

# Statistical Analysis of Network Data

David Marchette, Karen Kafadar, Ginger Davis, Charles Wright

July 24, 2007

# Meeting Notes

## Logistics

Next meeting: JSM Sunday afternoon

## Immediate Tasks List

- Dave
  1. Talk to Wendy about potential funding.
  2. Get data from Ed.
  3. Look for BAAs.
  4. Literature review
  5. Get more cool t-shirts.
  
- Ginger
  1. Write undergraduate capstone project description.
  2. Arrange next meeting, which will be at JSM.
  3. Organize notes/tasks.
  4. Look for BAAs.
  5. Literature review
  6. Harrass those who do not do their tasks.
  
- Karen
  1. Get website up for literature review papers.
  2. Look for BAAs.
  3. Literature review
  4. Bring cat pictures to JSM. We did not see any pictures!
  
- Charles
  1. Put VoIP slides and paper on website.

2. Get Lawrence Berkeley Lab packet flow data.
3. Find software to process GMU data into connections (SNORT?)
4. Look for BAAs.
5. Literature review
6. Write thesis.

## **Overview of Charles' Voice over IP (VoIP) project**

Developed a classifier for predicting spoken languages whose signals were compressed.

### Objectives

1. language determination
2. speaker identification

### Hypotheses

1. Packet size is linked to content.
2. The pattern of transfers can tell you which website you are viewing.

Dave's potential contribution using graphs: You can use languages to improve the groups; you can use groups to improve the languages. Basically, the graph (clusters of countries) can help improve prediction.

### Packet flow analysis objectives

1. How little information can we get away with?
2. How far can we push the information we have for learning?

## **Data Terminology**

Packets – Flows – Connections – Activities (Network Sessions) – User Session

Flow - Either source or destination activity between a source/destination

pair.

Connection - Collection of source and destination flows.

Activity - Collection of connections.

User session - Collection of activities.

Data Problem: We won't see the url-only the IP addresses.

## Grant Proposal

### Research Goals

- Analyze internet activity for the purposes of
  - anomaly detection
    - \* attacks
    - \* unusual behavior
    - \* masquerades
      - packets
      - sessions
      - activities
      - See Early/Brodley (ACSAC, 2005) and Schonlau et al. (Statistical Science, 2000)
    - \* network interruptions
  - user profiling
  - workload management: managing resources for particular purposes
  - marketing strategies: what leads people to view certain sites
  - application verification
- Subgoals
  - Develop strategies for managing huge data sets.
    - \* We need a scan statistic for data chunks.
    - \* similar idea to Charles' VoIP 9 sorted pictures. Can we cluster the data once a similar image is constructed?

- Dimension reduction
- Visualization of data sets
- Privacy issues
  - \* Which features of the data leak sensitive information?
  - \* How can we obfuscate the data in such a way that performance is not degraded?

## Data Sources

1. 135,000 records from Dave
2. GMU dataset (Ed Wegman)
3. Lawrence Berkeley Lab packet flow data

## Research Agenda

1. Transform packet level data to connection data
  - What information would be useful?
  - Which do we need—flow or connections for which purposes?
2. Transform connection data into activities
  - Construction
  - Analysis
  - Characterization
3. Visualization
  - Packet – Flow – Connection – Activity
4. Privacy issues
  - VoIP
  - Obfuscation
  - Anonymization: How much can we detect; therefore, how much do we need to hide?
  - Can web pages or behavior (e.g. purchasing) be revealed unintentionally? This would imply a need for better obfuscation.

## Research Activities

### Year 1: Setup

1. Process GMU data into connections. Charles will look into which software to use (SNORT?). This will not be trivial due to the size and storage issues. Ginger's undergraduate students will be working on this project starting in August.
2. Extract potentially additional useful information from packet data into current connections output. We will have 2 flows (from client to server and server to client) which we will collect and call a connection. Ginger and her students will be working on this after the processing in the previous step is progressing.
3. Transform connection data into activities. Karen will be working on this.
4. Perform a historical search of Internet attacks and outages (power, worms, viruses, etc.). Karen and her student will be working on this. CAUTION: There is potential bias here (in reporting).
5. Visualization. Dave will be working on this.
6. Look into potential vehicles (such as SILK) for the implementation of analysis tools.

### Year 2: Analysis

- Packet data analysis
  - Develop (time series) models for traffic flow.
  - Simulate from these models and compare to real data.
  - Statistics such as histograms, variance, reduced histogram, time series model parameters, etc.
  - Responsibility: Ginger
- Activity analysis
  - Visualization (Dave)

- Look for patterns (Charles)
- Time series models (Ginger)
- Data analysis (Karen)
- Identification of information leaks
  - For each data set, we will need to:
    - \* look for patterns
    - \* perform information extraction (of features)
    - \* see what kinds of problems can be addressed

### **Year 3: Implementation**

1. Evaluation of tools
2. Ensure availability of programs (website) and algorithms (papers).
3. Provide test data sets.
4. SILK: Potential implementation vehicle which already contains a large number of tools for network data that we may possibly be able to use.

### **Resources required**

1. host for website
2. data storage
3. 2 graduate students
4. postdoc
5. full-time programmer (research scientist) for years 2-3
6. clinic project
7. salaries
8. travel

## Potential Funding Sources

1. DHS
2. NSF
3. ARO
4. NSA
5. DARPA
6. MURI

## Survey Paper: Statistical Issues in Network Security

Purpose: To introduce statisticians to the problem and provide guidance for research efforts in this area.

## Literature Search

- Responsibility: everyone
- Journals
  - CITESEER as a source
  - IEEE-PAMI
  - IEEE-Security and Privacy
  - CERT
  - various conference proceedings
- Karen: Set up a website for literature review. Send papers to Dana Franklin (dana.franklin@cudenver.edu).

Outline

## 1. Introduction

- The traditional approach is to identify a specific problem and try to detect it (such as Denial-of-Service attacks), signatures (virus detectors, attacks), and anomaly detection.
- Responsibility: Karen

## 2. Domain-driven data aggregation

- levels and uses of data (packets, flows, connections, activities, superactivities)
- Responsibility: Dave
- This also comes up at NASA, JPL.

## 3. Packets

- e.g. VoIP
- Responsibility: Charles
- Can you grab something from your thesis?

## 4. Flows and connections

- examples of various types of attacks
- how people use protocols to launch attacks
- Responsibility: Dave

## 5. Activities and User Sessions (Ginger and Karen)

## 6. Research needs

- which statistical methods that have been used
- past and current applications of statistics to the problem (statistical quality control, data mining, time series modeling)
- Responsibility: Ginger

## 7. Research agenda

- See Grants section.

- Responsibility: Karen

## 8. Conclusion (Karen)

## 9. Appendix

- Available resources
  - websites
  - tools
- Responsibility: Dave

# Data Sources

LBL packet trace data: <http://www.icir.org/enterprise-tracing/index.html>

Important note: Although there's no license agreement on the page, or even a disclaimer, I've been made aware that the publishers of this dataset are NOT OK with it being used for research into flaws in the anonymization system that was used to sanitize the traces. As far as I know, it's fine to use it for anything else.

The toolkit (SiLK) for processing packets into NetFlows (or just "flows"), is here: <http://tools.netsa.cert.org/silk/>

# References