Data Mining and Machine Learning for Analysis of Network Traffic

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 collection, characterization, and modeling
- Case studies:
  - telecommunication network: BCNET
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions
Ihr: 535,102 nodes and 601,678 links

http://www.caida.org/home/
Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- Case studies:
  - telecommunication network: BCNET
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions
Measurements of network traffic

- **Traffic measurements:**
  - help understand characteristics of network traffic
  - are basis for developing traffic models
  - are used to evaluate performance of protocols and applications

- **Traffic analysis:**
  - provides information about the network usage
  - helps understand the behavior of network users

- **Traffic prediction:**
  - important to assess future network capacity requirements
  - used to plan future network developments
Traffic modeling: self-similarity

- Self-similarity implies a “fractal-like” behavior
- Data on various time scales have similar patterns
- Implications:
  - no natural length of bursts
  - bursts exist across many time scales
  - traffic does not become “smoother” when aggregated (unlike Poisson traffic)
  - it is unlike Poisson traffic used to model traffic in telephone networks
  - as the traffic volume increases, the traffic becomes more bursty and more self-similar
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:
  - $r^{(m)}(k) = r(k), \ k \geq 0, \ m = 1, 2, \ldots, \ n,$
  where $m$ is the level of aggregation.
Self-similar processes

- Properties:
  - slowly decaying variance
  - long-range dependence
  - *Hurst parameter* (*H*)

- Processes with only short-range dependence (Poisson): *H* = 0.5
- Self-similar processes: *0.5 < H < 1.0*
- *As the traffic volume increases, the traffic becomes more bursty, more self-similar, and the Hurst parameter increases*
Self-similarity: influence of time-scales

- **Genuine MPEG traffic trace**

Self-similarity: influence of time-scales

- Synthetically generated Poisson model

Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- Case studies:
  - telecommunication network: BCNET
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions
Case study: BCNET

- BCNET is the hub of advanced telecommunication network in British Columbia, Canada that offers services to research and higher education institutions.
- The BCNET network is high-speed fiber optic research network.
- British Columbia's network extends to 1,400 km and connects Kamloops, Kelowna, Prince George, Vancouver, and Victoria.
BCNET packet capture
BCNET packet capture

- BCNET transits have two service providers with 10 Gbps network links and one service provider with 1 Gbps network link
- Optical Test Access Point (TAP) splits the signal into two distinct paths
- The signal splitting ratio from TAP may be modified
- The Data Capture Device (NinjaBox 5000) collects the real-time data (packets) from the traffic filtering device
Net Optics Director 7400: application diagram

- Net Optics Director 7400 is used for BCNET traffic filtering
- It directs traffic to monitoring tools such as NinjaBox 5000 and FlowMon
Network monitoring and analyzing: Endace card

- Endace Data Acquisition and Generation (DAG) 5.2X card resides inside the NinjaBox 5000
- It captures and transmits traffic and has time-stamping capability
- DAG 5.2X is a single port Peripheral Component Interconnect Extended (PCIX) card and is capable of capturing on average Ethernet traffic of 6.9 Gbps
Real time network usage by BCNET members
Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- Case studies:
  - telecommunication network: BCNET
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions
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

Municipalities with all three emergency services (police, fire, ambulance) on the E-Comm Wide-Area Radio System
E-Comm network architecture

