CHARACTERIZING CYBERLOCKER TRAFFIC FLOWS

Aniket Mahanti

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Niklas Carlsson

Martin Arlitt

Carey Williamson

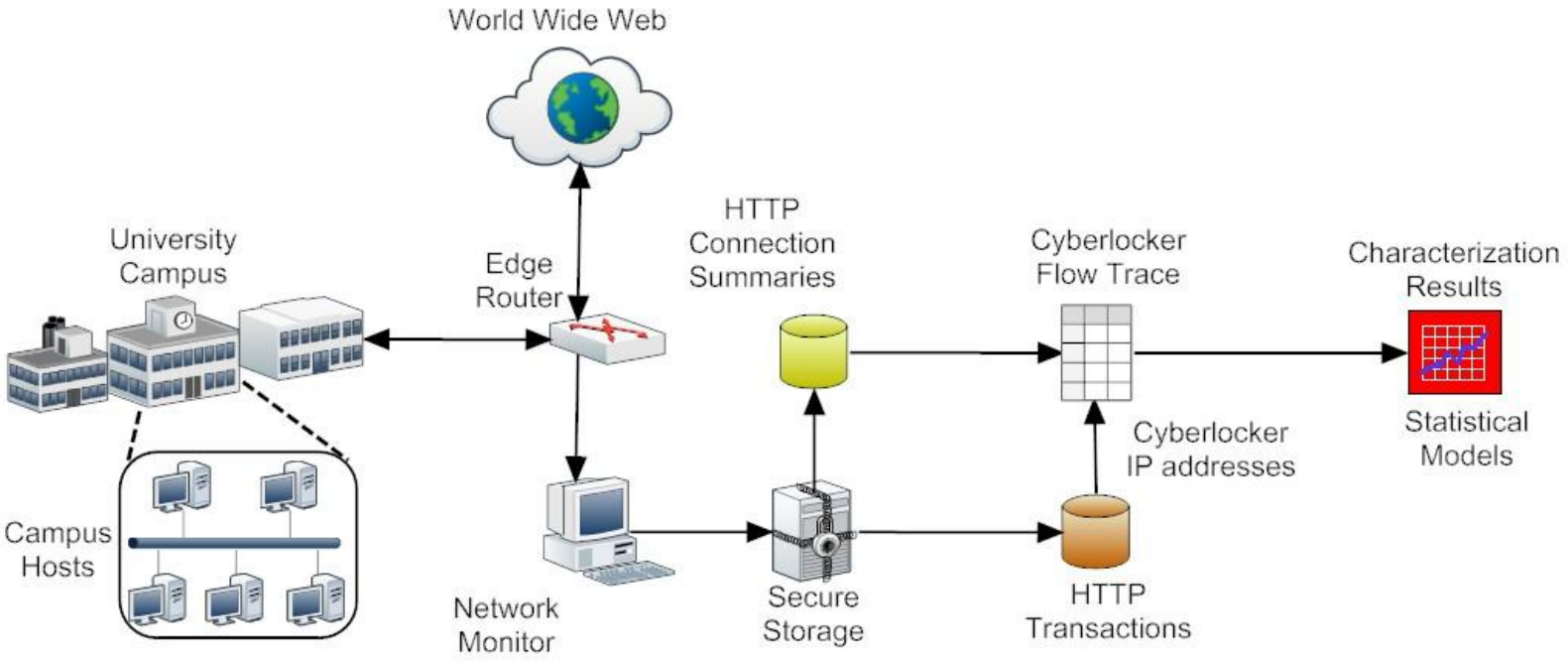
UNIVERSITY OF CALGARY

Introduction

- Cyberlocker services provide an easy Web interface to upload, manage, and share content.
- Recent academic and industry studies suggest that cyberlocker traffic account for a significant fraction of the Internet traffic volume.
- Usage, content characteristics, performance, and infrastructure of selected cyberlockers have been analyzed in previous work.
- In this work, we analyze flows originating from several cyberlockers, and study their **properties at the transport layer** and their **impact on edge network**.

