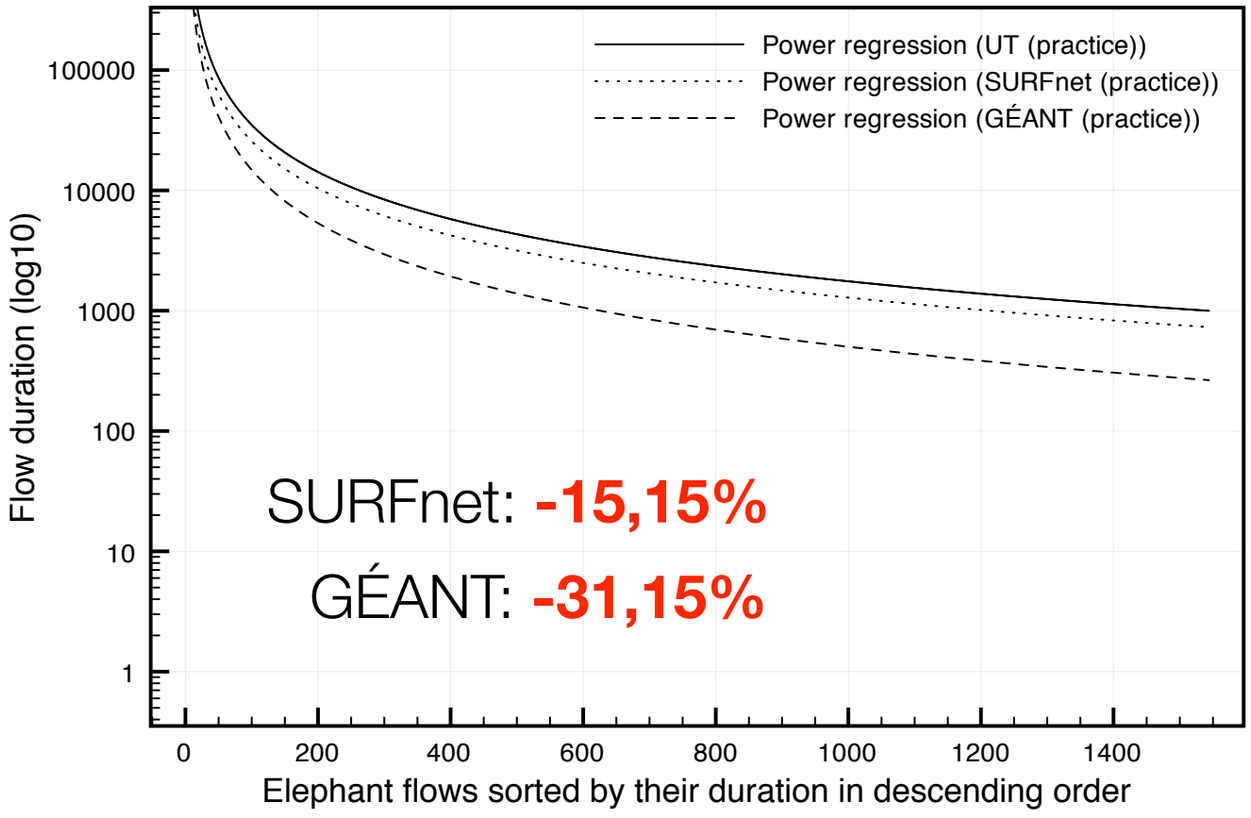
(b) GÉANT (1:1000)

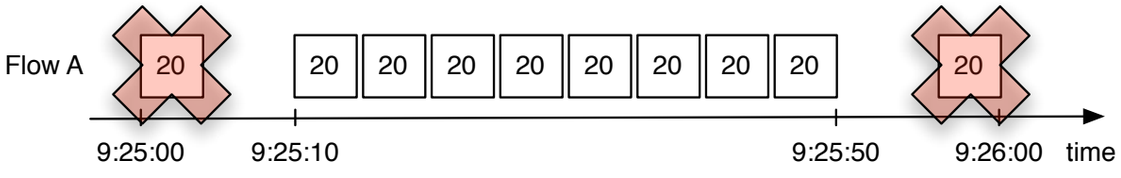


The deviation from the expected value increases when: 1) lesser packets are sampled and 2) bigger sampling ratios are employed

