Mining Network Traffic Data

Ljiljana Trajković
ljilja@cs.sfu.ca

Communication Networks Laboratory
http://www.ensc.sfu.ca/cnl
School of Engineering Science
Simon Fraser University, Vancouver, British Columbia
Canada
Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection, statistical analysis, clustering tools, prediction analysis
- Case studies:
  - wireless network: Telus Mobility
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions and references
Roadmap

Simon Fraser University, Burnaby Campus
Introduction

Communication Networks Laboratory
http://www.ensc.sfu.ca/~ljilja/cnl

Research interests:
- modeling and analysis of computer networks
- characterization and modeling of network traffic
- performance analysis of communication networks
- simulation of protocols and network control algorithms
- intelligent control of communication systems
Communication Networks Laboratory

People:
http://www.ensc.sfu.ca/~ljljila/cnl/alumni.html

- Ph.D.
- M.Sc.
- M.Eng.
- B.A.Sc.

Co-Op students:
L'Institut des Sciences de l'Ingénieur de Toulon et du Var, ISITV
Communication Networks Laboratory

Collaborations:

- C. K. Tse, The Hong Kong Polytechnic University, Hong Kong
- M. di Bernardo, University of Naples Federico II, Naples, Italy
- K. Okumura, Kyoto University, Japan
- G. Petrovic, Faculty of Electrical Engineering, University of Belgrade
- K. Mayaram, Oregon State University, USA
- W. Mathis, University of Hanover, Germany
Communication Networks Laboratory

Projects:
- Data Analysis in Wireless and Wireline Networks
- Intelligent Control of Communication Networks
- Simulation of Communication Networks
- OPNET-specific projects
Communication Networks Laboratory

Data Analysis in Wireless and Wireline Networks:

- Analysis of Internet topologies: a historical view
- Spectral analysis of the Internet topology
- Data mining on billing traces of wireless network
- Modeling and characterization of traffic in public safety wireless networks
- Adapting ad hoc network concepts to land mobile radio systems
- Wavelet-based analysis of long-range dependent video traces
- TCP session analysis and modeling of hybrid satellite-terrestrial Internet traffic
- Measurement and analysis of hybrid satellite-terrestrial Internet traffic
- Understanding network customers' behavior from billing traces
- Using AutoClass for exploring demographic structure of Internet users
Communication Networks Laboratory

Intelligent Control of Communication Networks:

- Stability study of the TCP-RED system using detrended fluctuation analysis,
- Stability analysis of RED gateway with multiple TCP Reno connections
- Discontinuity-induced bifurcations in TCP/RED communication algorithms
- Modeling TCP with active queue management schemes
- Characterization of a simple communication network using Legendre transform
- Delay and throughput differentiation mechanism for non-elevated services
- Simulation of loss patterns in video transfers over UDP and TCP
- Analysis and simulation of wireless data network traffic
Communication Networks Laboratory

Simulation of Communication Networks:

- Integrating ns-BGP with the ns-2.33 network simulator
- BGP route flap damping algorithms
- BGP with an adaptive minimal route advertisement interval (MRAI)
- Implementation of BGP in a network simulator
- Improving the performance of the Gnutella network
- Selective-TCP for wired/wireless networks
- TCP over wireless networks
- Modeling and performance evaluation of a General Packet Radio Services (GPRS) network using OPNET
- Traffic engineering prioritized IP packets over Multi-Protocol Label Switching (MPLS) network
- Enhancements and performance evaluation of wireless local area networks
- Route optimization of mobile IP over IPv4
Communication Networks Laboratory

OPNET-specific projects:
http://www.ensc.sfu.ca/~ljilja/opnet/

- Streaming video content over IEEE 802.16/WiMAX broadband access
- Performance evaluation of TCP Tahoe, Reno, Reno with SACK, and NewReno
- OPNET model of TCP with adaptive delay and loss response for broadband GEO satellite networks
- M-TCP+: using disconnection feedback to improve performance of TCP in wired/wireless networks
- Performance evaluation of M-TCP over wireless links with periodic disconnections
- General Packet Radio Service OPNET model
- Effect of cell update on performance of General Packet Radio Service
- OPNET implementation of the Megaco/H.248 Protocol
- Compressed Real-Time Transport Protocol (cRTP)
- Enhancements and performance evaluation of wireless local area networks
- Cellular Digital Packet Data (CDPD) MAC layer model
Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection, statistical analysis, clustering tools, prediction analysis
- Case studies:
  - wireless network: Telus Mobility
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions and references
Network traffic measurements

- **Traffic measurements** in operational networks help:
  - understand traffic characteristics in deployed networks
  - develop traffic models
  - evaluate performance of protocols and applications

- **Traffic analysis**:
  - provides information about the user behavior patterns
  - enables network operators to understand the behavior of network users

- **Traffic prediction**: important to assess future network capacity requirements and to plan future network developments
Self-similarity

- Self-similarity implies a “fractal-like” behavior: data on various time scales have similar patterns.
- A wide-sense stationary process $X(n)$ is called (exactly second order) self-similar if its autocorrelation function satisfies:
  - $r^{(m)}(k) = r(k), k \geq 0, m = 1, 2, ..., n,$
    where $m$ is the level of aggregation.
- Implications:
  - no natural length of bursts
  - bursts exist across many time scales
  - traffic does not become “smoother” when aggregated (unlike Poisson traffic)
Long-range dependence: properties

- High variability:
  - when the sample size increases, variance of the sample mean decays more slowly than expected
- Burstiness over a range of timescales:
  - long runs of large values followed by long runs of small values, repeated in aperiodic patterns