METHODOLOGY

Data Collection

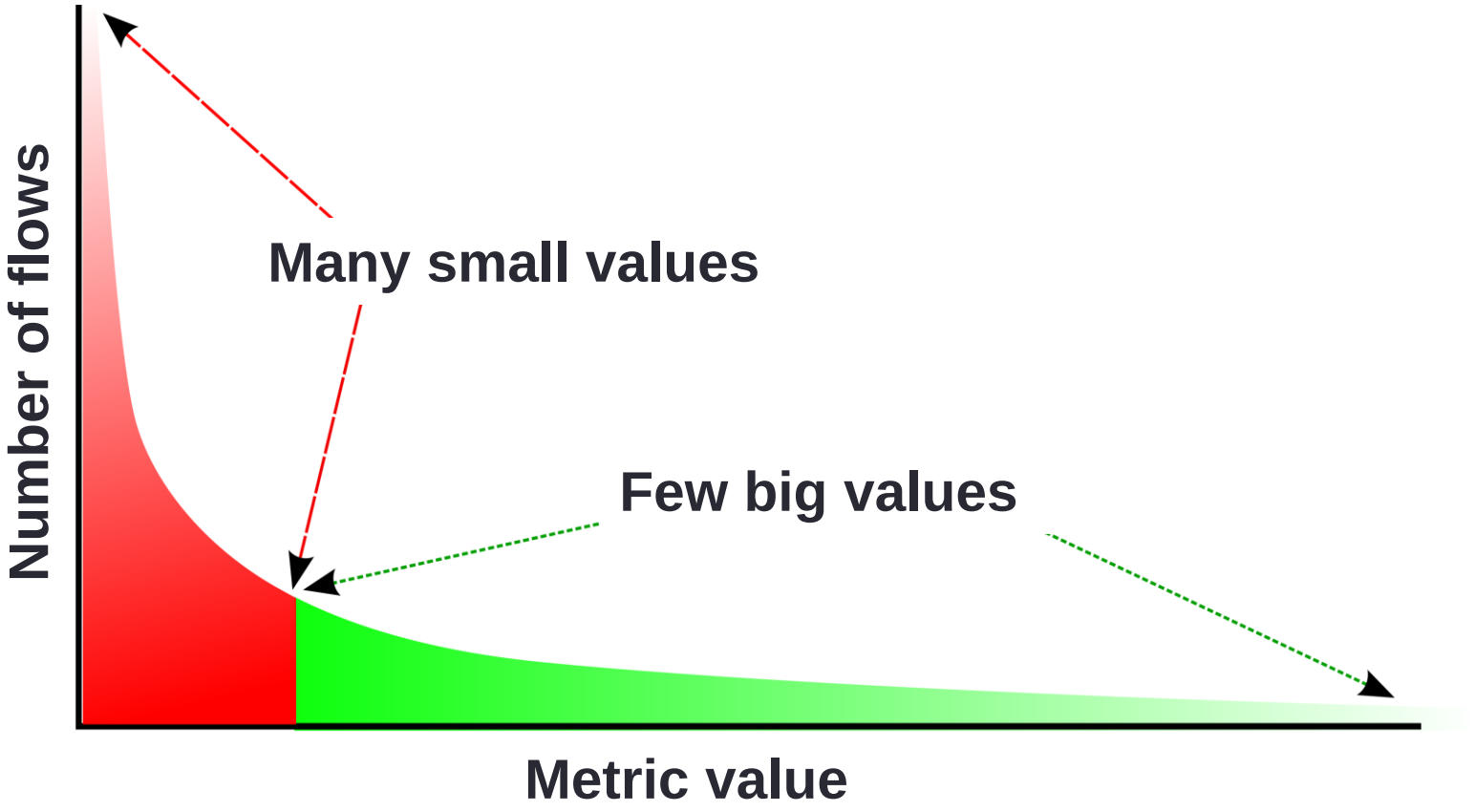


- **Flow-level summaries** were collected using Bro from a *large university edge router* between Jan. 2009 – Dec. 2009
- **HTTP transaction summaries** used to extract IP addresses of *top-10 cyberlocker* services for mapping the flows.

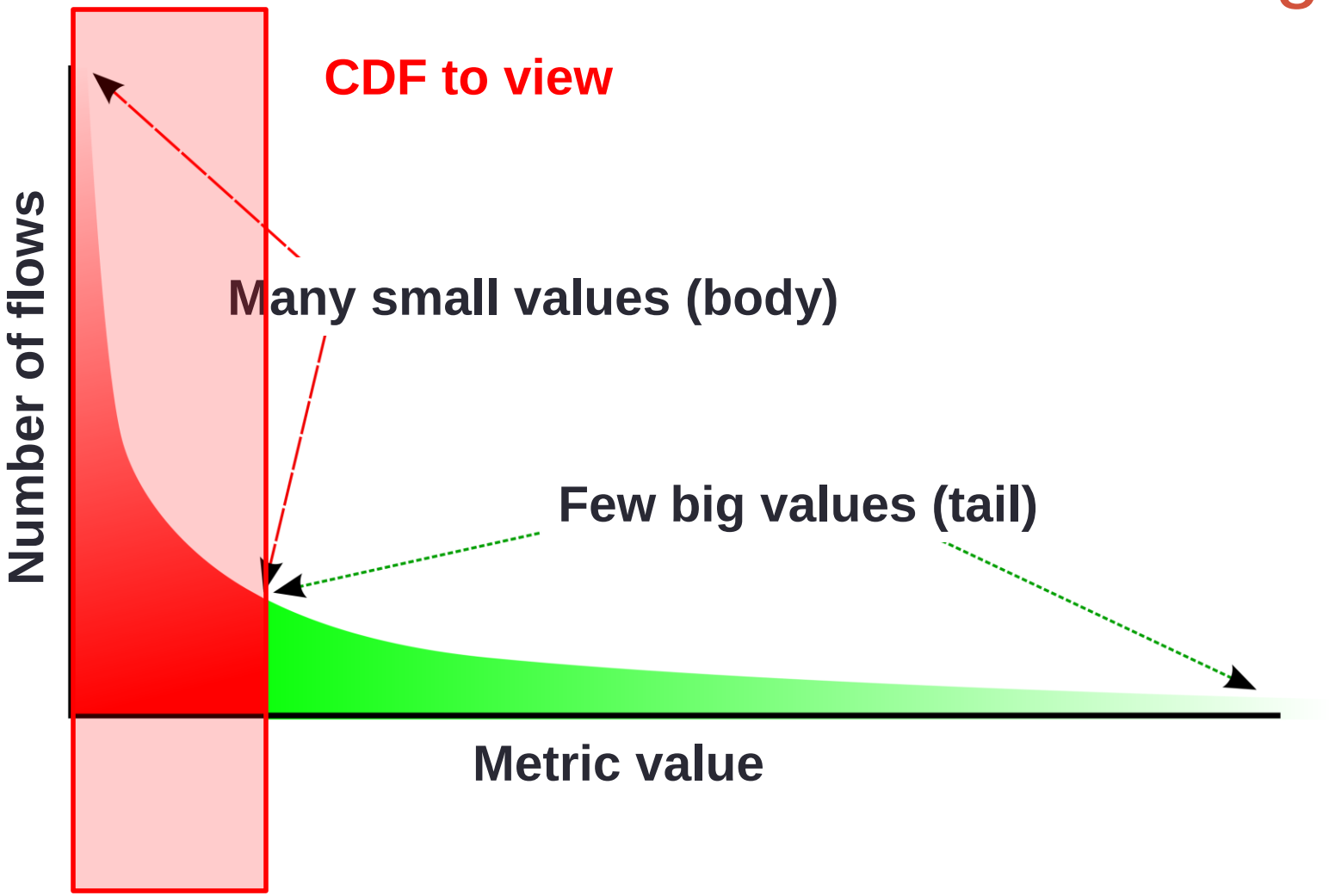
Characterization Metrics

- Flow-level characterization
  - Flow size: The total number of bi-directional bytes transferred within a single TCP flow.
  - Flow duration: The time between start and end of a flow.
  - Flow rate: The average data transfer rate of a TCP connection.
  - Flow inter-arrival time: The time between two consecutive flow arrivals.
- Host-level characterization
  - Transfer volume: The total traffic volume transferred by a campus host during the trace period.
  - On-time: The total time the campus host was active during the trace period.

