A Data Streaming Algorithm for Estimating Entropies of OD Flows

Chuck Zhao[1] **Ashwin Lall**[2] Jim Xu[1] Mitsu Ogihara[2]
Oliver Spatscheck[3] Jia Wang[3]

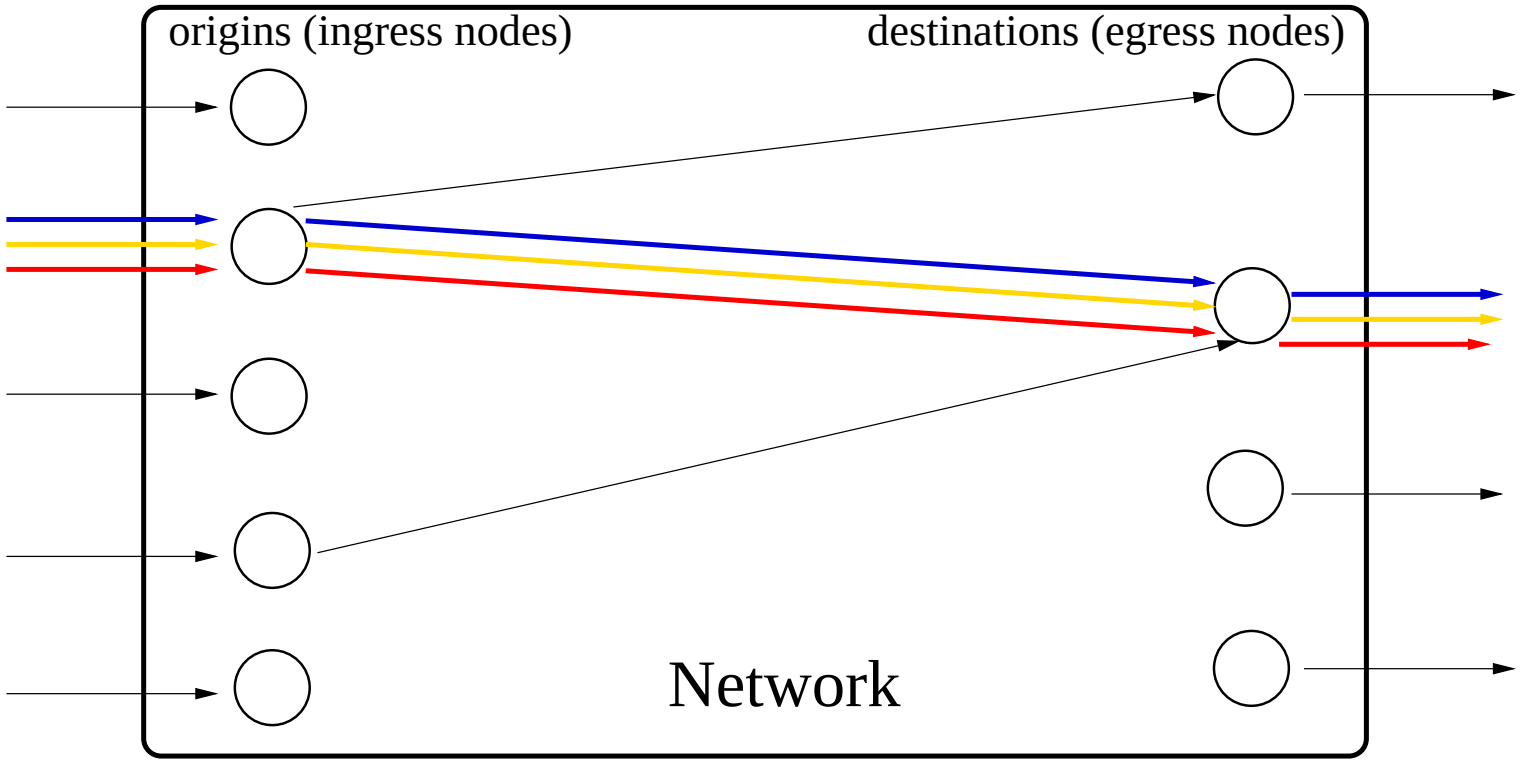
[1]Georgia Tech

[2]University of Rochester

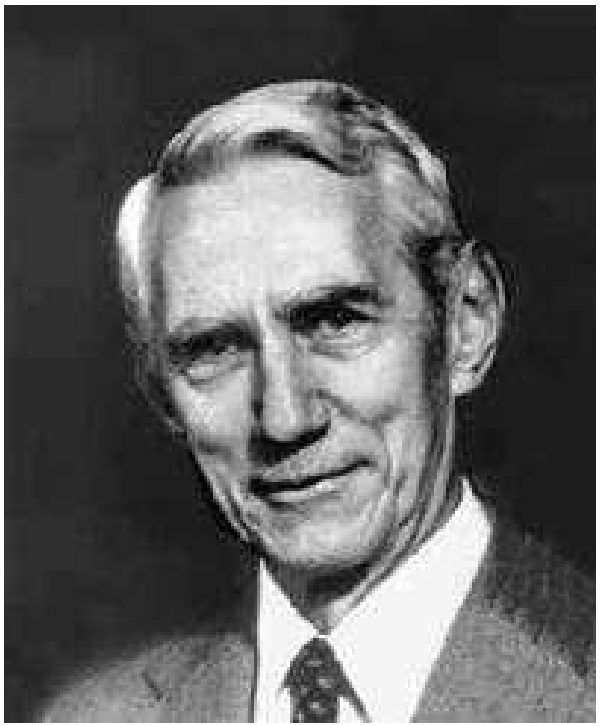
[3]AT&T Research

May 14th, 2008

Origin Destination Pairs



Entropy



Claude Shannon

Entropy: $H = -\sum_i p_i \log p_i$

$p_i =$ the fraction of traffic from the ith flow

Entropy norm: $S = \sum_i m_i \log m_i$

$m_i =$ frequency of ith flow

Motivation

- Anomaly Detection (Lakhina et al., SIGCOMM 2005)
- Traffic Clustering (Xu et al., SIGCOMM 2005)
- Redsitribution of traffic across many links
- DDoS attacks may not be detectable as simple volume changes
- Traffic Engineering

Stable Distributions

Definition

A distrbution D is called p-stable if, for any constants $a_1, \ldots, a_n$ and random variables $X, X_1, \ldots, X_n$ drawn from the distribution,

$$a_1 X_1 + \ldots + a_n X_n \quad \sim_d \quad (|a_1|^p + \ldots + |a_n|^p)^{1/p} X.$$

