Analysis of Traffic Data in Communication Networks

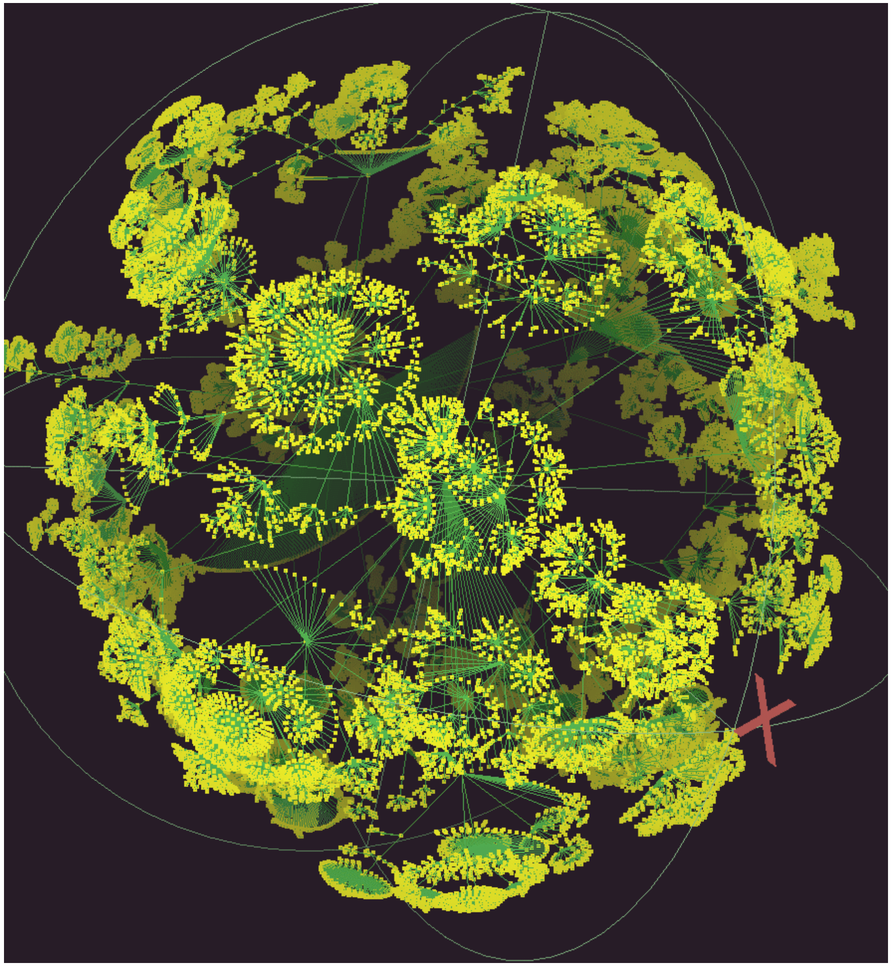
Ljiljana Trajković
ljilja@cs.sfu.ca

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

Roadmap

- Introduction
- Traffic measurements and analysis tools
- Case study:
  - public safety wireless network: E-Comm
- Collection of BCNET traffic
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- Conclusions

lhr: 535,102 nodes and 601,678 links



http://www.caida.org/home

Roadmap

- Introduction
- Traffic measurements and analysis tools
- Case study:
  - public safety wireless network: E-Comm
- 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

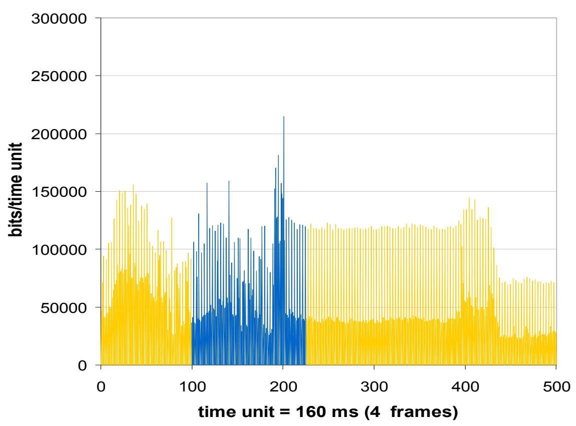
- Traffic measurements:
  - help understand characteristics of network traffic
  - are basis for developing traffic models
  - are used to evaluate performance of protocols and applications
- Traffic analysis:
  - provides information about the network usage
  - helps understand the behavior of network users
- Traffic prediction:
  - important to assess future network capacity requirements
  - used to plan future network developments