![fGn trace](image)
Estimation of H

Various estimators:
- variance-time plots
- R/S plots
- periodograms
- wavelets

Their performance often depends on the characteristics of the data trace under analysis
Clustering analysis

- Clustering analysis groups or segments a collection of objects into subsets or clusters based on similarity.
- An object can be described by a set of measurements or by its relations to other objects.
- Clustering algorithms can be employed to analyze network user behaviors.
- Network users are classified into clusters, according to the similarity of their behavior patterns.
- With user clusters, traffic prediction is reduced to predicting and aggregating users' traffic from few clusters.
Clustering algorithms

- Two approaches:
  - partitioning clustering (k-means)
  - hierarchical clustering

- Clustering tools:
  - k-means algorithm
  - AutoClass tool
Traffic prediction: ARIMA model

- Auto-Regressive Integrated Moving Average (ARIMA) model:
  - general model for forecasting time series
  - past values: AutoRegressive (AR) structure
  - past random fluctuant effect: Moving Average (MA) process
- ARIMA model explicitly includes differencing
- ARIMA \((p, d, q)\):
  - autoregressive parameter: \(p\)
  - number of differencing passes: \(d\)
  - moving average parameter: \(q\)
Traffic prediction: SARIMA model

- Seasonal ARIMA is a variation of the ARIMA model
- Seasonal ARIMA (SARIMA) model:
  \[(p,d,q) \times (P,D,Q)_S\]
  - captures seasonal pattern
- SARIMA additional model parameters:
  - seasonal period parameter: S
  - seasonal autoregressive parameter: P
  - number of seasonal differencing passes: D
  - seasonal moving average parameter: Q
SARIMA models: selection criteria

- Order \((p,d,q)\) is selected based on:
  - time series plot of traffic data
  - autocorrelation and partial autocorrelation functions

- Validity of parameter selection:
  - Akaike’s information criteria
  - Bayesian information criterion
Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection, statistical analysis, clustering tools, prediction analysis
- Case study:
  - wireless network: Telus Mobility
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions and references
Case study: E-Comm network

- E-Comm network: an operational trunked radio system serving as a regional emergency communication system
- The E-Comm network is capable of both voice and data transmissions
- Voice traffic accounts for over 99% of network traffic
- A group call is a standard call made in a trunked radio system
- More than 85% of calls are group calls
- A distributed event log database records every event occurring in the network: call establishment, channel assignment, call drop, and emergency call
E-Comm network

E-Comm’s Wide-Area Radio System: Police Customers

- Police departments using E-Comm’s Wide-Area Radio System
- GVTAPS not illustrated
E-Comm network
E-Comm network

E-Comm's Wide-Area Radio System: Ambulance Service

The BC Ambulance Service uses the E-Comm Radio System throughout the GVRD
E-Comm network architecture

Diagram:
- Users
- Transmitters/Repeaters
- PSTN
- PBX
- Dispatch console
- Network switch
- Database server
- Data gateway
- Management console
- Other EDACS systems
- Vancouver
- Burnaby
Traffic data

- 2001 data set:
  - 2 days of traffic data
    - 2001-11-01 to 2001-11-02 (110,348 calls)
- 2002 data set:
  - 28 days of continuous traffic data
    - 2002-02-10 to 2002-03-09 (1,916,943 calls)
- 2003 data set:
  - 92 days of continuous traffic data
    - 2003-03-01 to 2003-05-31 (8,756,930 calls)
Traffic data

- Records of network events:
  - established, queued, and dropped calls in the Vancouver cell
- Traffic data span periods during:

<table>
<thead>
<tr>
<th>Trace (dataset)</th>
<th>Time span</th>
<th>No. of established calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>November 1–2, 2001</td>
<td>110,348</td>
</tr>
<tr>
<td>2002</td>
<td>March 1–7, 2002</td>
<td>370,510</td>
</tr>
</tbody>
</table>
Observations

- Presence of daily cycles:
  - minimum utilization: ~2 PM
  - maximum utilization: 9 PM to 3 AM

- 2002 sample data:
  - cell 5 is the busiest
  - others seldom reach their capacities

- 2003 sample data:
  - several cells (2, 4, 7, and 9) have all channels occupied during busy hours
Performance analysis

- Modeling and Performance Analysis of Public Safety Wireless Networks
- WarnSim: a simulator for public safety wireless networks (PSWN)
- Traffic data analysis
- Traffic modeling
- Simulation and prediction

WarnSim overview

- Simulators such as OPNET, ns-2, and JSim are designed for packet-switched networks
- WarnSim is a simulator developed for circuit-switched networks, such as PSWN
- WarnSim:
  - publicly available simulator:
    http://www.ensc.sfu.ca/~ljilja/cnl/projects/warnsim
  - effective, flexible, and easy to use
  - developed using Microsoft Visual C# .NET
  - operates on Windows platforms
Call arrival rate in 2002 and 2003: cyclic patterns

- the busiest hour is around midnight
- the busiest day is Thursday
- useful for scheduling periodical maintenance tasks
Modeling and characterization of traffic

- We analyzed voice traffic from a public safety wireless network in Vancouver, BC
  - call inter-arrival and call holding times during five busy hours from each year (2001, 2002, 2003)
- Statistical distribution and the autocorrelation function of the traffic traces:
  - Kolmogorov-Smirnov goodness-of-fit test
  - autocorrelation functions
  - wavelet-based estimation of the Hurst parameter