Paul Lévy

Properties of Stable Distributions

Properties

- Stable distributions exist for $p \in (0, 2]$.
- Examples: The Gaussian distribution is 2-stable and the Cauchy distribution is 1-stable.
- Closed forms are known only for $p = 0.5, 1, 2$.
- There are known formulas (Chamber et al.) for generating samples from each p-stable distribution.

Stable Sketch

Theorem (Indyk '06)

The pth frequency moment of a stream $\sum_i m_i^p$ *can be* (ϵ, δ)*-approximated with* $O\left(\frac{1}{\epsilon^2} \log(1/\delta) \log m\right)$ *bits of storage.*

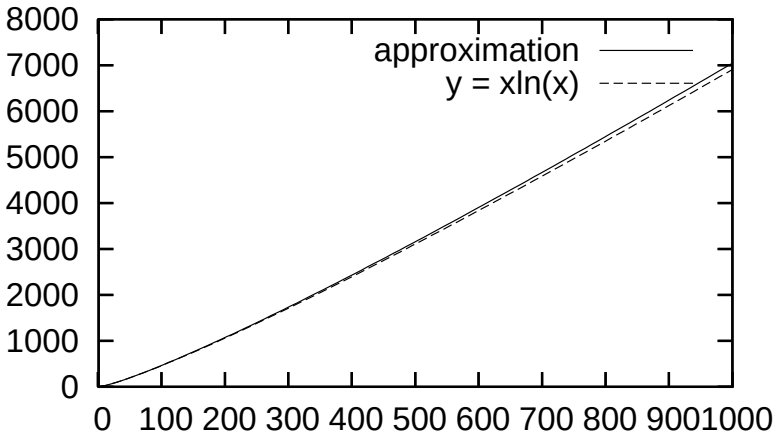
Algorithm

- For each flow i, draw a stably distributed X_i; set $V := 0$
- For each item i encountered in the stream, set $V := V + X_i$
- Now, V is distributed as $(|m_1|^p + \ldots + |m_n|^p)^{1/p} X$, where X has the p-stable distribution.
- To extract the quantity $(|m_1|^p + \ldots + |m_n|^p)^{1/p}$, we repeat this independently $O(\frac{1}{\epsilon^2} \log(1/\delta))$ times and take the median.