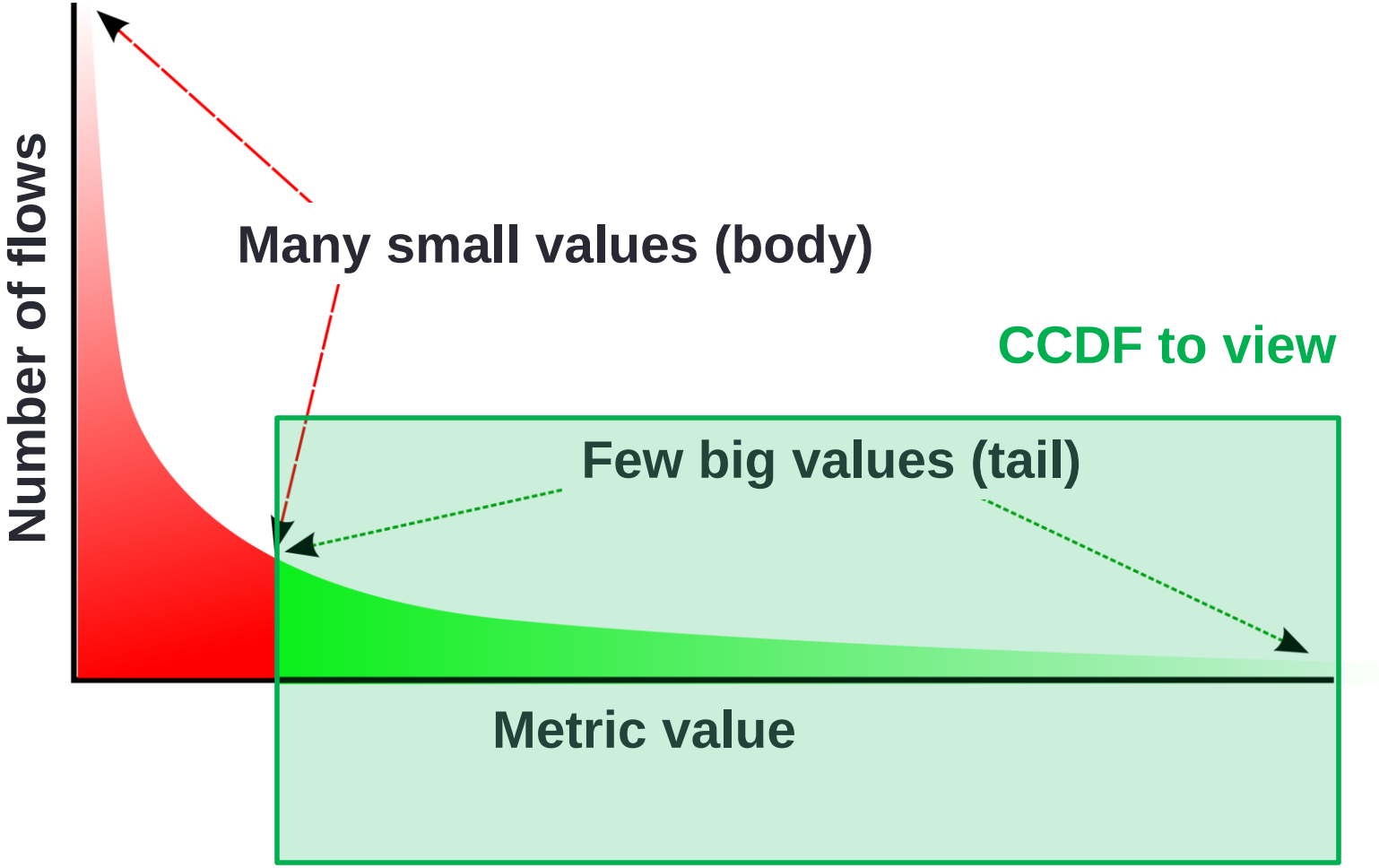
Distribution Characterization and Fitting