Erlang traffic models

Erlang B

\[ P_B = \frac{A^N}{N!} \sum_{x=0}^{N} \frac{A^x}{x!} \]

Erlang C

\[ P_C = \frac{A^N}{N!} \frac{N}{N - A} \sum_{x=0}^{N-1} \frac{A^x}{x!} + \frac{A^N}{N!} \frac{N}{N - A} \]

- \( P_B \): probability of rejecting a call
- \( P_C \): probability of delaying a call
- \( N \): number of channels/lines
- \( A \): total traffic volume
Hourly traces

- Call holding and call inter-arrival times from the five busiest hours in each dataset (2001, 2002, and 2003)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02.11.2001</td>
<td>3,718</td>
<td>01.03.2002</td>
<td>4,436</td>
<td>26.03.2003</td>
<td>4,919</td>
</tr>
<tr>
<td>15:00–16:00</td>
<td></td>
<td>04:00–05:00</td>
<td></td>
<td>22:00–23:00</td>
<td></td>
</tr>
<tr>
<td>01.11.2001</td>
<td>3,707</td>
<td>01.03.2002</td>
<td>4,314</td>
<td>25.03.2003</td>
<td>4,249</td>
</tr>
<tr>
<td>00:00–01:00</td>
<td></td>
<td>22:00–23:00</td>
<td></td>
<td>23:00–24:00</td>
<td></td>
</tr>
<tr>
<td>02.11.2001</td>
<td>3,492</td>
<td>01.03.2002</td>
<td>4,179</td>
<td>26.03.2003</td>
<td>4,222</td>
</tr>
<tr>
<td>16:00–17:00</td>
<td></td>
<td>23:00–24:00</td>
<td></td>
<td>23:00–24:00</td>
<td></td>
</tr>
<tr>
<td>01.11.2001</td>
<td>3,312</td>
<td>01.03.2002</td>
<td>3,971</td>
<td>29.03.2003</td>
<td>4,150</td>
</tr>
<tr>
<td>19:00–20:00</td>
<td></td>
<td>00:00–01:00</td>
<td></td>
<td>02:00–03:00</td>
<td></td>
</tr>
<tr>
<td>02.11.2001</td>
<td>3,227</td>
<td>02.03.2002</td>
<td>3,939</td>
<td>29.03.2003</td>
<td>4,097</td>
</tr>
<tr>
<td>20:00–21:00</td>
<td></td>
<td>00:00–01:00</td>
<td></td>
<td>01:00–02:00</td>
<td></td>
</tr>
</tbody>
</table>
Example: March 26, 2003

![Graph showing call holding times and call inter-arrival time](image)

- **Call holding times (s)**
- **Time (hh:mm:ss)**

- **Call inter-arrival time**

---

April 27, 2009  
University of California, Irvine  
38
Statistical distributions

- Fourteen candidate distributions:
  - exponential, Weibull, gamma, normal, lognormal, logistic, log-logistic, Nakagami, Rayleigh, Rician, t-location scale, Birnbaum-Saunders, extreme value, inverse Gaussian

- Parameters of the distributions: calculated by performing maximum likelihood estimation

- Best fitting distributions are determined by:
  - visual inspection of the distribution of the trace and the candidate distributions
  - Kolmogorov-Smirnov test of potential candidates
Call inter-arrival times: pdf candidates

![Probability density plot showing different models for call inter-arrival times](image)

- Traffic data
- Exponential model
- Lognormal model
- Weibull model
- Gamma model
- Rayleigh model
- Normal model
Call inter-arrival times: K-S test results (2003 data)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Parameter</th>
<th>26.03.2003, 22:00–23:00</th>
<th>25.03.2003, 23:00–24:00</th>
<th>26.03.2003, 23:00–24:00</th>
<th>29.03.2003, 02:00–03:00</th>
<th>29.03.2003, 01:00–02:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>h</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.0027</td>
<td>0.0469</td>
<td>0.4049</td>
<td>0.0316</td>
<td>0.1101</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>0.0283</td>
<td>0.0214</td>
<td>0.0137</td>
<td>0.0205</td>
<td>0.0185</td>
</tr>
<tr>
<td>Weibull</td>
<td>h</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.4885</td>
<td>0.4662</td>
<td>0.2065</td>
<td>0.286</td>
<td>0.2337</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>0.0130</td>
<td>0.0133</td>
<td>0.0164</td>
<td>0.014</td>
<td>0.0159</td>
</tr>
<tr>
<td>Gamma</td>
<td>h</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>0.3956</td>
<td>0.3458</td>
<td>0.127</td>
<td>0.145</td>
<td>0.1672</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>0.0139</td>
<td>0.0146</td>
<td>0.0181</td>
<td>0.0163</td>
<td>0.0171</td>
</tr>
<tr>
<td>Lognormal</td>
<td>h</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>1.015E-20</td>
<td>4.717E-15</td>
<td>2.97E-16</td>
<td>3.267E-23</td>
<td>4.851E-21</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>0.0689</td>
<td>0.0629</td>
<td>0.0657</td>
<td>0.0795</td>
<td>0.0761</td>
</tr>
</tbody>
</table>
Call inter-arrival times: estimates of $H$