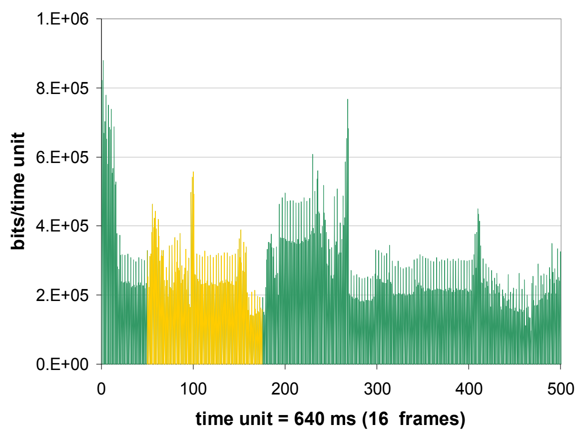
Traffic modeling: self-similarity

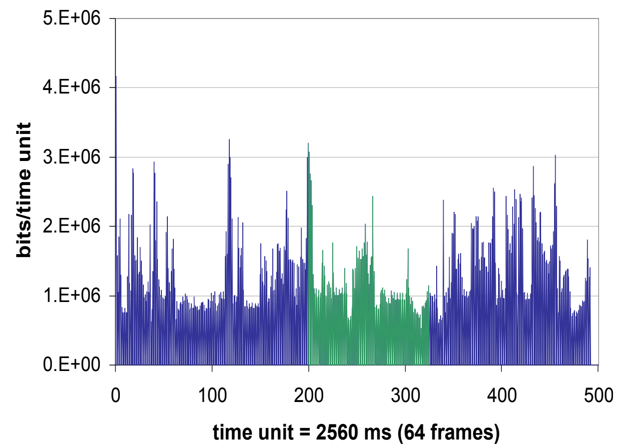
- Self-similarity implies a "fractal-like" behavior
- Data on various time scales have similar patterns
- Implications:
  - no natural length of bursts
  - bursts exist across many time scales
  - traffic does not become "smoother" when aggregated
  - it is unlike Poisson traffic used to model traffic 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



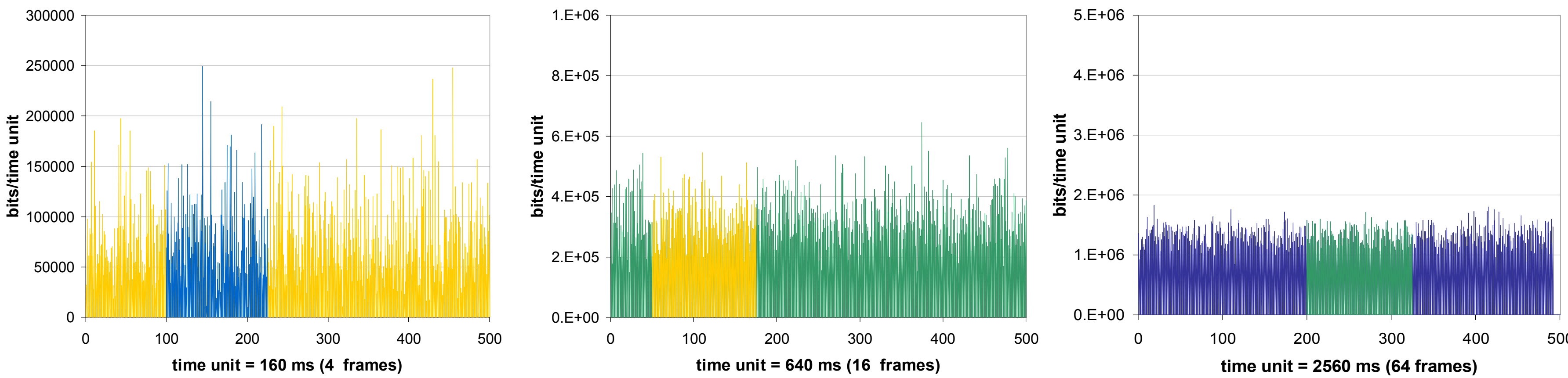




W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no 1, pp. 1-15, Feb. 1994.

Self-similarity: influence of time-scales

- Synthetically generated Poisson model



W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no 1, pp. 1-15, Feb. 1994.

