Monitoring of Internet traffic and applications

Chadi BARAKAT
INRIA Sophia Antipolis, France
Planète research group

ETH - Zurich - October 2009

Email: Chadi.Barakat@sophia.inria.fr
WEB: http://www.inria.fr/planete/chadi
Our goal

Efficient solutions for passive and active network monitoring

- Passive monitoring: use the existing, don’t inject more traffic
- Active monitoring: measure the Internet by injecting probes

Features:

- Reduce the overhead of passive monitoring
  - Volume of collected traffic, memory access, processing
- Reduce the volume of probes to be injected into the network
  - Targeted applications: network troubleshooting and topology mapping
- Congestion control for data collection and network probing
An example of two activities

- Application identification from packet measurements
  - What can we learn on applications from packet sizes?
  - Is it possible to avoid port numbers and payload inspection?
  - Networking 2009 in Aachen, Germany.

- Analysis of packet sampling in the frequency domain
  - How packet sampling impacts the spectrum of network traffic?
  - Is there a way to preserve frequencies?
  - Supported by the ECODE FP7 strep project with Alcatel-Lucent, LAAS, U. Lancaster, U. Liege, U. Louvain (Sep 2008 to Sep 2011)
    
http://www.ecode-project.eu/
Application identification from packet sizes: Learning phase

- Collect real packet traces where we know the reality of applications
- Construct density spaces for packet sizes
  - One space per packet size order (first packet of an application, second packet of an application, etc)
  - Plus and minus for the direction of the packet

\[ x: \text{size of packet of order } i \text{ of Application 1} \]
\[ x: \text{size of packet of order } i \text{ of Application 2} \]

Cluster the dots and calculate weights per cluster per application
Application identification from packet sizes: Classification phase

- **On the fly**
  - Capture a packet from an application, get its size
  - Go to the corresponding space and cluster, then calculate probability per application
  - Update a global likelihood function per application

\[
Pr(I/Result) = \frac{Pr(I) \times \prod_{k=1}^{N} Pr(i(k)/I)}{\sum_{I=1}^{A} Pr(I) \times \prod_{k=1}^{N} Pr(i(k)/I)}
\]

- Stop when either a threshold is reached
- Or a maximum number of iterations is reached
- Map the flow to the most likely application
Applications - Originality

- That remains a probabilistic method ...
  - But it works with encrypted packets and non standard ports

- Can help administrator to raise alarms and trigger further inspection of a given application flow

- Originality of the work:
  - A clustering space per packet order which allows the method to scale to further packets
  - At the expense of ignoring correlation between packet sizes (measurements show it to be low)
  - Current work focus on other compression/clustering methods

Prior work:
One joint space for all packets

The accuracy decreases after the fourth packet because of the complexity of the joint space.

Gain in accuracy when monitoring the first few packets.
Our case: One space per packet - Sequential testing

Global accuracy: True positive

Accuracy keeps improving and can reach very high value

Different confidence levels in the port number

Number of packets

Precision with alpha = 1/A
Precision with alpha = 0.5
Precision with alpha = 0.6
Precision with alpha = 0.7
Accuracy per application
False positive per application

![Graph showing false positive per application for different protocols over the number of packets. The graph includes lines for Web Trace II, https Trace II, IMAP Trace II, SSH Trace II, and SMTP Trace II. Each line represents a different protocol, with varying slopes indicating different false positive rates.]
Why? The Likelihood per application

After few packets tested, there is a clear separation in the likelihood to be WEB between applications.
The Likelihood per application

- For Web Flows
- For POP3 Flows
- For SMTP Flows

The likelihood to be POP3

Number of packets

Chadi Barakat - 12/10/2009
Time between packets adds noise

![Graph showing global accuracy over the number of packets]

- Green crosses: when using packet’s size only
- Pink squares: when using packet’s size and inter-packets time

Time between packets represents network conditions more than application behaviors.
Analysis of packet sampling in the frequency domain

Chadi BARAKAT
INRIA Sophia Antipolis, France
Planète research group

Joint work with Alfredo Grieco
Politecnico di Bari, Italy

Email: Chadi.Barakat@sophia.inria.fr
WEB: http://www.inria.fr/planete/chadi
Motivation: Packet sampling

- Packet sampling, a technique to reduce the monitoring load on routers
Packet sampling, a technique to reduce the monitoring load on routers
Motivation: Packet sampling

- Packet sampling, a technique to reduce the monitoring load on routers
Packet sampling, a technique to reduce the monitoring load on routers.
Packet sampling, a technique to reduce the monitoring load on routers.

