Communication Networks:
Traffic Data, Network Topologies,
and Routing Anomalies

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 study: Collection of BCNET traffic
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- Conclusions
Ihr: 535,102 nodes and 601,678 links

http://www.caida.org/home/
Ihr: 535,102 nodes and 601,678 links

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

- Introduction
- **Traffic collection, characterization, and modeling**
- Case study: Collection of BCNET traffic
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- 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
  - 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: 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 study: Collection of BCNET traffic**
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- Conclusions
BCNET packet capture: physical overview

- BCNET is the hub of advanced telecommunication network in British Columbia, Canada that offers services to research and higher education institutions.
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

- 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
Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- Case study: Collection of BCNET traffic
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- 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
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|
  ```
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
Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- Case study: Collection of BCNET traffic
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- Conclusions
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 algorithms

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

- **Train classifiers** for BGP anomaly detection using:
  - Support Vector Machines
  - Hidden Markov Models
  - Naive Bayes
  - Decision Tree
  - Extreme Learning Machine (ELM)
BGP: update 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: anomalies

<table>
<thead>
<tr>
<th>Anomaly</th>
<th>Date</th>
<th>Duration (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slammer</td>
<td>January 25, 2003</td>
<td>16</td>
</tr>
<tr>
<td>Nimda</td>
<td>September 18, 2001</td>
<td>59</td>
</tr>
<tr>
<td>Code Red I</td>
<td>July 19, 2001</td>
<td>10</td>
</tr>
</tbody>
</table>

#### Training Data

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slammer + Nimda</td>
<td>Dataset 1</td>
</tr>
<tr>
<td>Slammer + Code Red I</td>
<td>Dataset 2</td>
</tr>
<tr>
<td>Code Red I + Nimda</td>
<td>Dataset 3</td>
</tr>
<tr>
<td>Slammer</td>
<td>Dataset 4</td>
</tr>
<tr>
<td>Nimda</td>
<td>Dataset 5</td>
</tr>
<tr>
<td>Code Red I</td>
<td>Dataset 6</td>
</tr>
</tbody>
</table>
BGP: features

- 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>
<tr>
<td>Feature</td>
<td>Definition</td>
<td>Category</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>----------</td>
</tr>
<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: decision tree

- Commonly used algorithm in data mining
- Generates a model that predicts the value of a target variable based on several input variables
- A top-down approach is commonly used for constructing decision trees:
  - an appropriate variable is chosen to best split the set of items based on homogeneity of the target variable within subsets
- C5 software tool was used to generate decision trees

### 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 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
Feature selection: fuzzy rough sets

- Deal with the approximation of fuzzy sets in a fuzzy approximation space defined by a fuzzy similarity relation or by a fuzzy partition.

- The fuzzy similarity relation $Sim(C)$ is:
  - an $nxn$ matrix that describes similarities between any two samples
  - computed by the min operator

- Computational complexity: $O(n^2m)$
  - $n$ is the number of samples
  - $m$ is the number of features
Feature selection: fuzzy rough sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training data</th>