Traffic analysis: clustering analysis

- Clustering generates groups (clusters) of similar objects
- An object is described by a set of measurements
- Clustering algorithms can be used to analyze behavior of network users
- Users are grouped into clusters based on the similarity of their behavior
- Traffic prediction based on clusters is simplified to predicting users' traffic from few clusters
- Clustering tools:
  - k-means algorithm
  - AutoClass tool

Traffic prediction: SARIMA 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
- Seasonal ARIMA (SARIMA) is a variation of the ARIMA model:
  - it captures seasonal patterns

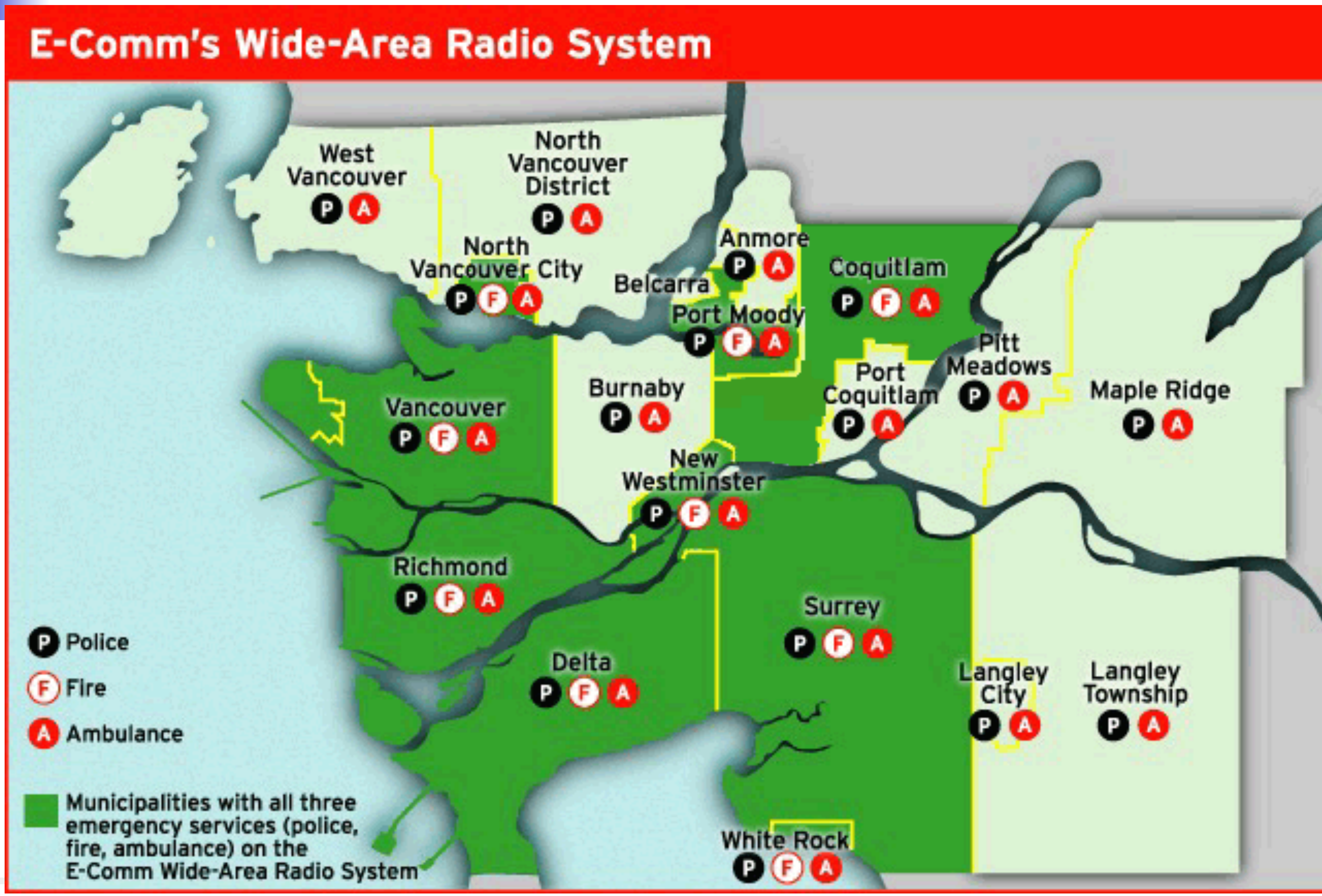
Roadmap

- Introduction
- Traffic measurements and analysis tools
- Case study:
  - public safety wireless network: E-Comm
- Collection of BCNET traffic
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- Conclusions