Approximating $x \ln x$

Theorem

For any $N > 1$, $\epsilon > 0$, there exists α, c, such that $f(x) = c(x^{1+\alpha} - x^{1-\alpha})$ approximates the entropy function $x \ln x$ for $x \in (1, N]$ within relative error bound ϵ.



Approximating $x \ln x$ with $\frac{(x^{1.05} - x^{0.95})}{2}$ on $[1, 1000]$

Approximating Entropy

$$\begin{aligned} m_1 \ln m_1 &\approx c(m_1^{1+\alpha} - m_1^{1-\alpha}) \\ &\dots \\ m_n \ln m_n &\approx c(m_n^{1+\alpha} - m_n^{1-\alpha}) \end{aligned}$$

So, $\sum m_i \ln m_i \approx c(\sum m_i^{1+\alpha} - \sum m_i^{1-\alpha})$.

In parallel, we have an elephant-detection module that handles (with high probability) all the flows of size greater than N.

Extracting OD Entropies

$$\begin{aligned} \overline{o} &= |f_1|^p + \ldots + |f_k|^p + |g_1|^p + \ldots + |g_l|^p \\ \overline{d} &= |f_1|^p + \ldots + |f_k|^p + |h_1|^p + \ldots + |h_m|^p \end{aligned}$$

$$\begin{aligned} \overline{o \ominus d} &= |g_1|^p + \ldots + |g_l|^p + |h_1|^p + \ldots + |h_m|^p \\ \overline{o \oplus d} &= |g_1|^p + \ldots + |g_l|^p + |h_1|^p + \ldots + |h_m|^p \\ &+ |2f_1|^p + \ldots + |2f_k|^p \end{aligned}$$

Hence, $\frac{\overline{o}+\overline{d}-\overline{o \ominus d}}{2}$ and $\frac{\overline{o \oplus d}-\overline{o \ominus d}}{2^p}$ give $|f_1|^p + \ldots + |f_k|^p$.