Duration



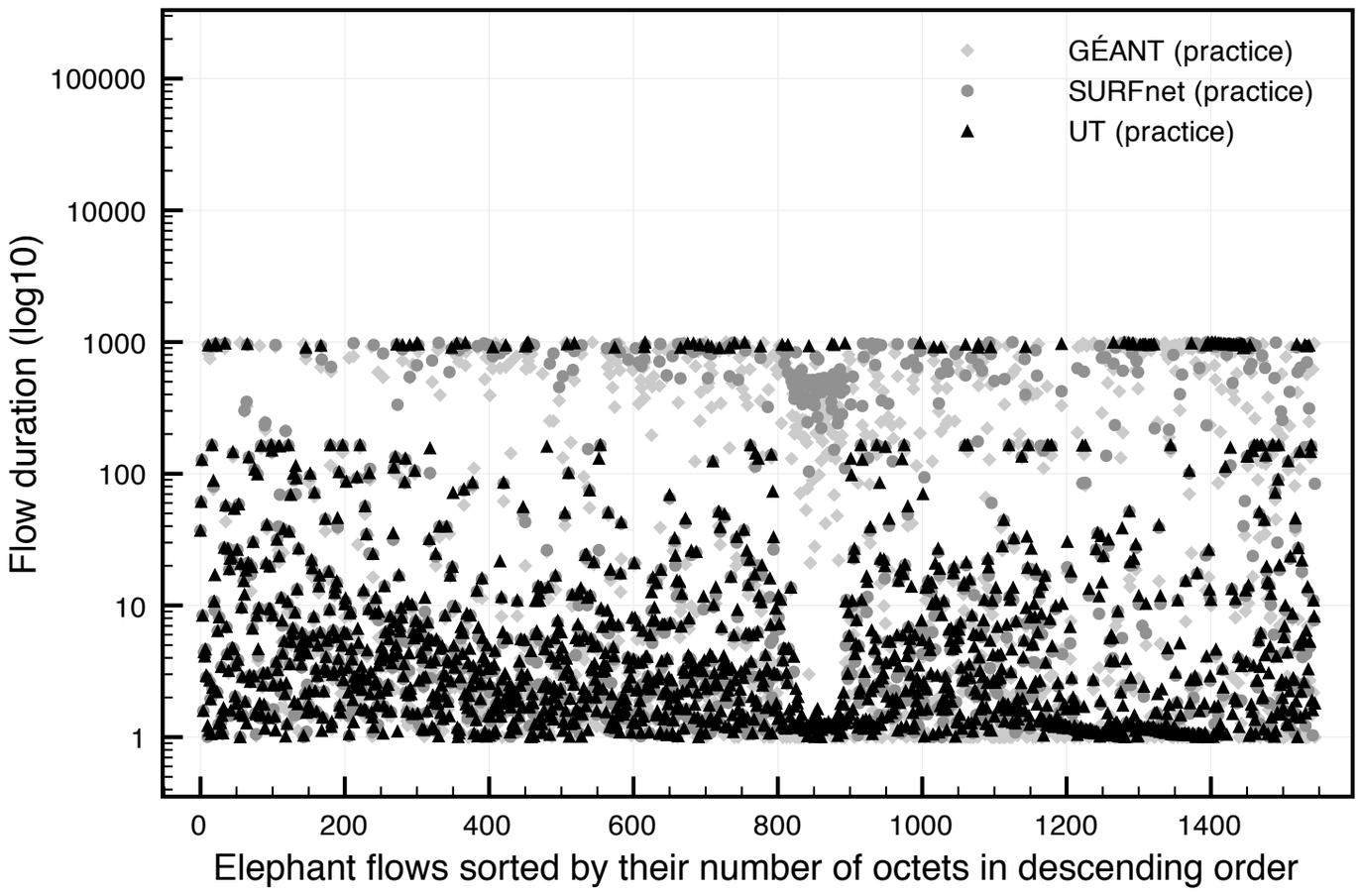




Flow A in reality	**Flow A sampled**
Octets: 200	Octets: 160 (-20%)
Packets: 10	Packets: 8 (-20%)
Duration: 60 sec	Duration: 40 sec (**-33%**)

Flow duration is more sensitive for missing packets, specially in the flow outermost