Case study: E-Comm network

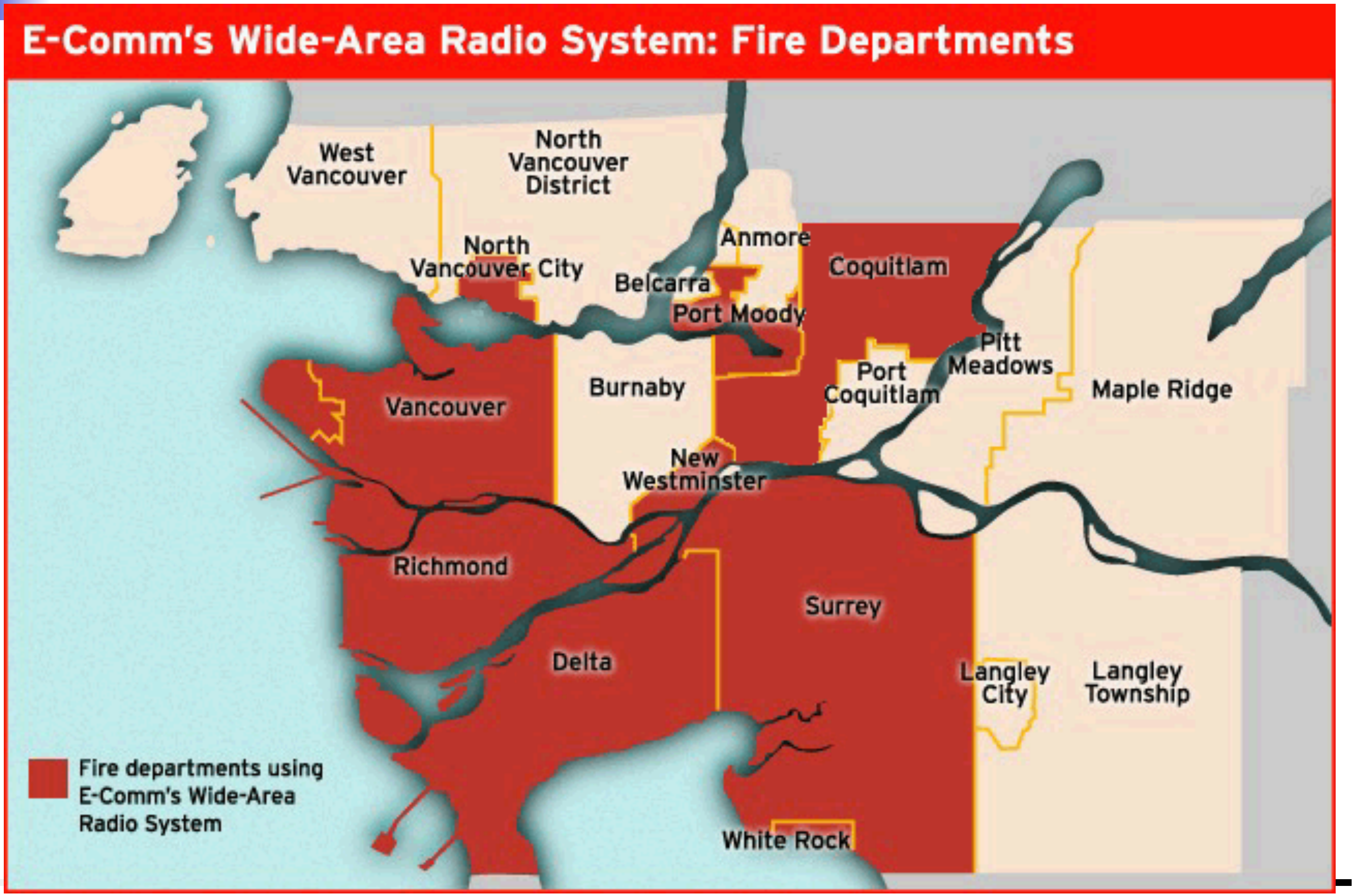
- An operational trunked radio system serving as a regional emergency communication system
- The E-Comm network is capable of both voice and data transmissions
- Voice traffic accounts for over 99% of network traffic
- A group call is a standard call made in a trunked radio system
- More than 85% of 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