<th>Selected Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 4</td>
<td>Slammer</td>
<td>1, 3-6, 9, 10, 13-32, 35</td>
</tr>
<tr>
<td>Dataset 5</td>
<td>Nimda</td>
<td>1, 3-4, 8-10, 12, 14-32, 35, 36</td>
</tr>
<tr>
<td>Dataset 6</td>
<td>Code Red I</td>
<td>3-4, 8-10, 12, 14-32, 35, 36</td>
</tr>
</tbody>
</table>

- Using combination of datasets, for example Slammer + Nimda for training leads to higher computational load
- Each dataset was used individually
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 tree may be used for classifying unseen samples.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Testing data</th>
<th>$\text{Acc}_{\text{train}}$</th>
<th>$\text{Acc}_{\text{test}}$</th>
<th>Training time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>Code Red I</td>
<td>90.7</td>
<td>78.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>Nimda</td>
<td>92.3</td>
<td>72.8</td>
<td>2.1</td>
</tr>
<tr>
<td>Dataset 3</td>
<td>Slammer</td>
<td>87.1</td>
<td>23.8</td>
<td>2.3</td>
</tr>
</tbody>
</table>
Anomaly classifier: ELM

- Used for learning with a single hidden layer feed forward neural network
- Weights connecting the input and hidden layers with the bias terms are initialized randomly
- Weights connecting the hidden and output layers are analytically determined
- Learns faster than SVMs by a factor of thousands
- Suitable for online applications
- We use all features (37), all continuous features (17), features selected by fuzzy rough sets (28 or 27), and continuous features selected by fuzzy rough sets (9 or 8)
Anomaly classifiers: ELM

<table>
<thead>
<tr>
<th>No. of features</th>
<th>Dataset</th>
<th>$\text{Acc}_{\text{train}}$</th>
<th>$\text{Acc}_{\text{test}}$</th>
<th>Training time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>Dataset 1</td>
<td>83.57 ± 0.11</td>
<td>80.01 ± 0.07</td>
<td>2.3043</td>
</tr>
<tr>
<td></td>
<td>Dataset 2</td>
<td>83.53 ± 0.12</td>
<td>79.75 ± 0.08</td>
<td>2.2756</td>
</tr>
<tr>
<td></td>
<td>Dataset 3</td>
<td>80.82 ± 0.09</td>
<td>21.65 ± 1.93</td>
<td>2.2747</td>
</tr>
<tr>
<td>17</td>
<td>Dataset 1</td>
<td>84.50 ± 0.07</td>
<td>79.91 ± 0.01</td>
<td>1.9268</td>
</tr>
<tr>
<td></td>
<td>Dataset 2</td>
<td>84.43 ± 0.12</td>
<td>79.53 ± 0.10</td>
<td>1.5928</td>
</tr>
<tr>
<td></td>
<td>Dataset 3</td>
<td>83.06 ± 0.07</td>
<td>51.56 ± 16.38</td>
<td>1.8882</td>
</tr>
</tbody>
</table>

- 195 hidden units
- The binary features 14–33 are removed to form a set of 17 features
## Anomaly classifiers: ELM

<table>
<thead>
<tr>
<th>No. of features</th>
<th>Dataset</th>
<th>$\text{Acc}_{\text{train}}$</th>
<th>$\text{Acc}_{\text{test}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>Dataset 4</td>
<td>83.08 ± 0.11</td>
<td>80.03 ± 0.06</td>
</tr>
<tr>
<td>28 (from 37)</td>
<td>Dataset 5</td>
<td>83.08 ± 0.09</td>
<td>79.78 ± 0.07</td>
</tr>
<tr>
<td>27</td>
<td>Dataset 6</td>
<td>80.05 ± 0.00</td>
<td>81.00 ± 1.41</td>
</tr>
<tr>
<td>9</td>
<td>Dataset 4</td>
<td>84.59 ± 0.07</td>
<td>80.00 ± 0.05</td>
</tr>
<tr>
<td>9 (from 17)</td>
<td>Dataset 5</td>
<td>84.25 ± 0.11</td>
<td>79.79 ± 0.12</td>
</tr>
<tr>
<td>8</td>
<td>Dataset 6</td>
<td>83.38 ± 0.04</td>
<td>49.24 ± 12.90</td>
</tr>
</tbody>
</table>
Roadmap

- Introduction
- Traffic collection, characterization, and modeling
- Case study: Collection of BCNET traffic
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- 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
Conclusions

- **Machine learning algorithms** (feature selection and classification algorithms) are used for detecting BGP anomalies.
- **Performance of classifiers** greatly depended on the employed datasets.
- **Feature selection algorithms** were used to improve the performance of classifiers.
- For smaller datasets, performance of the ELM classifier was improved by using fuzzy rough sets.
- Both **decision tree** and **ELM** are relatively fast classifiers with satisfactory accuracy.
References: sources of data

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