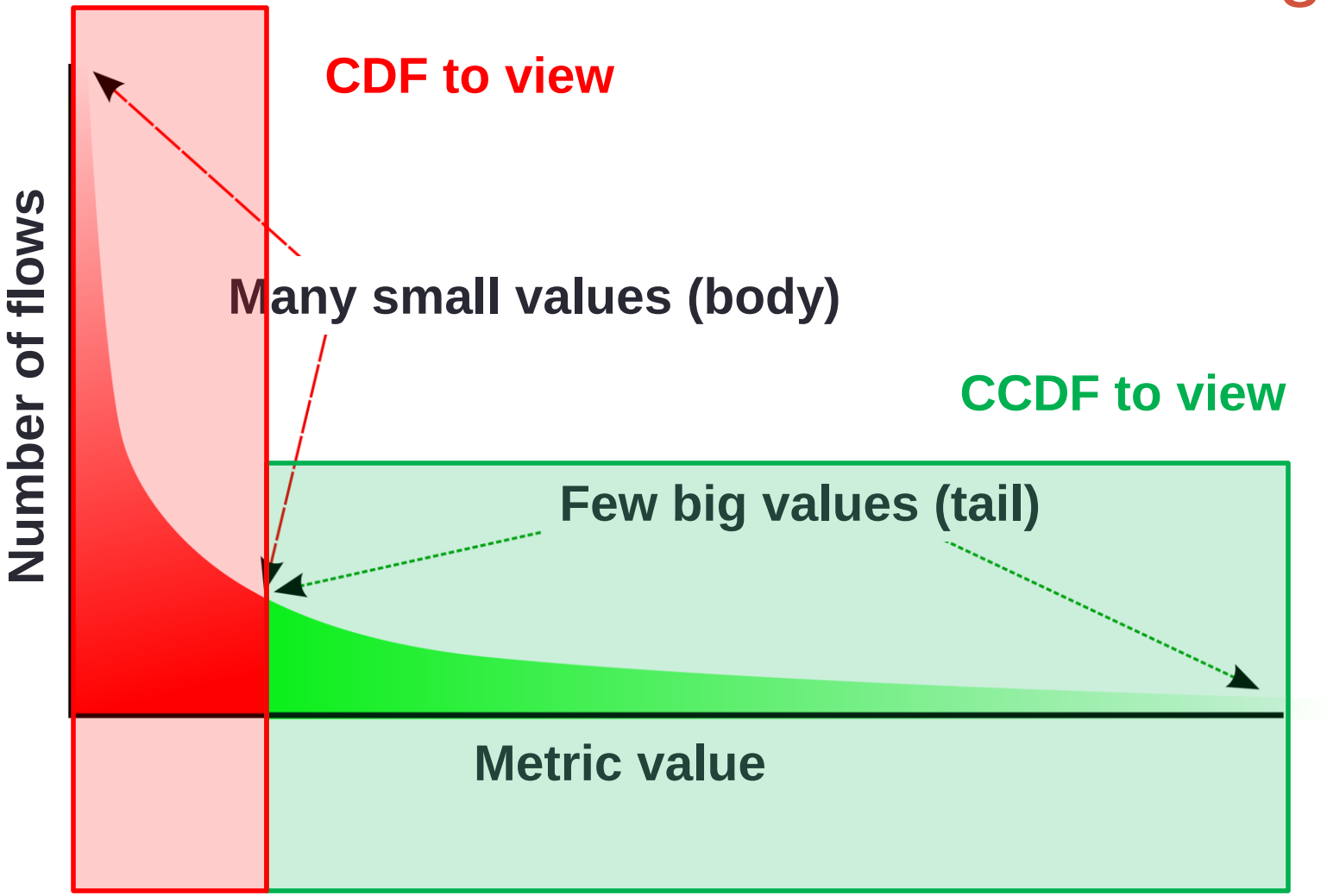
Distribution Characterization and Fitting



Distribution Characterization and Fitting



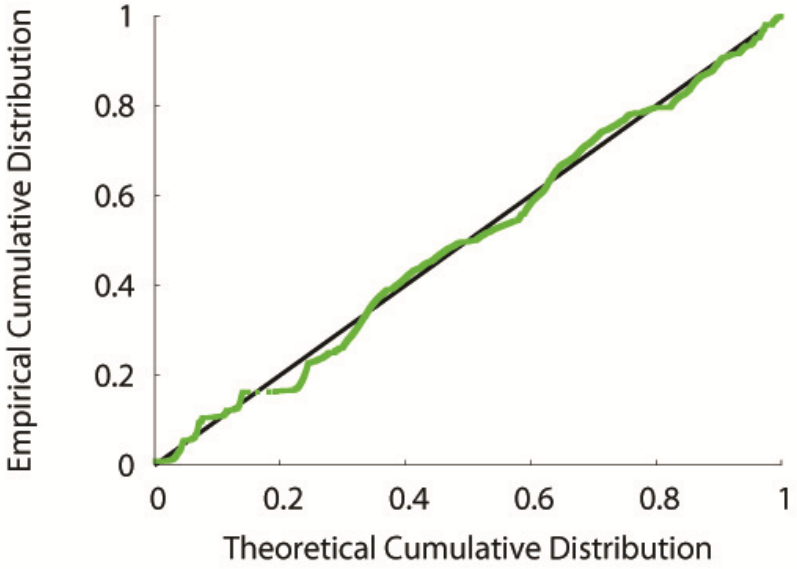
Distribution Characterization and Fitting



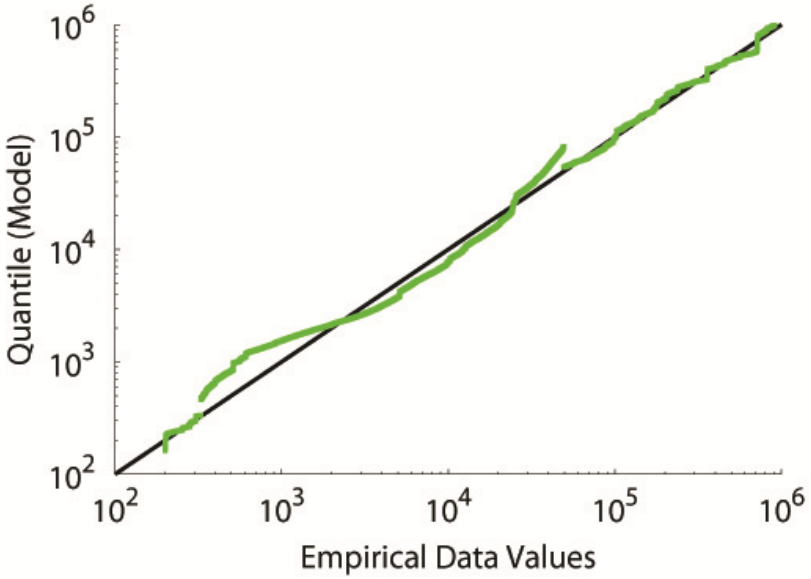
Distribution Fitting and Model Selection

- Complexity of the empirical distribution required us to apply hybrid fits of candidate distributions, where we fit the empirical distributions piece-wise.
- Each empirical distribution was divided into pieces based on manual inspection.
- We fitted seven well-known non-negative candidate statistical distributions (*Lognormal, Pareto, Gamma, Weibull, Levy,* and *Log Logistic*) to each piece and calculated the nonlinear sum of least square error.
- The statistical distribution with the lowest error was chosen.
- After fitting all the pieces of the empirical distribution, we generated the P-P and Q-Q plots; the goodness of the fit was determined by manually inspecting these plots.