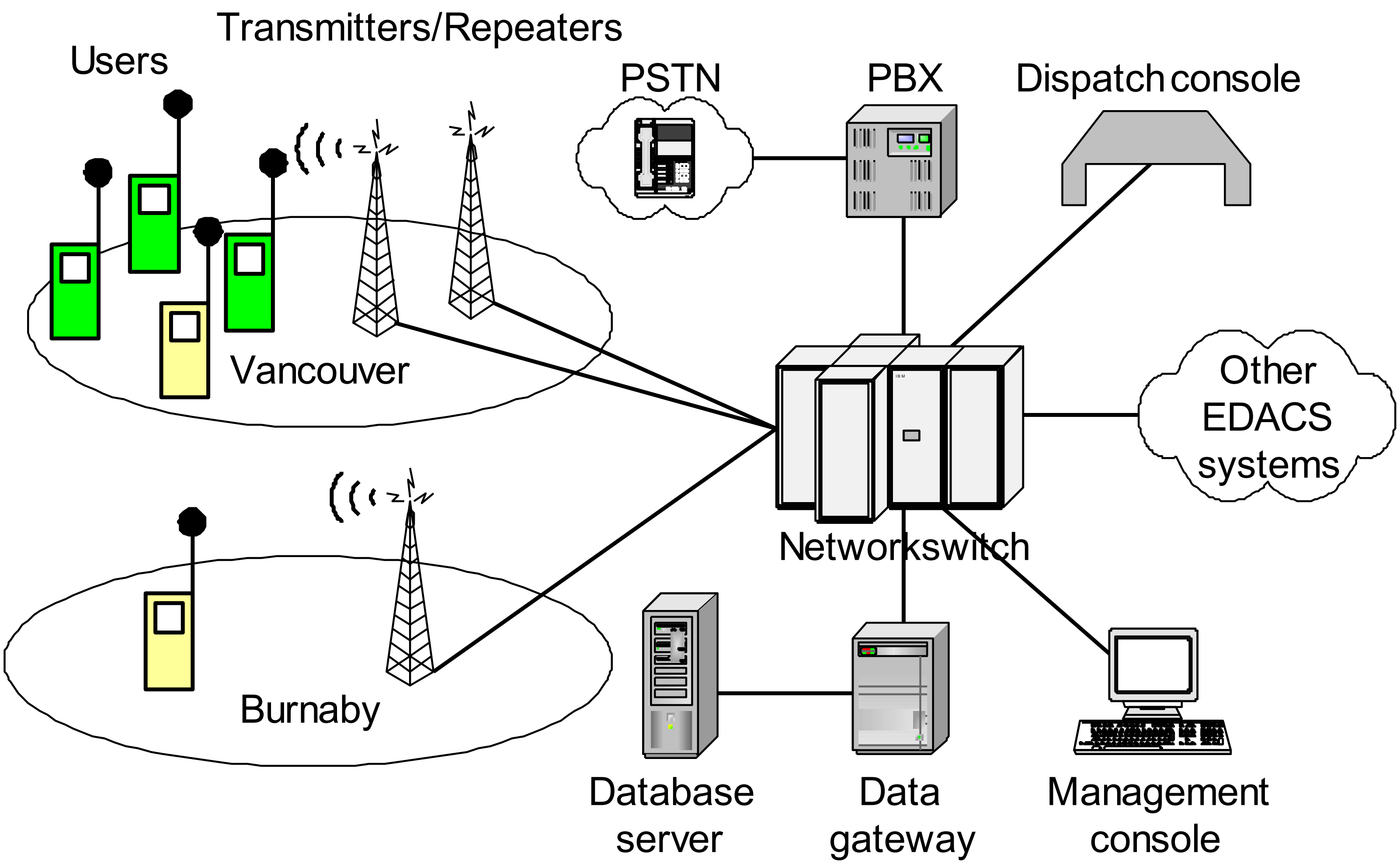
E-Comm network

E-Comm network



E-Comm network

E-Comm network architecture



E-Comm traffic data

- 2001 data set:
  - 2 days of traffic data
    - 2001-11-1 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)

E-Comm traffic data

- Records of network events:
  - established, queued, and dropped calls in the Vancouver cell
- Traffic data span periods during:
  - 2001, 2002, 2003

Trace (dataset)	Time span	No. of established calls
2001	November 1–2, 2001	110,348
2002	March 1–7, 2002	370,510
2003	March 24–30, 2003	387,340

E-Comm traffic: 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
- The busiest hour: around midnight
- The busiest day: Thursday
- Useful for scheduling periodical maintenance tasks