Estimating the Traffic Matrix

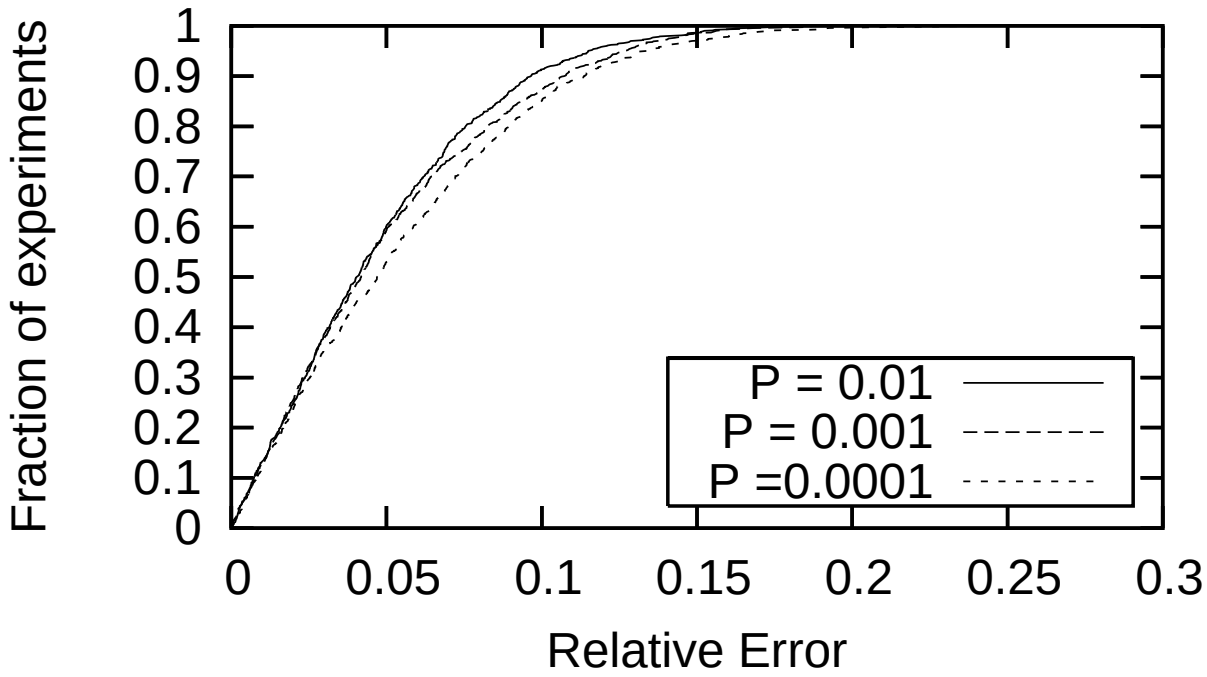
Theorem

Fix α, ϵ, and N as in the approximation of $x \ln x$. Then $(x^{1+\alpha} + x^{1-\alpha})/2$ estimates $f(x) = x$ with relative error at most 3ϵ in the range $(1, N]$.

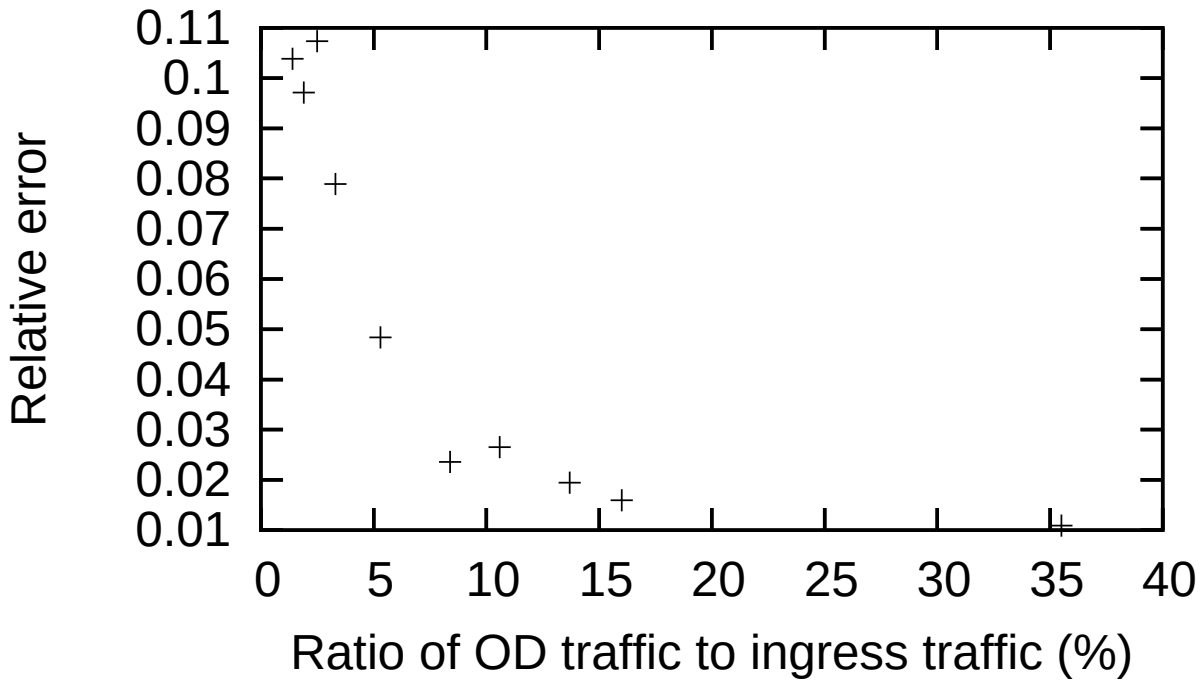
Data Collection

- Data collected at an AT&T data center with 1 Gbps access links
- Several 5-minute collected on April 25, 2007
- The traces had $\sim$400 million packets belonging to $\sim$1.8 million flows.
- All measurements were done at a single ingress router and egress traces were generated using the routing table