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)
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
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>2001</th>
<th></th>
<th>2002</th>
<th></th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day/hour</td>
<td>No.</td>
<td>Day/hour</td>
<td>No.</td>
<td>Day/hour</td>
<td>No.</td>
</tr>
<tr>
<td>02.11.2001 15:00–16:00</td>
<td>3,718</td>
<td>01.03.2002 04:00–05:00</td>
<td>4,436</td>
<td>26.03.2003 22:00–23:00</td>
<td>4,919</td>
</tr>
<tr>
<td>01.11.2001 00:00–01:00</td>
<td>3,707</td>
<td>01.03.2002 22:00–23:00</td>
<td>4,314</td>
<td>25.03.2003 23:00–24:00</td>
<td>4,249</td>
</tr>
<tr>
<td>02.11.2001 16:00–17:00</td>
<td>3,492</td>
<td>01.03.2002 23:00–24:00</td>
<td>4,179</td>
<td>26.03.2003 23:00–24:00</td>
<td>4,222</td>
</tr>
<tr>
<td>01.11.2001 19:00–20:00</td>
<td>3,312</td>
<td>01.03.2002 00:00–01:00</td>
<td>3,971</td>
<td>29.03.2003 02:00–03:00</td>
<td>4,150</td>
</tr>
<tr>
<td>02.11.2001 20:00–21:00</td>
<td>3,227</td>
<td>02.03.2002 00:00–01:00</td>
<td>3,939</td>
<td>29.03.2003 01:00–02:00</td>
<td>4,097</td>
</tr>
</tbody>
</table>
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

Call inter-arrival time (s)

Probability density

0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6

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>Year</th>
<th>Date</th>
<th>Time</th>
<th>$H$</th>
<th>Year</th>
<th>Date</th>
<th>Time</th>
<th>$H$</th>
<th>Year</th>
<th>Date</th>
<th>Time</th>
<th>$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>02.11.2001</td>
<td>15:00–16:00</td>
<td>0.907</td>
<td>2002</td>
<td>01.03.2002</td>
<td>04:00–05:00</td>
<td>0.679</td>
<td>2003</td>
<td>26.03.2003</td>
<td>22:00–23:00</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>01.11.2001</td>
<td>00:00–01:00</td>
<td>0.802</td>
<td></td>
<td>01.03.2002</td>
<td>22:00–23:00</td>
<td>0.757</td>
<td></td>
<td>25.03.2003</td>
<td>23:00–24:00</td>
<td>0.832</td>
</tr>
<tr>
<td></td>
<td>02.11.2001</td>
<td>16:00–17:00</td>
<td>0.770</td>
<td></td>
<td>01.03.2002</td>
<td>23:00–24:00</td>
<td>0.780</td>
<td></td>
<td>26.03.2003</td>
<td>23:00–24:00</td>
<td>0.699</td>
</tr>
<tr>
<td></td>
<td>01.11.2001</td>
<td>19:00–20:00</td>
<td>0.774</td>
<td></td>
<td>01.03.2002</td>
<td>00:00–01:00</td>
<td>0.741</td>
<td></td>
<td>29.03.2003</td>
<td>02:00–03:00</td>
<td>0.696</td>
</tr>
<tr>
<td></td>
<td>02.11.2001</td>
<td>20:00–21:00</td>
<td>0.663</td>
<td></td>
<td>02.03.2002</td>
<td>00:00–01:00</td>
<td>0.747</td>
<td></td>
<td>29.03.2003</td>
<td>01:00–02:00</td>
<td>0.705</td>
</tr>
</tbody>
</table>
Call holding times: pdf candidates

![Graph showing probability density against call holding time with various models compared to traffic data.]

- Traffic data
- Lognormal model
- Gamma model
- Weibull model
- Exponential model
- Normal model
- Rayleigh model
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></th>
<th>2001</th>
<th></th>
<th>2002</th>
<th></th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day/hour</td>
<td>$H$</td>
<td>Day/hour</td>
<td>$H$</td>
<td>Day/hour</td>
<td>$H$</td>
</tr>
<tr>
<td>02.11.2001 15:00–16:00</td>
<td>0.493</td>
<td>01.03.2002 04:00–05:00</td>
<td>0.490</td>
<td>26.03.2003 22:00–23:00</td>
<td>0.483</td>
</tr>
<tr>
<td>01.11.2001 00:00–01:00</td>
<td>0.471</td>
<td>01.03.2002 22:00–23:00</td>
<td>0.460</td>
<td>25.03.2003 23:00–24:00</td>
<td>0.483</td>
</tr>
<tr>
<td>02.11.2001 16:00–17:00</td>
<td>0.462</td>
<td>01.03.2002 23:00–24:00</td>
<td>0.489</td>
<td>26.03.2003 23:00–24:00</td>
<td>0.463</td>
</tr>
</tbody>
</table>
* |
| 01.11.2001 19:00–20:00 | 0.467 | 01.03.2002 00:00–01:00 | 0.508 | 29.03.2003 02:00–03:00 | 0.526 |
| 02.11.2001 20:00–21:00 | 0.479 | 02.03.2002 00:00–01:00 | 0.503 | 29.03.2003 01:00–02:00 | 0.466 |
### Call inter-arrival and call holding times