- Traces pass the test for time constancy of $a$: estimates of $H$ are reliable

<table>
<thead>
<tr>
<th></th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Day/hour</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02.11.2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00–16:00</td>
<td>0.907</td>
<td>0.679</td>
<td>0.788</td>
</tr>
<tr>
<td>01.11.2001</td>
<td>0.802</td>
<td>0.757</td>
<td>0.832</td>
</tr>
<tr>
<td>00:00–01:00</td>
<td>0.770</td>
<td>0.780</td>
<td>0.699</td>
</tr>
<tr>
<td>02.11.2001</td>
<td>0.774</td>
<td>0.741</td>
<td>0.696</td>
</tr>
<tr>
<td>16:00–17:00</td>
<td>0.663</td>
<td>0.747</td>
<td>0.705</td>
</tr>
<tr>
<td>01.11.2001</td>
<td>0.774</td>
<td>0.741</td>
<td>0.696</td>
</tr>
<tr>
<td>19:00–20:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02.11.2001</td>
<td>0.663</td>
<td>0.747</td>
<td>0.705</td>
</tr>
<tr>
<td>20:00–21:00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Call holding times: pdf candidates

![Diagram showing probability density functions for call holding times with various models: Traffic data, Lognormal model, Gamma model, Weibull model, Exponential model, Normal model, and Rayleigh model.](chart.png)
Call holding times: estimates of $H$

- All (except one) traces pass the test for constancy of a
- only one unreliable estimate (*): consistent value

<table>
<thead>
<tr>
<th>Day/hour</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H$</td>
<td>$H$</td>
<td>$H$</td>
</tr>
<tr>
<td>02.11.2001</td>
<td>0.493</td>
<td>0.490</td>
<td>0.483</td>
</tr>
<tr>
<td>15:00–16:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>01.11.2001</td>
<td>0.471</td>
<td>0.460</td>
<td>0.483</td>
</tr>
<tr>
<td>00:00–01:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02.11.2001</td>
<td>0.462</td>
<td>0.489</td>
<td>0.463</td>
</tr>
<tr>
<td>16:00–17:00</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>01.11.2001</td>
<td>0.467</td>
<td>0.508</td>
<td>0.526</td>
</tr>
<tr>
<td>19:00–20:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02.11.2001</td>
<td>0.479</td>
<td>0.503</td>
<td>0.466</td>
</tr>
<tr>
<td>20:00–21:00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Call inter-arrival and call holding times

<table>
<thead>
<tr>
<th>Date/Time</th>
<th>Avg. (s)</th>
<th>Date/Time</th>
<th>Avg. (s)</th>
<th>Date/Time</th>
<th>Avg. (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>02.11.2001 15:00–16:00</td>
<td>0.97</td>
<td>01.03.2002 04:00–05:00</td>
<td>0.81</td>
<td>26.03.2003 22:00–23:00</td>
<td>0.73</td>
</tr>
<tr>
<td>01.11.2001 00:00–01:00</td>
<td>3.95</td>
<td>01.03.2002 22:00–23:00</td>
<td>3.84</td>
<td>25.03.2003 23:00–24:00</td>
<td>4.12</td>
</tr>
<tr>
<td>02.11.2001 16:00–17:00</td>
<td>1.03</td>
<td>01.03.2002 23:00–24:00</td>
<td>3.88</td>
<td>26.03.2003 23:00–24:00</td>
<td>4.04</td>
</tr>
<tr>
<td>01.11.2001 19:00–20:00</td>
<td>1.09</td>
<td>01.03.2002 00:00–01:00</td>
<td>3.95</td>
<td>29.03.2003 02:00–03:00</td>
<td>4.14</td>
</tr>
<tr>
<td>02.11.2001 20:00–21:00</td>
<td>1.12</td>
<td>02.03.2002 00:00–01:00</td>
<td>4.06</td>
<td>29.03.2003 01:00–02:00</td>
<td>4.25</td>
</tr>
</tbody>
</table>

**Avg. call inter-arrival times:** 1.08 s (2001), 0.86 s (2002), 0.84 s (2003)

## Busy hour: best fitting distributions

<table>
<thead>
<tr>
<th>Busy hour</th>
<th>Distribution</th>
<th>Call inter-arrival times</th>
<th>Call holding times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Call holding times</td>
<td></td>
<td>Call holding times</td>
</tr>
<tr>
<td></td>
<td>Weibull</td>
<td>Gamma</td>
<td>Lognormal</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>02.11.2001 15:00–16:00</td>
<td>0.9785</td>
<td>1.1075</td>
<td>1.0326</td>
</tr>
<tr>
<td>01.11.2001 00:00–01:00</td>
<td>0.9907</td>
<td>1.0517</td>
<td>1.0818</td>
</tr>
<tr>
<td>02.11.2001 16:00–17:00</td>
<td>1.0651</td>
<td>1.0826</td>
<td>1.1189</td>
</tr>
<tr>
<td>01.03.2002 04:00–05:00</td>
<td>0.8313</td>
<td>1.0603</td>
<td>1.1096</td>
</tr>
<tr>
<td>01.03.2002 22:00–23:00</td>
<td>0.8532</td>
<td>1.0542</td>
<td>1.0931</td>
</tr>
<tr>
<td>01.03.2002 23:00–24:00</td>
<td>0.8877</td>
<td>1.0790</td>
<td>1.1308</td>
</tr>
<tr>
<td>26.03.2003 22:00–23:00</td>
<td>0.7475</td>
<td>1.0475</td>
<td>1.0910</td>
</tr>
<tr>
<td>25.03.2003 23:00–24:00</td>
<td>0.8622</td>
<td>1.0376</td>
<td>1.0762</td>
</tr>
<tr>
<td>26.03.2003 23:00–24:00</td>
<td>0.8579</td>
<td>1.0092</td>
<td>1.0299</td>
</tr>
</tbody>
</table>
Traffic prediction