Goodness of Fit



(a) Fit of body (majority of flows)



(b) Fit of tail (rare-extreme values)

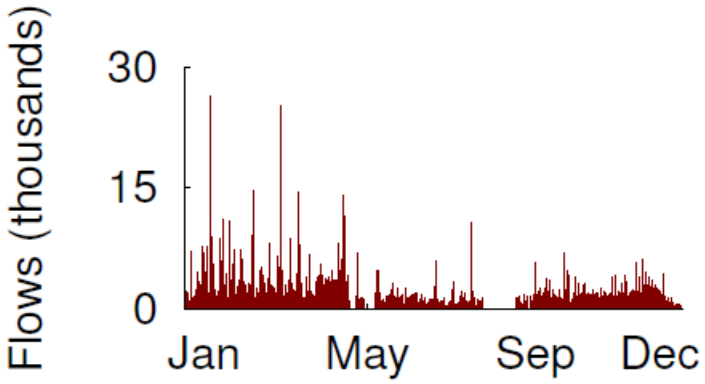
DATASET OVERVIEW

Trace Summary

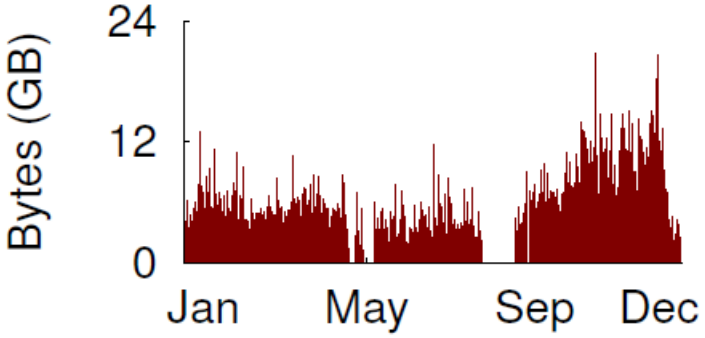
Characteristic	Count
Flow summary log size	1 TB
HTTP traffic	4 billion flows
HTTP traffic volume	488 TB
Top-10 cyberlockers	7 million flows (0.19%)
Top-10 cyberlocker traffic volume	22 TB (4.5%)
Campus hosts using cyberlockers	13,000 hosts

Service	Host	Flows	Bytes
Mega Network (%)	75	43	68
RapidShare (%)	41	42	13
zSHARE (%)	35	4	8
MediaFire (%)	34	8	3
Hotfile (%)	5	0	2
Enterupload (%)	30	1	2
Sendspace (%)	11	1	1
2Shared (%)	7	0	1
Depositfiles (%)	8	1	1
Uploading (%)	5	0	0
Top-10 cyberlockers	13K	7 mil	22 TB