<table>
<thead>
<tr>
<th></th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day/hour</td>
<td>Avg. (s)</td>
<td>Day/hour</td>
</tr>
<tr>
<td>inter-arrival</td>
<td>02.11.2001</td>
<td>0.97</td>
<td>01.03.2002</td>
</tr>
<tr>
<td></td>
<td>15:00–16:00</td>
<td>3.78</td>
<td>04:00–05:00</td>
</tr>
<tr>
<td>holding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter-arrival</td>
<td>01.11.2001</td>
<td>0.97</td>
<td>01.03.2002</td>
</tr>
<tr>
<td></td>
<td>00:00–01:00</td>
<td>3.95</td>
<td>22:00–23:00</td>
</tr>
<tr>
<td>holding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter-arrival</td>
<td>02.11.2001</td>
<td>1.03</td>
<td>01.03.2002</td>
</tr>
<tr>
<td></td>
<td>16:00–17:00</td>
<td>3.99</td>
<td>23:00–24:00</td>
</tr>
<tr>
<td>holding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter-arrival</td>
<td>01.11.2001</td>
<td>1.09</td>
<td>01.03.2002</td>
</tr>
<tr>
<td></td>
<td>19:00–20:00</td>
<td>3.97</td>
<td>00:00–01:00</td>
</tr>
<tr>
<td>holding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter-arrival</td>
<td>02.11.2001</td>
<td>1.12</td>
<td>02.03.2002</td>
</tr>
<tr>
<td></td>
<td>20:00–21:00</td>
<td>3.84</td>
<td>00:00–01:00</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>Call inter-arrival times</th>
<th>Call holding times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weibull</td>
<td>Gamma</td>
</tr>
<tr>
<td></td>
<td>( a )</td>
<td>( b )</td>
</tr>
<tr>
<td>02.11.2001 15:00–16:00</td>
<td>0.9785</td>
<td>1.1075</td>
</tr>
<tr>
<td>01.11.2001 00:00–01:00</td>
<td>0.9907</td>
<td>1.0517</td>
</tr>
<tr>
<td>02.11.2001 16:00–17:00</td>
<td>1.0651</td>
<td>1.0826</td>
</tr>
<tr>
<td>01.03.2002 04:00–05:00</td>
<td>0.8313</td>
<td>1.0603</td>
</tr>
<tr>
<td>01.03.2002 22:00–23:00</td>
<td>0.8532</td>
<td>1.0542</td>
</tr>
<tr>
<td>01.03.2002 23:00–24:00</td>
<td>0.8877</td>
<td>1.0790</td>
</tr>
<tr>
<td>26.03.2003 22:00–23:00</td>
<td>0.7475</td>
<td>1.0475</td>
</tr>
<tr>
<td>25.03.2003 23:00–24:00</td>
<td>0.8622</td>
<td>1.0376</td>
</tr>
<tr>
<td>26.03.2003 23:00–24:00</td>
<td>0.8579</td>
<td>1.0092</td>
</tr>
</tbody>
</table>
Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- **Case studies:**
  - telecommunication network: BCNET
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions
Case study: ChinaSat DirecPC system

- 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 DirecPC system
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
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
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
Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- Case studies:
  - telecommunication network: BCNET
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions
Internet topology

- 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 ASes and their interconnections
- Analyzing the Internet topology and finding properties of associated graphs rely on mining data and capturing information about Autonomous Systems (ASes)
Variety of graphs

- **Random graphs:**
  - nodes and edges are generated by a random process
  - Erdős and Rényi model