- E-Comm network and traffic data:
  - data preprocessing and extraction
- Data clustering
- Traffic prediction:
  - based on aggregate traffic
  - cluster based

Traffic data: preprocessing

- Original database: ~6 GBytes, with 44,786,489 record rows
- Data pre-processing:
  - cleaning the database
  - filtering the outliers
  - removing redundant records
  - extracting accurate user calling activity
- After the data cleaning and extraction, number of records was reduced to only 19% of original records
## Data preparation

<table>
<thead>
<tr>
<th>Date</th>
<th>Original</th>
<th>Cleaned</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003/03/01</td>
<td>466,862</td>
<td>204,357</td>
<td>91,143</td>
</tr>
<tr>
<td>2003/03/02</td>
<td>415,715</td>
<td>184,973</td>
<td>88,014</td>
</tr>
<tr>
<td>2003/03/03</td>
<td>406,072</td>
<td>182,311</td>
<td>76,310</td>
</tr>
<tr>
<td>2003/03/04</td>
<td>464,534</td>
<td>207,016</td>
<td>84,350</td>
</tr>
<tr>
<td>2003/03/05</td>
<td>585,561</td>
<td>264,226</td>
<td>97,714</td>
</tr>
<tr>
<td>2003/03/06</td>
<td>605,987</td>
<td>271,514</td>
<td>104,715</td>
</tr>
<tr>
<td>2003/03/07</td>
<td>546,230</td>
<td>247,902</td>
<td>94,511</td>
</tr>
<tr>
<td>2003/03/08</td>
<td>513,459</td>
<td>233,982</td>
<td>90,310</td>
</tr>
<tr>
<td>2003/03/09</td>
<td>442,662</td>
<td>201,146</td>
<td>79,815</td>
</tr>
<tr>
<td>2003/03/10</td>
<td>419,570</td>
<td>186,201</td>
<td>76,197</td>
</tr>
<tr>
<td>2003/03/11</td>
<td>504,981</td>
<td>225,604</td>
<td>88,857</td>
</tr>
<tr>
<td>2003/03/12</td>
<td>516,306</td>
<td>233,140</td>
<td>94,779</td>
</tr>
<tr>
<td>2003/03/13</td>
<td>561,253</td>
<td>255,840</td>
<td>95,662</td>
</tr>
<tr>
<td>2003/03/14</td>
<td>550,732</td>
<td>248,828</td>
<td>99,458</td>
</tr>
</tbody>
</table>

Total 92 Days: 44,786,489 Original, 20,130,718 Cleaned, 8,663,586 Combined

44.95% Original, 19.34% Combined

![Graph showing data comparison](image)
User clusters with K-means: $k = 6$
Clustering results

- Larger values of silhouette coefficient produce better results:
  - values between 0.7 and 1.0 imply clustering with excellent separation between clusters
- Cluster sizes:
  - 17, 31, and 569 for K = 3
  - 17, 33, 4, and 563 for K = 4
  - 13, 17, 22, 3, 34, and 528 for K = 6
- K = 3 produces the best clustering results (based on overall clustering quality and silhouette coefficient)
- Interpretations of three clusters have been confirmed by the E-Comm domain experts
### K-means clusters of talk groups: k = 3

<table>
<thead>
<tr>
<th>Cluster size</th>
<th>Minimum number of calls</th>
<th>Maximum number of calls</th>
<th>Average number of calls</th>
<th>Total number of calls</th>
<th>Total number of calls (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>0-6</td>
<td>352-700</td>
<td>94-208</td>
<td>5,091,695</td>
<td>59</td>
</tr>
<tr>
<td>31</td>
<td>0-3</td>
<td>135-641</td>
<td>17-66</td>
<td>2,261,055</td>
<td>26</td>
</tr>
<tr>
<td>569</td>
<td>0</td>
<td>1-1613</td>
<td>0-16</td>
<td>1,310,836</td>
<td>15</td>
</tr>
</tbody>
</table>
Traffic prediction

- Traffic prediction: important to assess future network capacity requirements and to plan future network developments
- A network traffic trace consists of a series of observations in a dynamical system environment
- Traditional prediction: considers aggregate traffic and assumes a constant number of network users
- Approach that focuses on individual users has high computational cost for networks with thousands of users
- Employing clustering techniques for predicting aggregate network traffic bridges the gap between the two approaches
SARIMA models: selection criteria

- **Order (0,1,1)** is used for seasonal part \((P,D,Q)\):
  - cyclical seasonal pattern is usually random-walk
  - may be modeled as MA process after one-time differencing
- Model’s goodness-of-fit is validated using null hypothesis test:
  - time plot analysis and autocorrelation of model residual
Prediction quality

- Models $(2,0,9) \times (0,1,1)_{24}$ and $(2,0,1) \times (0,1,1)_{168}$ have smallest criterion values based on 1,680 training data.
- Normalized mean square error (\textit{nmse}) is used to measure prediction quality by comparing deviation between predicted and observed data.
- The \textit{nmse} of forecast is equal to ratio of normalized sum of variance of forecast to squared bias of forecast.
- Smaller values of \textit{nmse} indicate better prediction model.
Prediction: based on the aggregate traffic