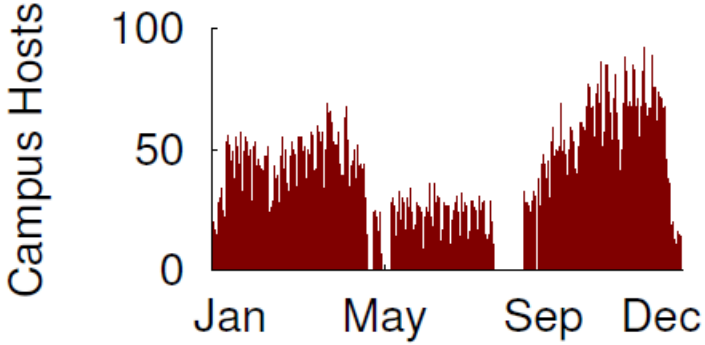
Campus Usage Trends



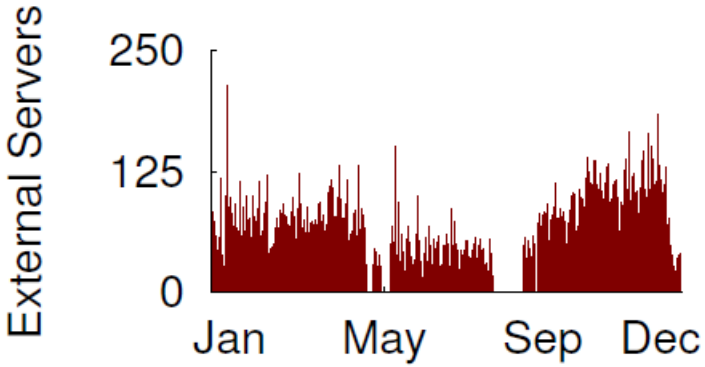
(a) Cyberlocker flows



(b) Bytes transferred by content flow



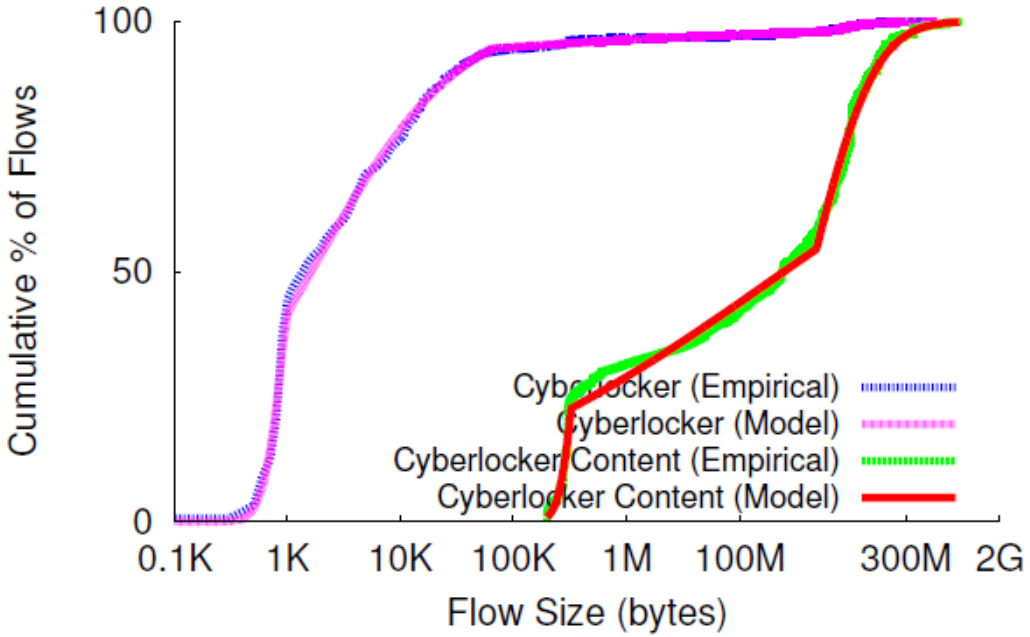
(c) Content flow hosts



(d) Content flow external servers

FLOW-LEVEL CHARACTERIZATION

Flow Size

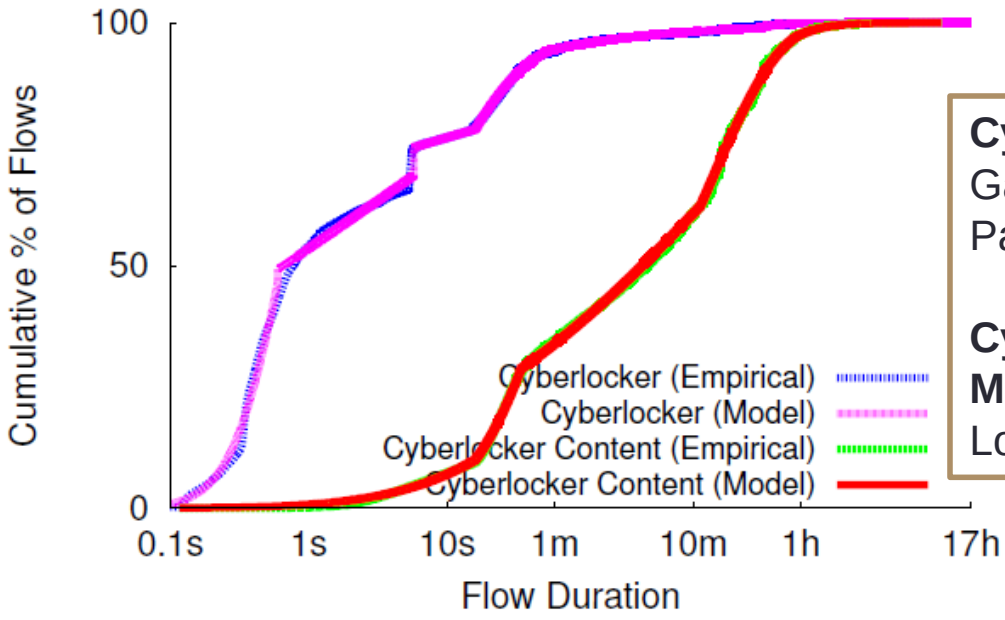


Cyberlocker Model: Lognormal-Pareto

Cyberlocker Content Model: Lognormal

- Content flows only represent 5% of the cyberlocker flows, they consume over 99% of the total traffic volume.
- Content flows are orders of magnitude larger as they transfer large content hosted on the sites.
- Significantly larger flows than typical Web object.

Flow Duration

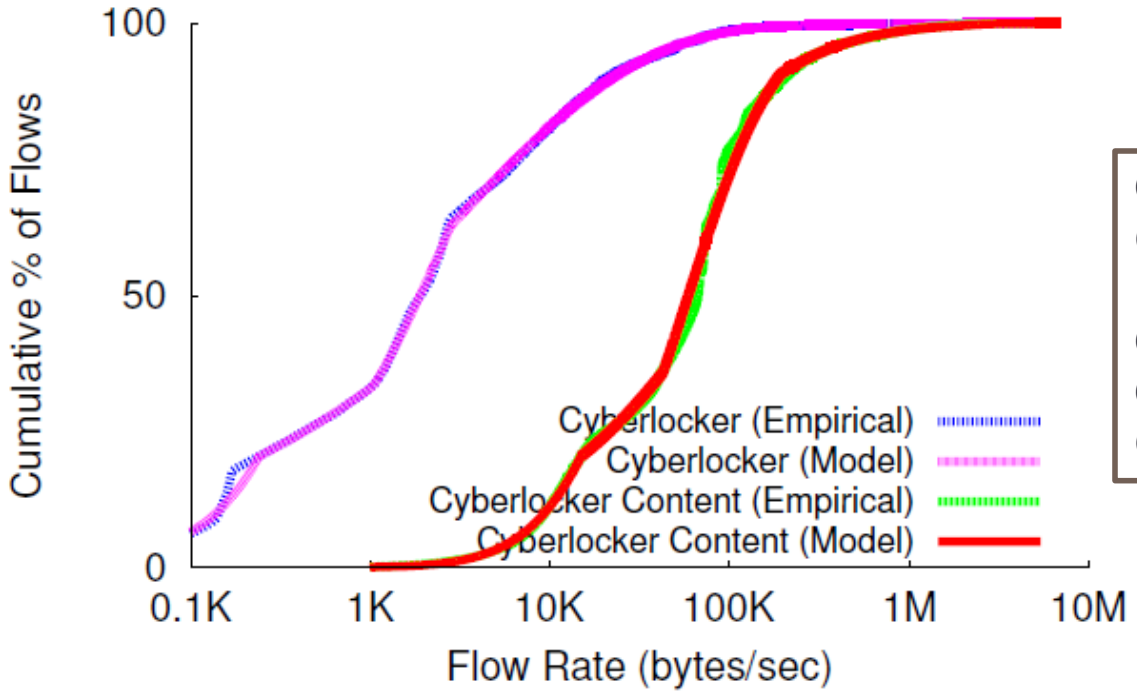


Cyberlocker Model: Gamma-Lognormal-Pareto

Cyberlocker Content Model: Lognormal-Gamma

- Content flows are long-lived, partly due to wait times and bandwidth throttling.
- Most content flows have duration less than 10 minutes due to medium-sized content downloads.