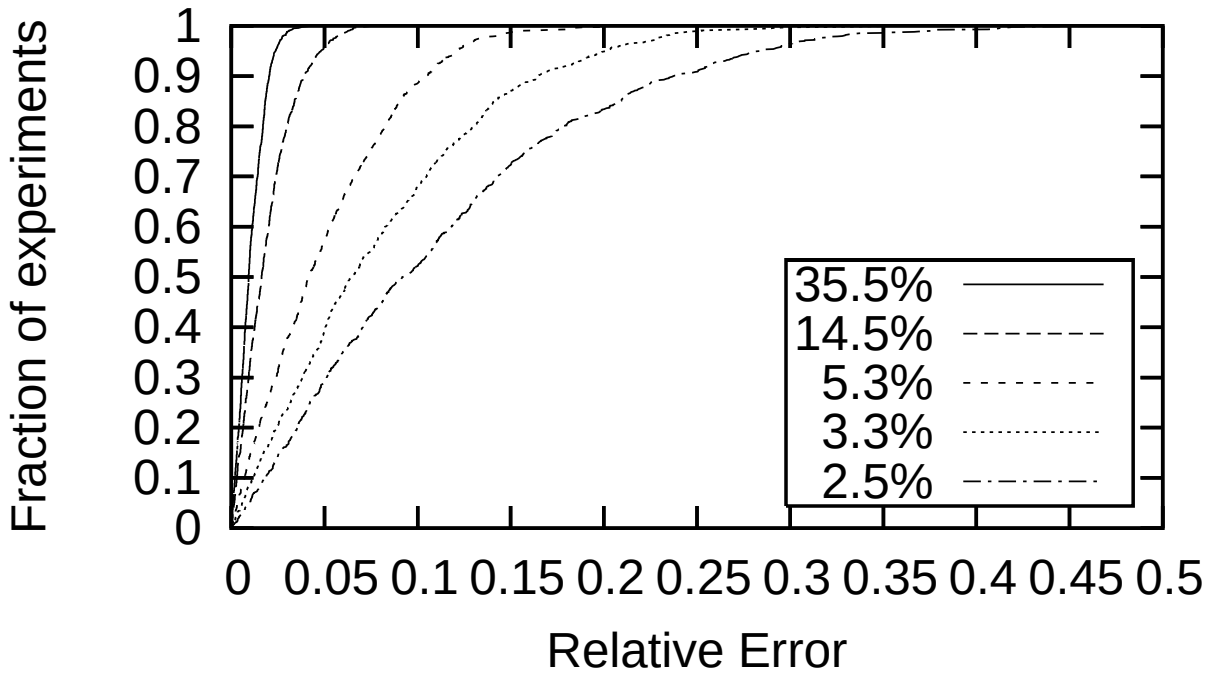
Varying Sampling Rates for Elephant Detection



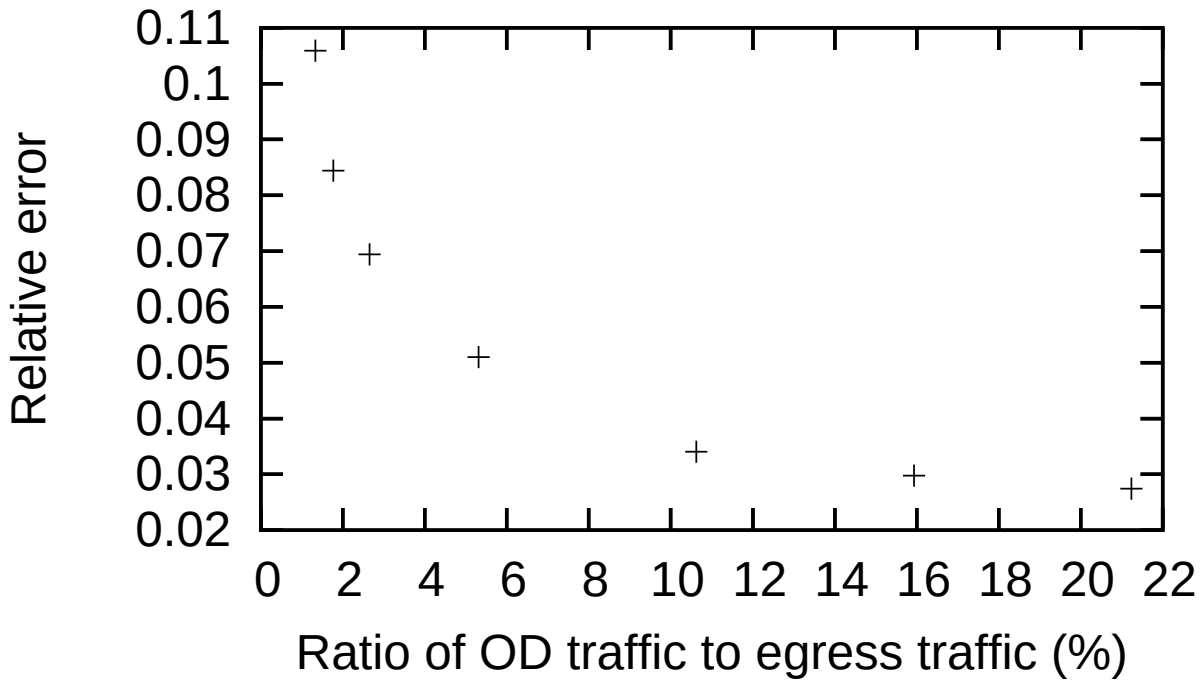
Varying Fraction of Traffic from Ingress