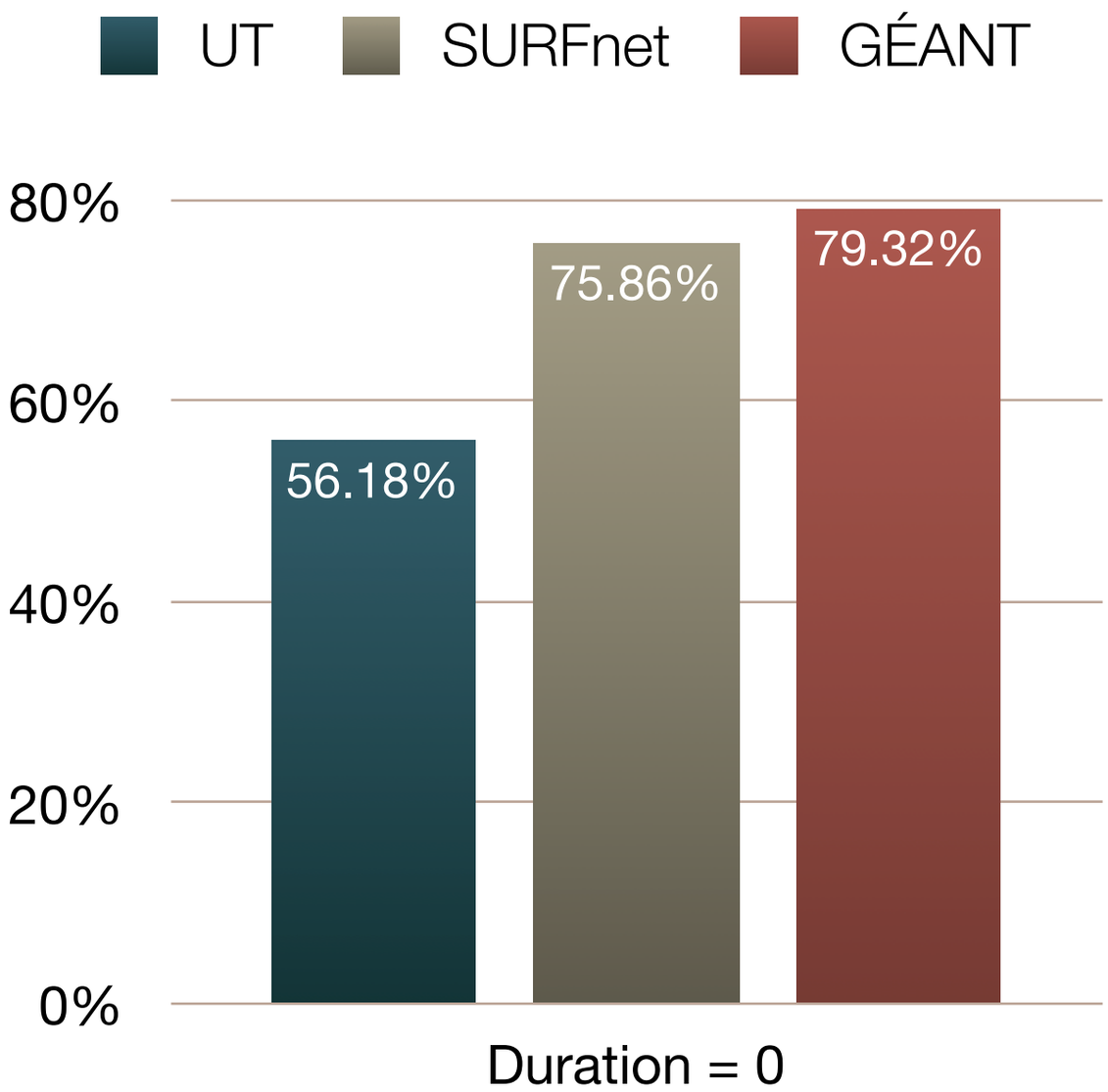
Flow size vs flow duration



There is no relation between flow volume and its duration!!!

A flow can generate lots of packets in a short duration, few packets in a long duration and and all in between.

Tons of flows with duration equal to zero



- ***Main reasons for that:***
  - ***The existence of applications (e.g., NTP) that periodically send control messages ->*** If this periodicity is bigger than the NetFlow inactive timeout, thousands of flows in cache may expire and be exported with duration equal to zero
  - ***The usage of sampling ->*** sampling increases the chance of not inspecting a packet belonging to an existent flow in cache and therefore increases the chance a flow with a single packet is exported due to inactivity

Conclusions and future work

- ***Can we trust (sampled) NetFlow data?***
  - Yes, BUT it depends on the parameters you rely on!!!
  - ***Octets*** and ***packets*** have a good reliability specially if only elephant flows are considered (more packets generated, bigger the chances are the packets will be inspected)
  - On the contrary, flow ***duration*** is considerably affected (specially when outermost packets are missed) and it should be used with precaution (e.g, flow throughput calculation)
- As a future work, a further investigation on how background traffic may influence sampling would be interesting as well.