Flow Rate

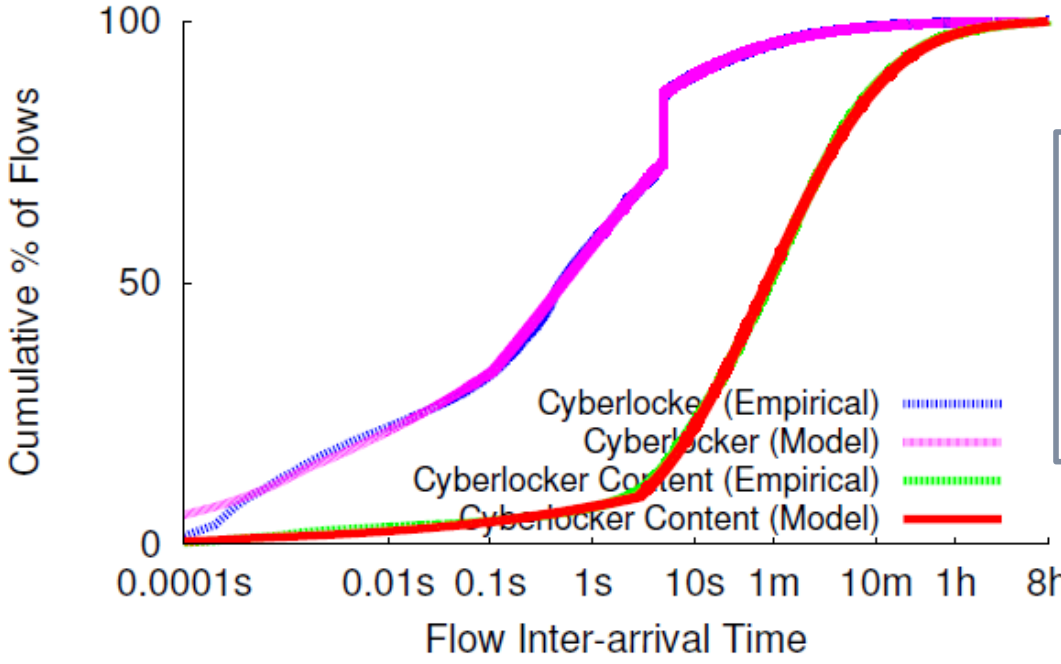


Cyberlocker Model: Gamma

Cyberlocker Content Model: Gamma-Lognormal

- Cyberlocker content flows are larger and long-lived and receive higher flow rates.
- There is presence of both free and premium hosts that download content from the services.

Flow Inter-arrival



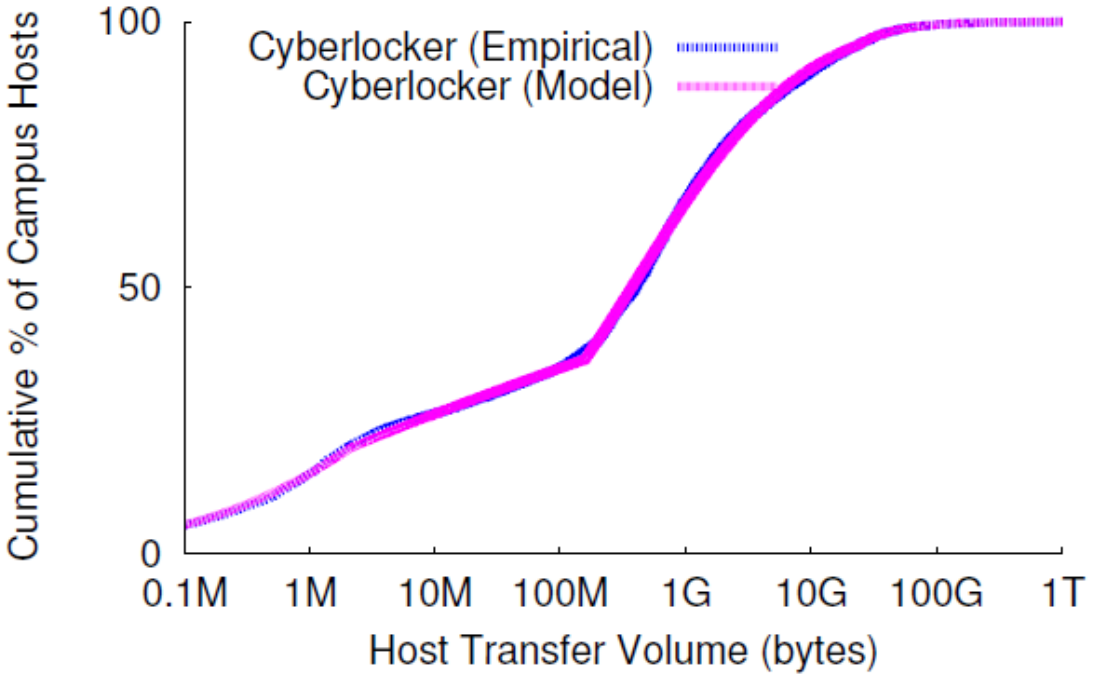
Cyberlocker Model: Lognormal-Gamma

Cyberlocker Content Model: Gamma-Lognormal

- Parallel downloading increases flow concurrency and decreases flow inter-arrivals.
- Content flow inter-arrivals are longer because there are far fewer such flows; most of the flows are due to objects being retrieved from sites.