<table>
<thead>
<tr>
<th>No.</th>
<th>p</th>
<th>d</th>
<th>q</th>
<th>P</th>
<th>D</th>
<th>Q</th>
<th>S</th>
<th>m</th>
<th>n</th>
<th>nmse</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>1512</td>
<td>672</td>
<td>0.3790</td>
</tr>
<tr>
<td>A2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>1512</td>
<td>672</td>
<td>0.3803</td>
</tr>
<tr>
<td>A3</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>168</td>
<td>1512</td>
<td>672</td>
<td>0.1742</td>
</tr>
<tr>
<td>A4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>168</td>
<td>1512</td>
<td>672</td>
<td>0.1732</td>
</tr>
<tr>
<td>B1</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>1680</td>
<td>168</td>
<td>0.3790</td>
</tr>
<tr>
<td>B2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>1680</td>
<td>168</td>
<td>0.4079</td>
</tr>
<tr>
<td>B3</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>168</td>
<td>1680</td>
<td>168</td>
<td>0.1736</td>
</tr>
<tr>
<td>B4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>168</td>
<td>1680</td>
<td>168</td>
<td>0.1745</td>
</tr>
<tr>
<td>C1</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>2016</td>
<td>168</td>
<td>0.3384</td>
</tr>
<tr>
<td>C2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>2016</td>
<td>168</td>
<td>0.3433</td>
</tr>
<tr>
<td>C3</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>168</td>
<td>2016</td>
<td>168</td>
<td>0.1282</td>
</tr>
<tr>
<td>C4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>168</td>
<td>2016</td>
<td>168</td>
<td>0.1178</td>
</tr>
</tbody>
</table>

Models forecast future \( n \) traffic data based on \( m \) past traffic data samples
Prediction: based on the aggregate traffic

- Two groups of models, with 24-hour and 168-hour seasonal periods:
  - SARIMA \( (2, 0, 9) \times (0, 1, 1)_{24} \) and \( 168 \)
  - SARIMA \( (2, 0, 1) \times (0, 1, 1)_{24} \) and \( 168 \)

- Comparisons:
  - rows A1 with A2, B1 with B2, and C1 with C2
  - SARIMA \( (2, 0, 9) \times (0, 1, 1)_{24} \) gives better prediction results than SARIMA \( (2, 0, 1)\times(0, 1, 1)_{24} \)

- Models with a 168-hour seasonal period provided better prediction than the four 24-hour period based models, particularly when predicting long term traffic data
Prediction of 168 hours of traffic based on 1,680 past hours: sample

Comparison of the 24-hour and the 168-hour models
- Solid line: observation
- o: prediction of 168-hour seasonal model
- *: prediction of 24-hour seasonal model
Prediction of 168 hours of traffic based on 1,680 past hours

Comparisons: model \((1,0,1)\times(0,1,1)_{168}\)
* observation
* prediction without clustering
* prediction with clustering
Traffic prediction with user clusters

- 57% of cluster-based predictions perform better than aggregate-traffic-based prediction with SARIMA model \((2,0,1) \times (0,1,1)_{168}\).
- Prediction of traffic in networks with a variable number of users is possible, as long as the new user groups could be classified into the existing user clusters.
Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection, statistical analysis, clustering tools, prediction analysis
- Case study:
  - wireless network: Telus Mobility
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions and references
ChinaSat data: analysis

- Analysis of network traffic:
  - characteristics of TCP connections
  - network traffic patterns
  - statistical and cluster analysis of traffic
- anomaly detection:
  - statistical methods
  - wavelets
  - principle component analysis

TCP: transport control protocol
Network and traffic data

- **ChinaSat**: network architecture and TCP
- Analysis of **billing** records:
  - aggregated traffic
  - user behavior
- Analysis of **tcpdump** traces:
  - general characteristics
  - TCP options and operating system (OS) fingerprinting
  - network anomalies
DirecPC system diagram
Characteristics of satellite links

- ChinaSat hybrid satellite network
  - Employs geosynchronous satellites deployed by Hughes Network Systems Inc.
  - Provides data and television services:
    - DirecPC (Classic): unidirectional satellite data service
    - DirecTV: satellite television service
    - DirecWay (Hughnet): new bi-directional satellite data service that replaces DirecPC
- DirecPC transmission rates:
  - 400 kb/s from satellite to user
  - 33.6 kb/s from user to network operations center (NOC) using dial-up
- Improves performance using TCP splitting with spoofing
ChinaSat data: analysis

- ChinaSat traffic is self-similar and non-stationary
- Hurst parameter differs depending on traffic load
- Modeling of TCP connections:
  - inter-arrival time is best modeled by the Weibull distribution
  - number of downloaded bytes is best modeled by the lognormal distribution
- The distribution of visited websites is best modeled by the discrete Gaussian exponential (DGX) distribution
ChinaSat data: analysis

- Traffic prediction:
  - autoregressive integrative moving average (ARIMA) was successfully used to predict uploaded traffic (but not downloaded traffic)
  - wavelet + autoregressive model outperforms the ARIMA model

Analysis of collected data

- Analysis of patterns and statistical properties of two sets of data from the ChinaSat DirecPC network:
  - billing records
  - tcpdump traces

- Billing records:
  - daily and weekly traffic patterns
  - user classification:
    - single and multi-variable k-means clustering based on average traffic
    - hierarchical clustering based on user activity
Analysis of collected data

- Analysis of `tcpdump` trace
  - `tcpdump` trace:
    - protocols and applications
    - TCP options
    - operating system fingerprinting
    - network anomalies
- Developed C program `pcapread`:
  - processes `tcpdump` files
  - produces custom output
  - eliminates the need for packet capture library `libpcap`
Network anomalies

- Scans and worms
- Denial of service
- Flash crowd
- Traffic shift
- Alpha traffic
- Traffic volume anomalies
Network anomalies