E-Comm traffic: hourly traces

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

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

E-Comm traffic: 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 and call holding times: observations

	2001		2002		2003	
	Day/hour	Avg. (s)	Day/hour	Avg. (s)	Day/hour	Avg. (s)
inter-arrival	02.11.2001 15:00–16:00	0.97	01.03.2002 04:00–05:00	0.81	26.03.2003 22:00–23:00	0.73
holding		3.78		4.07		4.08
inter-arrival	01.11.2001 00:00–01:00	0.97	01.03.2002 22:00–23:00	0.83	25.03.2003 23:00–24:00	0.85
holding		3.95		3.84		4.12
inter-arrival	02.11.2001 16:00–17:00	1.03	01.03.2002 23:00–24:00	0.86	26.03.2003 23:00–24:00	0.85
holding		3.99		3.88		4.04
inter-arrival	01.11.2001 19:00–20:00	1.09	01.03.2002 00:00–01:00	0.91	29.03.2003 02:00–03:00	0.87
holding		3.97		3.95		4.14
inter-arrival	02.11.2001 20:00–21:00	1.12	02.03.2002 00:00–01:00	0.91	29.03.2003 01:00–02:00	0.88
holding		3.84		4.06		4.25

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

Avg. call holding times: 3.91 s (2001), 3.96 s (2002), 4.13 s (2003)

Busy hour: best fitting distributions

Busy hour	Distribution					
	Call inter-arrival times				Call holding times	
	Weibull		Gamma		Lognormal	
	a	b	a	b	μ	σ
02.11.2001 15:00–16:00	0.9785	1.1075	1.0326	0.9407	1.0913	0.6910
01.11.2001 00:00–01:00	0.9907	1.0517	1.0818	0.8977	1.0801	0.7535
02.11.2001 16:00–17:00	1.0651	1.0826	1.1189	0.9238	1.1432	0.6803
01.03.2002 04:00–05:00	0.8313	1.0603	1.1096	0.7319	1.1746	0.6671
01.03.2002 22:00–23:00	0.8532	1.0542	1.0931	0.7643	1.1157	0.6565
01.03.2002 23:00–24:00	0.8877	1.0790	1.1308	0.7623	1.1096	0.6803
26.03.2003 22:00–23:00	0.7475	1.0475	1.0910	0.6724	1.1838	0.6553
25.03.2003 23:00–24:00	0.8622	1.0376	1.0762	0.7891	1.1737	0.6715
26.03.2003 23:00–24:00	0.8579	1.0092	1.0299	0.8292	1.1704	0.6696