- **Small world graphs:**
  - nodes and edges are generated so that most of the nodes are connected by a small number of nodes in between
  - Watts and Strogatz model (1998)
Scale-free graphs

- **Scale-free graphs:**
  - graphs whose node degree distribution follow power-law
  - rich get richer
  - Barabási and Albert model (1999)

- **Analysis of complex networks:**
  - discovery of spectral properties of graphs
  - constructing matrices describing the network connectivity
Analyzed datasets

- Sample datasets:
  - Route Views:
    TABLE_DUMP| 1050122432| B| 204.42.253.253| 267| 3.0.0.0/8| 267 2914 174 701| IGP|
    204.42.253.253| 0| 0| 267:2914 2914:420 2914:2000 2914:3000| NAG|
  - RIPE:
    TABLE_DUMP| 1041811200| B| 212.20.151.234| 13129| 3.0.0.0/8| 13129 6461 7018 | IGP|
    212.20.151.234| 0| 0| 6461:5997 13129:3010| NAG|
Internet topology at AS level

- Datasets collected from Border Gateway Protocols (BGP) routing tables are used to infer the Internet topology at AS-level
The Internet topology is characterized by the presence of various power-laws:

- node degree vs. node rank
- eigenvalues of the matrices describing Internet graphs (adjacency matrix and normalized Laplacian matrix)

Power-laws exponents have not significantly changed over the years

Spectral analysis reveals new historical trends and notable changes in the connectivity and clustering of AS nodes over the years
Traffic anomalies

- Slammer, Nimda, and Code Red I anomalies affected performance of the Internet Border Gateway Protocol (BGP)
- BGP anomalies also include: Internet Protocol (IP) prefix hijacks, miss-configurations, and electrical failures
- Techniques for detecting BGP anomalies have recently gained visible attention and importance
Anomaly detection techniques

- **Classification problem:**
  - assigning an “anomaly” or “regular” label to a data point
- **Accuracy of a classifier depends on:**
  - extracted features
  - combination of selected features
  - underlying model

**Goal:**
- Detect Internet routing anomalies using the Border Gateway Protocol (BGP) update messages
BGP features

Approach:

- Define a set of 37 features based on BGP update messages
- Extract the features from available BGP update messages that are collected during the time period when the Internet experienced anomalies:
  - Slammer
  - Nimda
  - Code Red I
Feature selection

- Select the most relevant features for classification using:
  - Fisher
  - Minimum Redundancy Maximum Relevance (mRMR)
  - Odds Ratio
  - Decision Tree
  - Fuzzy Rough Sets
Anomaly classification

- Train classifiers for BGP anomaly detection using:
  - Support Vector Machines (SVM)
  - Long Short-Term Memory (LSTM) Neural Network
  - Hidden Markov Models (HMM)
  - Naive Bayes (NB)
  - Decision Tree
  - Extreme Learning Machine (ELM)
Feature extraction: BGP messages

- **Border Gateway Protocol (BGP)** enables exchange of routing information between gateway routers using update messages.

- **BGP update message collections:**
  - Réseaux IP Européens (RIPE) under the Routing Information Service (RIS) project
  - Route Views
  - Available in multi-threaded routing toolkit (MRT) binary format
### BGP: known anomalies