- Records were collected during the continuous period from 23:00 on Oct. 31, 2002 to 11:00 on Jan. 10, 2003
- Each file contains the hourly traffic summary for each user
- Fields of interests:
  - SiteID (user identification)
  - Start (record start time)
  - CTxByt (number of bytes downloaded by a user)
  - CRxByt (number of bytes uploaded by a user)
  - CTxPkt (number of packets downloaded by a user)
  - CRxPkt (number of packets uploaded by a user)

download: satellite to user
upload: user to NOC
Aggregated hourly traffic
Aggregated daily traffic
Daily diurnal traffic: average downloaded bytes

![Graph showing daily diurnal traffic]

- Average downloaded traffic (packets)
- Average uploaded traffic (packets)
Weekly traffic: 
average downloaded bytes

- **Average downloaded traffic (bytes)**
- **Average uploaded traffic (bytes)**

![Weekly Traffic Graph]

- The graph shows the average downloaded and uploaded traffic bytes per day of the week.
- The x-axis represents the days of the week (Sun to Sat).
- The y-axis represents the average traffic in bytes, scaled by $10^8$.
- There are noticeable peaks and troughs throughout the week, indicating varying levels of traffic.

---

April 27, 2009  
University of California, Irvine  
75
Ranking of user traffic

- Users are ranked according to the traffic volume
- The top user downloaded 78.8 GB, uploaded 11.9 GB, and downloaded/uploaded ~205 million packets
- Most users download/uploaded little traffic
- Cumulative distribution functions (CDFs) are constructed from the ranks:
  - top user accounts for 11% of downloaded bytes
  - top 25 users contributed 93.3% of downloaded bytes
  - top 37 users contributed 99% of total traffic (packets and bytes)
Refinement: clustering results

<table>
<thead>
<tr>
<th>Traffic pattern</th>
<th>Number of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle</td>
<td>162</td>
</tr>
<tr>
<td>Active</td>
<td>16</td>
</tr>
<tr>
<td>Semi-active</td>
<td>8</td>
</tr>
<tr>
<td>Total number of users</td>
<td>186</td>
</tr>
</tbody>
</table>
Tcpdump traces

- Traces were continuously collected from 11:30 on Dec. 14, 2002 to 11:00 on Jan. 10, 2003 at the NOC.
- The first 68 bytes of each TCP/IP packet were captured.
- ~63 GB of data contained in 127 files.
- User IP address is not constant due to the use of the private IP address range and dynamic IP.
- Majority of traffic is TCP:
  - 94% of total bytes and 84% of total packets.
  - HTTP (port 80) accounts for 90% of TCP connections and 76% of TCP bytes.
  - FTP (port 21) accounts for 0.2% of TCP connections and 11% of TCP bytes.
Network anomalies

- Ethereal/Wireshark, tcptrace, and pcapread
- Four types of network anomalies were detected:
  - invalid TCP flag combinations
  - large number of TCP resets
  - UDP and TCP port scans
  - traffic volume anomalies
## Analysis of TCP flags

<table>
<thead>
<tr>
<th>TCP flag</th>
<th>Packet count</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN only</td>
<td>19,050,849</td>
<td>48.500</td>
</tr>
<tr>
<td>RST only</td>
<td>7,440,418</td>
<td>18.900</td>
</tr>
<tr>
<td>FIN only</td>
<td>12,679,619</td>
<td>32.300</td>
</tr>
<tr>
<td>*SYN+FIN</td>
<td>408</td>
<td>0.001</td>
</tr>
<tr>
<td>*RST+FIN (no PSH)</td>
<td>85,571</td>
<td>0.200</td>
</tr>
<tr>
<td>*RST+PSH (no FIN)</td>
<td>18,111</td>
<td>0.050</td>
</tr>
<tr>
<td>*RST+FIN+PSH</td>
<td>8,329</td>
<td>0.020</td>
</tr>
<tr>
<td>*Total number of packets with invalid TCP flag combinations</td>
<td>112,419</td>
<td>0.300</td>
</tr>
<tr>
<td>Total packet count</td>
<td>39,283,305</td>
<td>100.000</td>
</tr>
</tbody>
</table>
UDP port scans direct to the ChinaSat network

- External address (210.x.x.23) scans for port (137) (NETBEUI) response within the ChinaSat network from source port (1035):
  - > 200 are recorded within a three-hour period
  - targets IP addresses are not sequential
  - may correspond to Bugbear, OpaSoft, or other worms

210.x.x.23:1035 - 192.168.1.121:137
210.x.x.23:1035 - 192.168.1.63:137
210.x.x.23:1035 - 192.168.2.11:137
210.x.x.23:1035 - 192.168.1.250:137
210.x.x.23:1035 - 192.168.1.25:137
210.x.x.23:1035 - 192.168.2.79:137
210.x.x.23:1035 - 192.168.1.52:137
210.x.x.23:1035 - 192.168.6.191:137
210.x.x.23:1035 - 192.168.1.241:137
210.x.x.23:1035 - 192.168.2.91:137
210.x.x.23:1035 - 192.168.1.15:137
210.x.x.23:1035 - 192.168.6.127:137
210.x.x.23:1035 - 192.168.1.201:137
210.x.x.23:1035 - 192.168.6.179:137
210.x.x.23:1035 - 192.168.2.82:137
210.x.x.23:1035 - 192.168.1.239:137
210.x.x.23:1035 - 192.168.1.87:137
210.x.x.23:1035 - 192.168.1.90:137
210.x.x.23:1035 - 192.168.1.177:137
210.x.x.23:1035 - 192.168.1.39:137
Detection of traffic volume anomalies using wavelets