How does the monitoring of the sampled traffic compare to the original one?
How to perform the inversion?
Motivation: Related work

- Many papers have studied the problem with stochastic tools (Duffield et al, Veitch et al, Estan et al, Diot et al, Zseby et al)
  - Packets or flows form a population
  - Sampled randomly then measured
  - Inversion aim at reducing some error function
    - Minimize mean square error
    - Maximize likelihood
    - Preserve some ranking measure
  - Inverted metrics: traffic volume, flow size distribution, heavy hitter detection, flow counting, etc
- How does packet sampling impact the spectrum of the traffic?
Outline

- Models for traffic and spectrum
- Analysis of packet sampling
- Aliasing noise and its removal by low pass filtering
- The Filter-Bank solution
- Simulation results
- Conclusions
Traffic model and spectrum

- **Traffic**: A time series of packets of different sizes $d_n$

- **Measured traffic rate**:
  - Divide time into small bins
  - Volume of bytes per bin divided by bin length $T$
  - The larger the bin the coarser the measurement

- **Targeted traffic spectrum**:
  - Spectrum of the binned traffic rate
  - Energy of different frequency components
Spectrum and sampling

- No sampling:
  - Spectrum depends on the binning interval $T$
  - Binning with time window $T$ == low pass filtering with band $0.445/T$
  - The bin defined the maximum frequency of interest
    
    All frequency oscillations less than $0.445/T$ are left

- With packet sampling:
  - Less packets
  - Different spectrum of binned traffic
  - For some bin $T$, are frequencies preserved?
  - Given sampling rate, is there any minimum $T$ to use?
Analysis: No Sampling

- Let $D(f)$ be the spectrum of the original traffic
  - Traffic discretized in tiny time slots $t_0$
  - Periodic spectrum of period $1/t_0$

- Suppose the existence of a maximum frequency $f_M$ with $0 < f_M < 1/t_0$

- An example of a real baseband
Analysis: No Sampling, With Binning

- Binning equivalent to low pass filtering

Convolution with a low pass filter of band $0.445/T$

Energy of signal of interest
Analysis: Sampling

- Traffic sampled with rate $p < 1$
- Let $D_p(f)$ be the spectrum of the sampled traffic

  - Result: A replication of $D_0(f)$ with period $p/t_0$ in the band of interest
  - Scaled down by $p$

\[ D(f) \approx p \sum D_0(f + n \cdot p/t_0) \]
Analysis: Sampling, With Binning

- By binning and scaling up by $1/p$, one can recover the signal of interest.
Aliasing for small sampling rates

- The smaller the sampling rate, the closer the replicas
  - There is a sampling rate below which they overlap

- If the binning is not coarse enough, aliasing occurs. We get a noisy signal.
Aliasing in the baseband

Baseband component of $D_p(f)/p$: (a) $p = 1$; (b) $p = 0.1$; (c) $p = 0.03$; (d) $p = 0.005$. 
Aliasing noise elimination

For a traffic of maximum frequency $f_M$ in the baseband

- Either increase the sampling rate to avoid the overlap of replicas in the band of interest
  - Always work
- Or increase the binning interval $T$
  - Will not work if $p/t_0 < f_M$

General result: Spectrum of the binned traffic is preserved upon traffic sampling if and only if

$$0.445 / T < p/t_0 - f_M$$
Determining the bin to use

- **A traffic already sampled**
  - Further downsampling possible, but not upsampling
  - No information on the maximum frequency in the baseband
  - How to know the right bin?

- **Increasing the bin size alone is not enough**
  - The energy decreases with

- **Our solution: Filter-Bank to check Traffic Variance (Energy)**
  - Take a bin size
  - Further increase the sampling rate
  - If energy (variance) quickly drops, aliasing exists
  - If energy (variance) slowly decays, the bin size is fine
Sampling rates vs bin sizes

- Over a long trace from the Japanese MAWI project
Conclusions

- A better method for classifying applications using their packet sizes
- An analysis of packet sampling in the frequency domain
  - An expression relating:
    - Sampling rate
    - Maximum frequency in the baseband
    - Minimum binning interval
  in order to avoid aliasing and sampling noise

- Future plans:
  - More applications to classify, especially P2P applications
  - Estimate the amount of noise caused by aliasing