<table>
<thead>
<tr>
<th>Anomaly</th>
<th>Date</th>
<th>Duration (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slammer</td>
<td>January 25, 2003</td>
<td>869</td>
</tr>
<tr>
<td>Nimda</td>
<td>September 18-20, 2001</td>
<td>3,521</td>
</tr>
<tr>
<td>Code Red I</td>
<td>July 19, 2001</td>
<td>600</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
<th>Peers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moscow power blackout</td>
<td>May 2005</td>
<td>AS 1853, AS 12793, AS 13237</td>
</tr>
<tr>
<td>AS 9121 routing table leak</td>
<td>Dec. 2004</td>
<td>AS 1853, AS 12793, AS 13237</td>
</tr>
<tr>
<td>AS 3561 improper filtering</td>
<td>Apr. 2001</td>
<td>AS 3257, AS 3333, AS 286</td>
</tr>
<tr>
<td>Panix domain hijack</td>
<td>Jan. 2006</td>
<td>AS 12956, AS 6762, AS 6939, AS 3549</td>
</tr>
<tr>
<td>As-path error</td>
<td>Oct. 2001</td>
<td>AS 3257, AS 3333, AS 6762, AS 9057</td>
</tr>
</tbody>
</table>
## Training and test datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training dataset</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Slammer and Nimda</td>
<td>Code Red I</td>
</tr>
<tr>
<td>2</td>
<td>Slammer and Code Red I</td>
<td>Nimda</td>
</tr>
<tr>
<td>3</td>
<td>Nimda and Code Red I</td>
<td>Slammer</td>
</tr>
<tr>
<td>4</td>
<td>Slammer</td>
<td>Nimda and Code Red I</td>
</tr>
<tr>
<td>5</td>
<td>Nimda</td>
<td>Slammer and Code Red I</td>
</tr>
<tr>
<td>6</td>
<td>Code Red I</td>
<td>Slammer and Nimda</td>
</tr>
<tr>
<td>7</td>
<td>Slammer, Nimda, and Code Red I</td>
<td>RIPE or BCNET</td>
</tr>
</tbody>
</table>
Slammer worm

- Sends its replica to randomly generated IP addresses
- Destination IP address gets infected if:
  - it is a Microsoft SQL server
  or
  - a personal computer with the Microsoft SQL Server Data Engine (MSDE)
Nimda worm

- Propagates through email messages, web browsers, and file systems
- Viewing the email message triggers the worm payload
- The worm modifies the content of the web document files in the infected hosts and copies itself in all local host directories
Code Red I worm

- Takes advantage of vulnerability in the Microsoft Internet Information Services (IIS) indexing software
- It triggers a buffer overflow in the infected hosts by writing to the buffers without checking their limit
Feature extraction: BGP messages

- Define 37 features
- Sample every minute during a five-day period:
  - the peak day of an anomaly
  - two days prior and two days after the peak day
- 7,200 samples for each anomalous event:
  - 5,760 regular samples (non-anomalous)
  - 1,440 anomalous samples
  - Imbalanced dataset
### BGP features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Definition</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of announcements</td>
<td>Volume</td>
</tr>
<tr>
<td>2</td>
<td>Number of withdrawals</td>
<td>Volume</td>
</tr>
<tr>
<td>3</td>
<td>Number of announced NLRI prefixes</td>
<td>Volume</td>
</tr>
<tr>
<td>4</td>
<td>Number of withdrawn NLRI prefixes</td>
<td>Volume</td>
</tr>
<tr>
<td>5</td>
<td>Average AS-PATH length</td>
<td>AS-path</td>
</tr>
<tr>
<td>6</td>
<td>Maximum AS-PATH length</td>
<td>AS-path</td>
</tr>
<tr>
<td>7</td>
<td>Average unique AS-PATH length</td>
<td>AS-path</td>
</tr>
<tr>
<td>8</td>
<td>Number of duplicate announcements</td>
<td>Volume</td>
</tr>
<tr>
<td>9</td>
<td>Number of duplicate withdrawals</td>
<td>Volume</td>
</tr>
<tr>
<td>10</td>
<td>Number of implicit withdrawals</td>
<td>Volume</td>
</tr>
</tbody>
</table>
# BGP features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Definition</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Average edit distance</td>
<td>AS-path</td>
</tr>
<tr>
<td>12</td>
<td>Maximum edit distance</td>
<td>AS-path</td>
</tr>
<tr>
<td>13</td>
<td>Inter-arrival time</td>
<td>Volume</td>
</tr>
<tr>
<td>14-24</td>
<td>Maximum edit distance = n, where n = (7, ..., 17)</td>
<td>AS-path</td>
</tr>
<tr>
<td>25-33</td>
<td>Maximum AS-path length = n, where n = (7, ..., 15)</td>
<td>AS-path</td>
</tr>
<tr>
<td>34</td>
<td>Number of IGP packets</td>
<td>Volume</td>
</tr>
<tr>
<td>35</td>
<td>Number of EGP packets</td>
<td>Volume</td>
</tr>
<tr>
<td>36</td>
<td>Number of incomplete packets</td>
<td>Volume</td>
</tr>
<tr>
<td>37</td>
<td>Packet size (B)</td>
<td>Volume</td>
</tr>
</tbody>
</table>
Feature selection algorithms

