Data Needs for Sampling the Internet to Measure Performance

Juana Sanchez

UCLA Statistics

In this talk, I will give a brief survey of the work that statisticians are doing to try to model the Internet with statistical models.

Objective of my interest in Internet data analysis: introduce undergraduate/graduate students in Statistics courses to the field and motivate them to propose ideas and solutions.

ISMA Data Catalog 2004 Workshop. UCSD, Thursday, June 3rd 2004

Outline

1. Probabilistic modeling
2. Single node data analysis
3. Network tomography
4. Network Topology Identification
5. Sampling
6. Other
7. Conclusions

1. Probabilistic modeling

- Assume a probability model for the process, e.g., packet counts follow a mixture of Poisson distributions with parameters $\lambda_i \qquad i = 1, ..., k$. Or bytes counts follow an infinite source Poisson model...

- Use a random sample to estimate the parameter λ

- Attach a standard error to the estimate, express degree of confidence in the estimate.

- Properties of the estimators usually large sample properties.

- Many attempts to model the Internet until now are attached to known probability models, no matter how complex the process. Also attached to independence assumptions to large extent.

2. Single node (link) data analysis

- SAMSI (Statistical And Applied Mathematical Sciences Institute) Program: Network Modeling for the Internet. 2002-2004. Internet Statistics and Research Consortium (forthcoming)
  - ⋆ Probabilists (heavy traffic queueing theory and fluid models), Internet measurers, statisticians
  - ⋆ Statistical characterization of traces. Long range dependence property and scale invariance -Estimate hurst parameter. What causes the burstiness of traces? Synthetic and real traces. Bytes and packet counts. Effect of time scales. Which traces are similar?
  - ⋆ Wavelet spectrum of byte counts $\neq$ spectrum of packet counts. SIZER-helps see how wavelet features correspond to trace features. For example, does a burst correspond to a bump in the spectrum?
  - ⋆ Experiments: see if changing parameters of a synthetic network changes the burstiness..

  - ⋆ Study trace driven queues; effects of different utilization or buffer size scales on packet loss and queue length. Trace is the customers arriving at rate r(t). Many problems with assumptions. Traces that look similar under various statistical measures (such as the Hurst index) can exhibit rather different behavior under queueing simulation.

- Streaming data graphics (Wegman, E et al.2003).

- Long tradition of research in this direction: Taqqu, Willinger, Vexson,etc...

- CAIDA–Broido et al... New traces show signs of Poisson assumption. Are we back to old queueing models for networks?

2.General Network Tomography Models

- Objective: Estimating source-destination traffic intensities from link data (i.e., counts)." (Vardi, 1996, JASA 91 (433),pp 365-377

- Inference of the internal link delay distribution through multicast end-to-end measurements.

- Estimate packet loss.

- There are many possible combinations of internal link delays. The point is to estimate the most likely combination.

- Pseudo-Maximum Likelihood estimation of the intermediate paths

- OD matrix inference of counts through link-based counts

- Intricate details regarding network transportation are ignored.

- Bin Yu and associates at Berkeley (Sprint Europe to compare with AT&T data sets, Lucent Technologies network-4 nodes). Use a pseudo-likelihood approach (likelihood for smaller subparts,ignoring dependence between the subparts). The multicast tree is broken into parts.

  - ⋆ Fixed routing matrix unknown $Y = AX$. Y and A are observed. X is the unknown. A is the routing parameters.
  - ⋆ Normal models with variances function of the mean (to mimic Poisson..).

3.- Network Topology Identification

- Characterizing the structure of the network
- The network structure determines the delays. So, from the final delays, they do agglomerative cluster analysis and find the link tree-structure in the middle.
- Object similarities rather than object features determine the clusters.
- Probing non-TCP traffic in small university network.
- Rui Castro and R.Nowack, Rice University. IEEE Transactions, Coates.

Sampling

- Provisioning of information about a specific characteristic of the parent population at a lower cost than a full census would demand. Could use filtering (mask/match, hash-based), or sampling algorithms such as: Systematic sampling, random sampling, probabilistic sampling, etc.

- Sampling already used in routers. But it is important to infer about the unsampled flows using what we know about the sampled ones.

- Crucial to determine the needed type of information and the desired degree of accuracy in advance. E.g., What kind of metric? number of packets, packet size distribution?

- Internet Engineering Task Force (Duffield, N.G. et al.-AT&T Labs), K.C. Claffy (ISMA)

Other

- Computer Intrusion Detection-Marchette et al.

- Visualization tools for streaming data from internet packet headers. Evolutionary graphics that discard data as it is analyzed because it can not be stored -Wegman et al.

- Kolaczyk et al (Boston University): Principal components analyisis of complete set of OD flow time series from Sprint-Europe and Abilene. Find small dimensionality. Suggest decomposition of series into common periodic trends, short-lived bursts and noise.

- Other work at CAIDA

Conclusions

- Statisticians have not studied an awful lot of “whole network” data sets.
- Whole estimation of network-wide characteristics done usually in collaboration with engineers and computer scientists.
- more network-wide data sets similar to those already studied would help validate existing probability models.
- Perhaps more joint work of statisticians with the engineers/computer scientists may lead to more useful probability models.
- Good way to start, maybe: if we repeat the same models with wider network data, what would happen?
- Having the priority questions clear helps determine data needs and useful sampling.

Thank you for going through this document.
You may proceed reading the report on PPower4, the software tool which was used to prepare this demo. Please note that the report only describes the initial development, not the current state.

Hit Esc to leave FullScreen mode.
Select the appropriate entry in the View menu to return to this mode.

Back to the page displayed before.