HOST-LEVEL CHARACTERIZATION

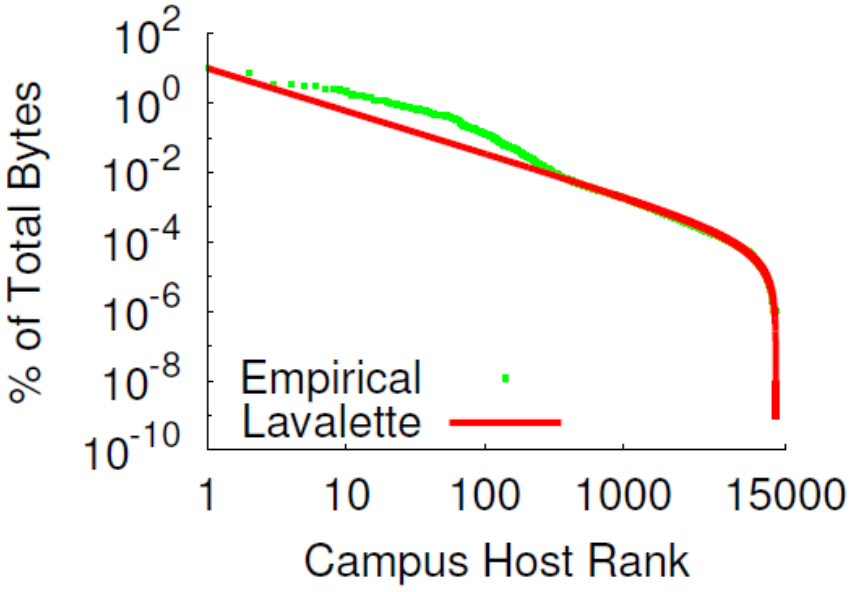
Host Transfer Volume



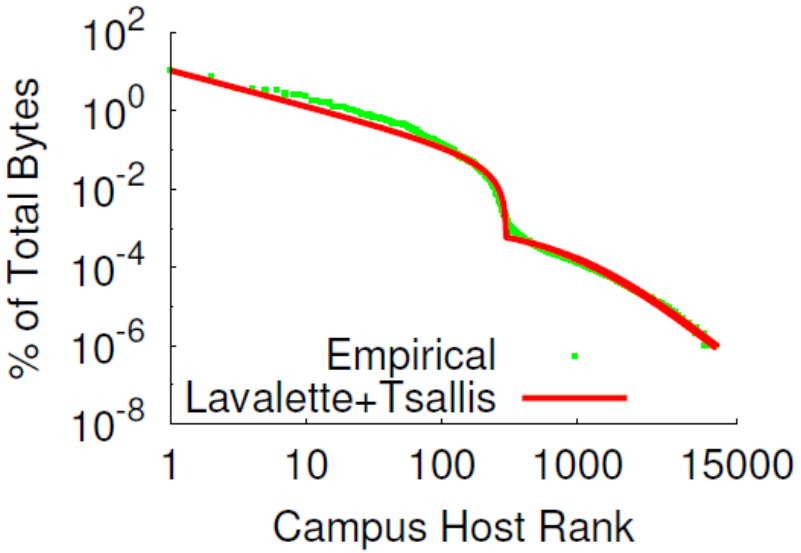
Cyberlocker Model:
Lognormal-Pareto

- There is presence of some hosts that transfer a lot of data as well as hosts that transfer less data.
- Most of the transfer volume is due to content flows.

Heavy Hitters



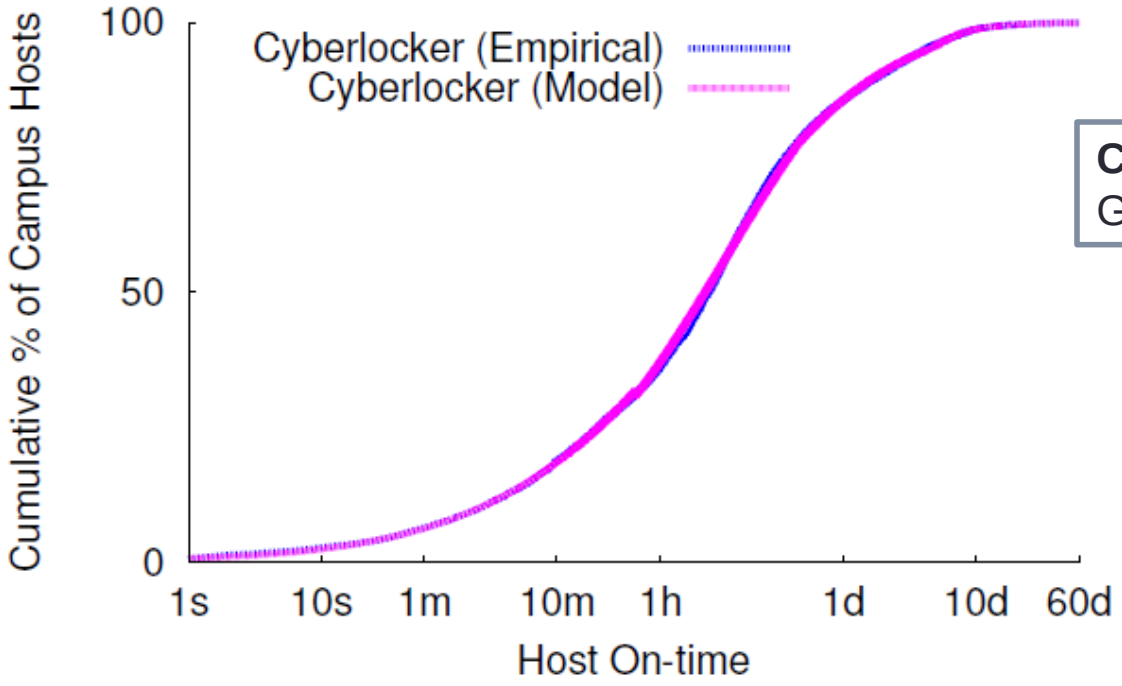
(a) Ranked cyberlocker hosts



(b) Ranked cyberlocker content hosts

- The top-100 ranked hosts account for more than 85% of the cyberlocker and cyberlocker content traffic volume.
- The high skews are well-modeled by non-linear power-law distributions.

Host On-time



Cyberlocker Model:
Gamma-Lognormal

- On-times of cyberlocker hosts are heavy-tailed
- Most of the time spent by hosts is for downloading content.
- Users with premium subscription may spend less time since they can download more content in less time.

CONCLUDING REMARKS

Conclusions

- Cyberlockers introduced many small and large flows.
- Most cyberlocker content flows are long-lived and durations follow a heavy-tailed distribution.
- Cyberlocker flows achieved high transfer rates.
- Cyberlocker heavy-hitter transfers followed power-law distributions.
- Increased cyberlocker usage can have significant impact on edge networks.
- Long-lived content flows transferring large amounts of data can strain network resources.

Aniket Mahanti – University of Auckland, New Zealand
Niklas Carlsson – Linkoping University, Sweden
Martin Arlitt – HP Labs, USA
Carey Williamson – University of Calgary, Canada

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

LINKÖPINGS UNIVERSITET

LABS hp

UNIVERSITY OF CALGARY

QUESTIONS?