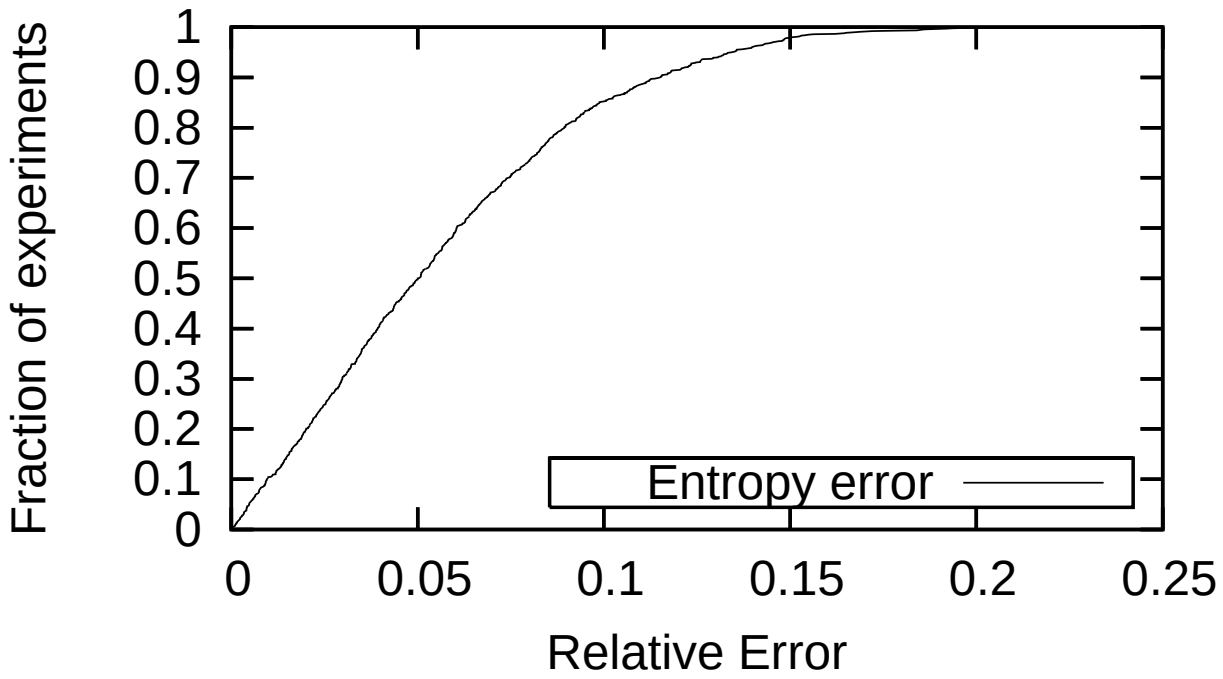
Varying Fraction of Traffic from Ingress



Varying Fraction of Traffic from Egress



Error Distribution for Actual Entropy



Conclusion

- We propose extending the study of entropy to origin-destination flows as a new tool for network tomography.
- We present an algorithm to solve this problem in practice, with low relative error and reasonable resource usage.
- We introduced a new type of distributed streaming problem: estimating statistics of origin-destination flows.

Thanks for your attention!