E-Comm traffic: clustering

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

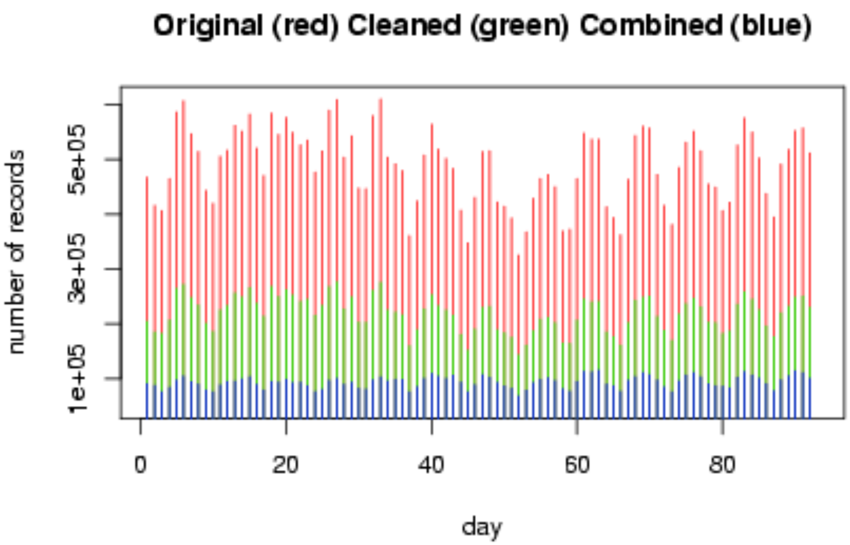
E-Comm traffic: preprocessing

- 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

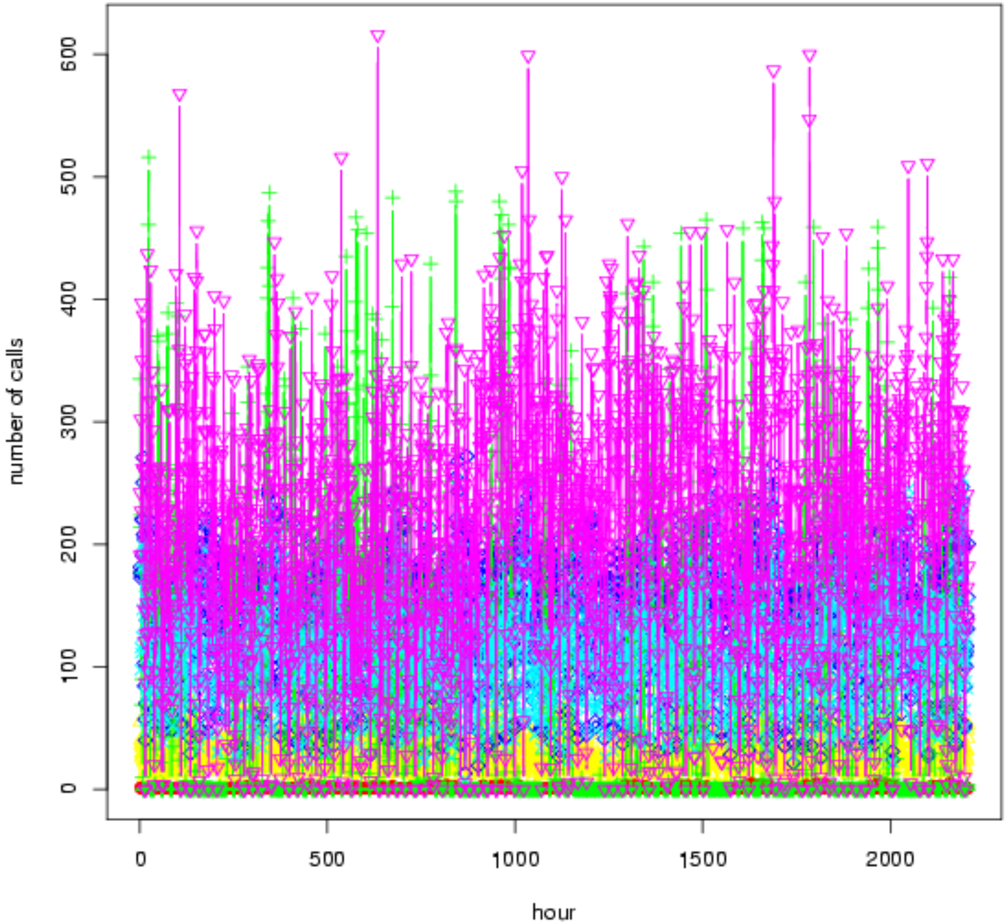
E-Comm traffic: data preparation

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

Total 92 Days	44,786,489	20,130,718	8,663,586
		44.95%	19.34%



User clusters with K-means: k = 6



Clustering results

- 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

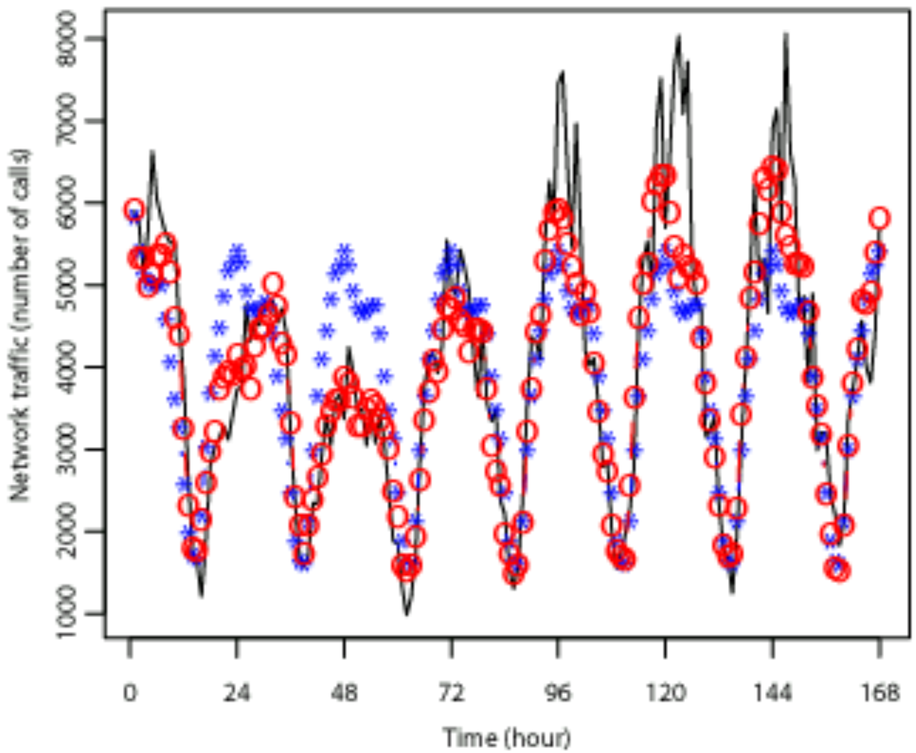
E-Comm traffic: prediction

- 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