- Traffic is decomposed into various frequencies using the wavelet transform
- Traffic volume anomalies are identified by the large variation in wavelet coefficient values
- The coarsest scale level where the anomalies are found indicates the time scale of an anomaly
Detection of traffic volume anomalies using wavelets

- **tcpdump** traces are binned in terms of packets or bytes (each second)
- Wavelet transform of 12 levels is employed to decompose the traffic
- The coarsest level approximately represents the hourly traffic
- Anomalies are:
  - detected with a moving window of size 20 and by calculating the mean and standard deviation ($\sigma$) of the wavelet coefficients in each window
  - identified when wavelet coefficients lie outside the $\pm 3\sigma$ of the mean value
Wavelet approximation coefficients
Wavelet detail coefficients: $d_9$
Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection, statistical analysis, clustering tools, prediction analysis
- Case studies:
  - wireless network: Telus Mobility
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions and references
Autonomous System (AS)

- Internet is a network of Autonomous Systems:
  - groups of networks sharing the same routing policy
  - identified with Autonomous System Numbers (ASN)
- Autonomous System Numbers: [http://www.iana.org/assignments/as-numbers](http://www.iana.org/assignments/as-numbers)
- Internet topology on AS-level:
  - the arrangement of ASs and their interconnections
- Border Gateway Protocol (BGP):
  - inter-AS protocol
  - used to exchange network reachability information among BGP systems
  - reachability information is stored in routing tables
Internet AS-level data

Source of data are routing tables:

- Route Views: http://www.routeviews.org
  - most participating ASs reside in North America
- RIPE (Réseaux IP européens): http://www.ripe.net/ris
  - most participating ASs reside in Europe
Internet AS-level data

Data used in prior research (partial list):

<table>
<thead>
<tr>
<th></th>
<th>Route Views</th>
<th>RIPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faloutsos, 1999</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Chang, 2001</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Vukadinovic, 2001</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Mihail, 2003</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Research results have been used in developing Internet simulation tools:
- power-laws are employed to model and generate Internet topologies: BA model, BRITE, Inet2
Spectral analysis of graphs

- Normalized Laplacian matrix $N(G)$ [Chung, 1997]:

$$N(i, j) = \begin{cases} 
1 & \text{if } i = j \text{ and } d_i \neq 0 \\
-\frac{1}{\sqrt{d_i d_j}} & \text{if } i \text{ and } j \text{ are adjacent} \\
0 & \text{otherwise}
\end{cases}$$

$d_i$ and $d_j$ are degrees of node $i$ and $j$, respectively

- The second smallest eigenvalue [Fiedler, 1973]
- The largest eigenvalue [Chung, 1997]
- Characteristic valuation [Fiedler, 1975]
Spectral analysis of topology data

- Consider only ASs with the first 30,000 assigned AS numbers
- AS degree distribution in Route Views and RIPE datasets:
Before the sort

(a) RouteViews_original

(b) RIPE_original

After the sort

(c) RouteViews_min

(d) RIPE_min
Before the sort

(a) RouteViews_original

(b) RIPE_original

After the sort

(c) RouteViews_max

(d) RIPE_max
Data analysis results

- The second smallest eigenvector:
  - separates connected ASs from disconnected ASs
  - Route Views and RIPE datasets are similar on a coarser scale

- The largest eigenvector:
  - reveals highly connected clusters
  - Route Views and RIPE datasets differ on a finer scale
Observations

- The two datasets are similar on coarse scales:
  - number of ASs, number of AS connections, core ASs
- They exhibit different clustering characteristics:
  - Route Views data contain larger AS clusters
  - core ASs in Route Views have larger degrees than core ASs in RIPE
  - core ASs in Route Views connect a larger number of smaller ASs
Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection
  - statistical analysis, clustering tools, prediction analysis
- Case studies:
  - wireless network: Telus Mobility
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data network: Internet
- Conclusions, future work, and references
Conclusions

- Traffic data from deployed networks (Telus Mobility, E-Comm, ChinaSat, the Internet) were used to:
  - evaluate network performance
  - characterize and model traffic (inter-arrival and call holding times)
  - classify network users using clustering algorithms
  - predict network traffic by employing SARIMA models based on aggregate user traffic and user clusters
  - detect network anomalies using wavelet analysis
Current project

- Measuring traffic from BC.NET: http://www.bc.net/
  BCNET builds high-performance networks for British Columbia's research and education institutes. A not-for-profit society, BCNET is collectively funded by BC's universities, federal and provincial governments.
- Collecting user traffic and BGP data from routing tables
- Measuring equipment:
  - Endace Ninjabox 5000 (10 Gbps): 16 GB RAM, 16 TB RAID storage with write-to-disk performance of 5 Gbps
  - Endace Ninjabox 504 (1 Gbps): 8 GB RAM, 8 TB RAID storage with write-to-disk performance of 2 Gbps

BGP: border gateway protocol
April 27, 2009

BC.NET traffic measurements

Current transit loads:
500 Mbps  In
500 Mbps  Out

Supplier 1

Supplier 2

Supplier 3

10000 Capture system

10/100/1000/10000 Ethernet switch 1

10/100/1000 Capture system

10/100/1000 Ethernet switch 1

10/100/1000 Ethernet switch 2

BC.NET router

10 GigE link

1 GigE link

Proposed additions
References: downloads

http://www.ensc.sfu.ca/~ljilja/publications_date.html