- Employed to select the most relevant features:
  - Fisher
  - Minimum Redundancy Maximum Relevance (mRMR)
  - Odds Ratio
  - Decision Tree
  - Fuzzy Rough Sets
### Feature selection: decision tree

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training data</th>
<th>Selected Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>Slammer + Nimda</td>
<td>1-21, 23-29, 34-37</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>Slammer + Code Red I</td>
<td>1-22, 24-29, 34-37</td>
</tr>
<tr>
<td>Dataset 3</td>
<td>Code Red I + Nimda</td>
<td>1-29, 34-37</td>
</tr>
</tbody>
</table>

- Either four (30, 31, 32, 33) or five (22, 30, 31, 32, 33) features are removed in the constructed trees mainly because:
  - features are numerical and some are used repeatedly
Anomaly classification

- Train classifiers for BGP anomaly detection using:
  - Support Vector Machines (SVM)
  - Long Short-Term Memory (LSTM) Neural Network
  - Hidden Markov Models (HMM)
  - Naive Bayes (NB)
  - Decision Tree
  - Extreme Learning Machine (ELM)
Repeating modules for the LSTM neural network: input layer, LSTM layer with one LSTM cell, and output layer.
## Anomaly classifiers: LSTM

<table>
<thead>
<tr>
<th></th>
<th>Test dataset</th>
<th>Accuracy (%)</th>
<th>F-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RIPE</td>
<td>BCNET</td>
<td>RIPE</td>
</tr>
<tr>
<td>LSTMu 1</td>
<td>Code Red I</td>
<td>95.22</td>
<td>65.49</td>
</tr>
<tr>
<td>LSTMu 2</td>
<td>Nimda</td>
<td>53.94</td>
<td>51.53</td>
</tr>
<tr>
<td>LSTMu 3</td>
<td>Slammer</td>
<td>95.87</td>
<td>56.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Test dataset</th>
<th>Accuracy (%)</th>
<th>F-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RIPE</td>
<td>BCNET</td>
<td>RIPE</td>
</tr>
<tr>
<td>LSTMb 1</td>
<td>Code Red I</td>
<td>56.43</td>
<td>60.48</td>
</tr>
<tr>
<td>LSTMb 2</td>
<td>Nimda</td>
<td>53.32</td>
<td>44.27</td>
</tr>
<tr>
<td>LSTMb 3</td>
<td>Slammer</td>
<td>82.98</td>
<td>55.00</td>
</tr>
</tbody>
</table>
Anomaly classifiers: decision tree

Each path from the root node to a leaf node may be transformed into a decision rule.

A set of rules that are obtained from a trained decision decision tree may be used for classifying unseen samples.

<table>
<thead>
<tr>
<th>Training dataset</th>
<th>Test dataset</th>
<th>Accuracy (%)</th>
<th>F-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>Code Red I</td>
<td>85.36</td>
<td>89.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>77.22</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>Nimda</td>
<td>58.13</td>
<td>94.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>81.18</td>
</tr>
<tr>
<td>Dataset 3</td>
<td>Slammer</td>
<td>95.89</td>
<td>89.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>77.78</td>
</tr>
</tbody>
</table>
Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- Case studies:
  - telecommunication network: BCNET
  - public safety wireless network: E-Comm
  - satellite network: ChinaSat
  - packet data networks: Internet
- Conclusions
Conclusions

- Data collected from deployed networks are used to:
  - evaluate network performance
  - characterize and model traffic (inter-arrival and call holding times)
  - identify trends in the evolution of the Internet topology
  - classify traffic and network anomalies
References: sources of data


References:
http://www.sfu.ca/~ljilja/cnl


lhr: 535,102 nodes and 601,678 links

http://www.caida.org/home/