Acknowledgements & contact information

- Special thanks to Roel Hoek (University of Twente), Hans Trompert (SURFnet), Maurizio Molina (GÉANT) as well as some of DACS students for their valuable contribution to this work.

Tiago Fioreze

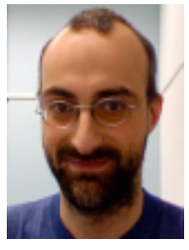
t.fioreze@utwente.nl

Lisandro Granville

granville@inf.ufrgs.br

Aiko Pras

a.pras@utwente.nl



Ramin Sadre

r.sadre@utwente.nl

Anna Sperotto

a.sperotto@utwente.nl

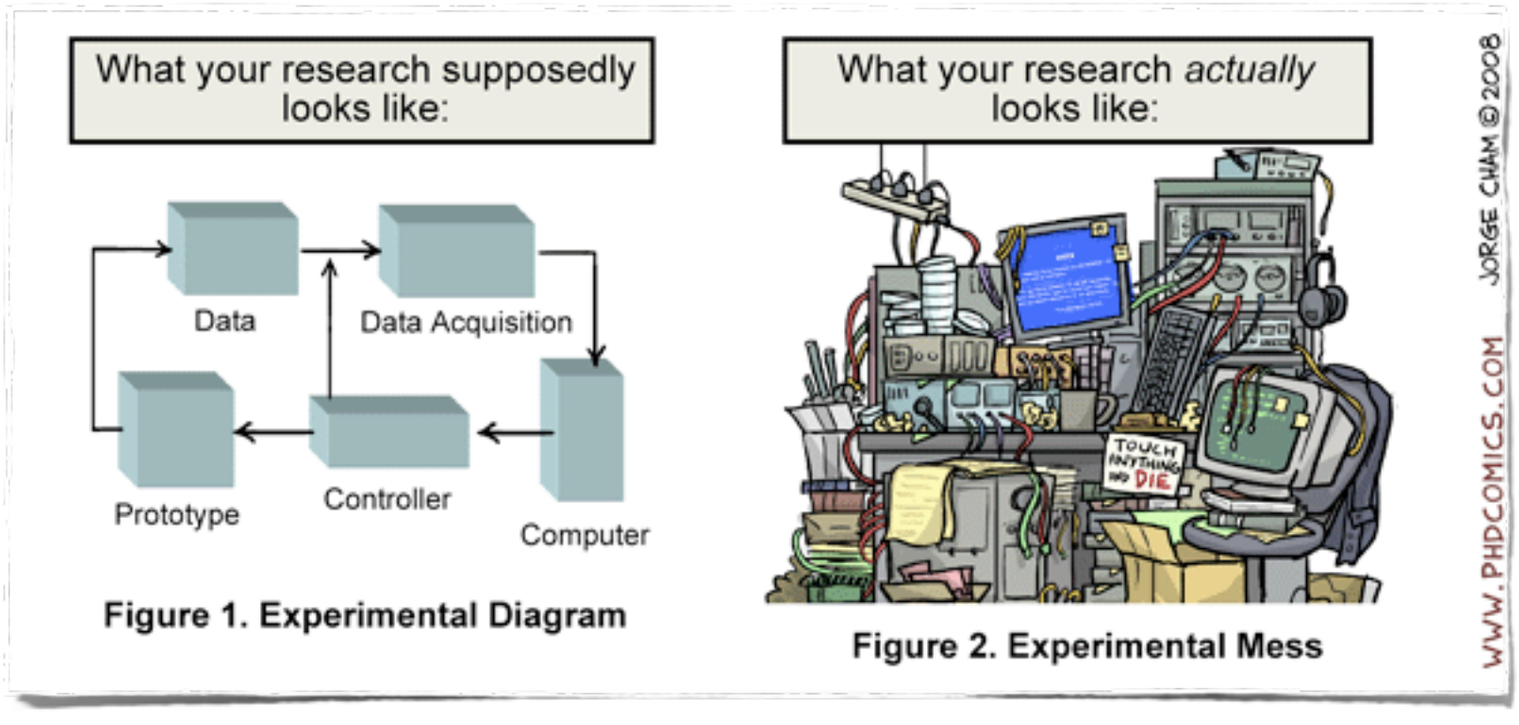


Figure 1. Experimental Diagram

Figure 2. Experimental Mess