Prediction: based on the aggregate traffic

- Two groups of models, with 24-hour and 168-hour seasonal periods:
  - SARIMA $(2, 0, 9) \times (0, 1, 1)_{24 \text{ and } 168}$
  - SARIMA $(2, 0, 1) \times (0, 1, 1)_{24 \text{ and } 168}$
- 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 traffic in networks with a variable number of users is possible, as long as the new users could be classified within the existing clusters

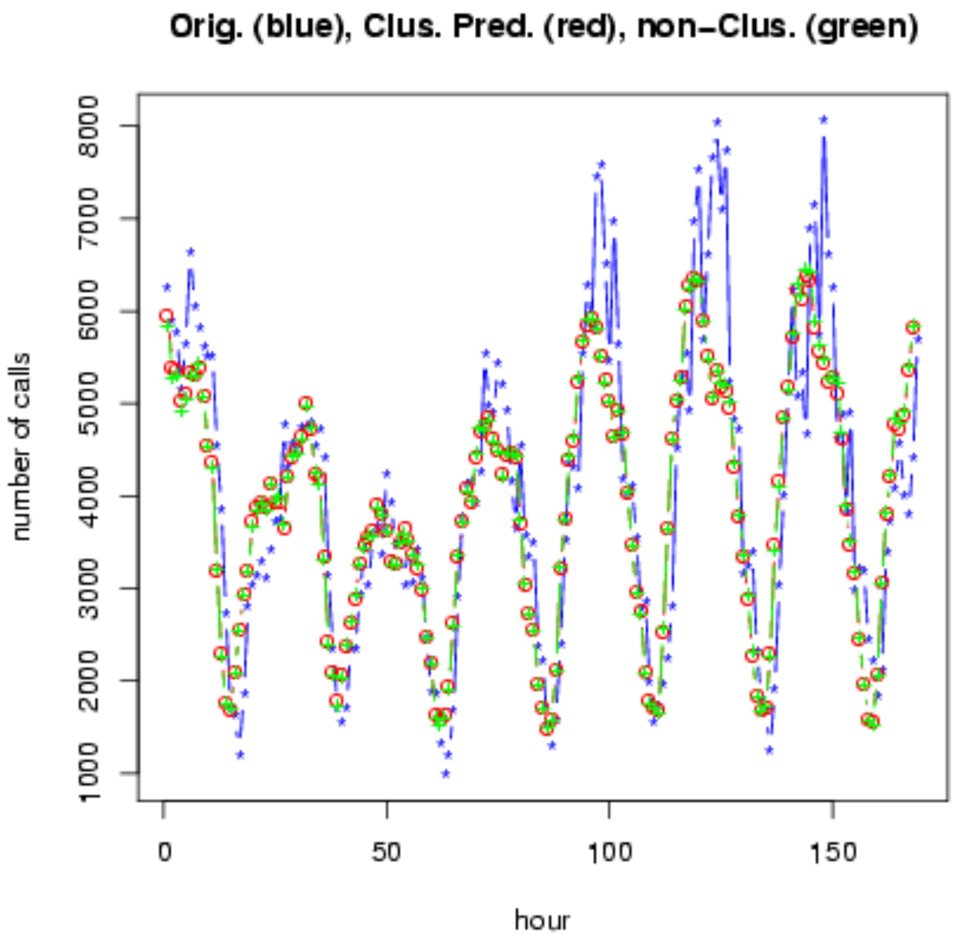
Prediction of 168 hours of traffic based on 1,680 past hours: sample



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 $(1,0,1)\times(0,1,1)_{168}$

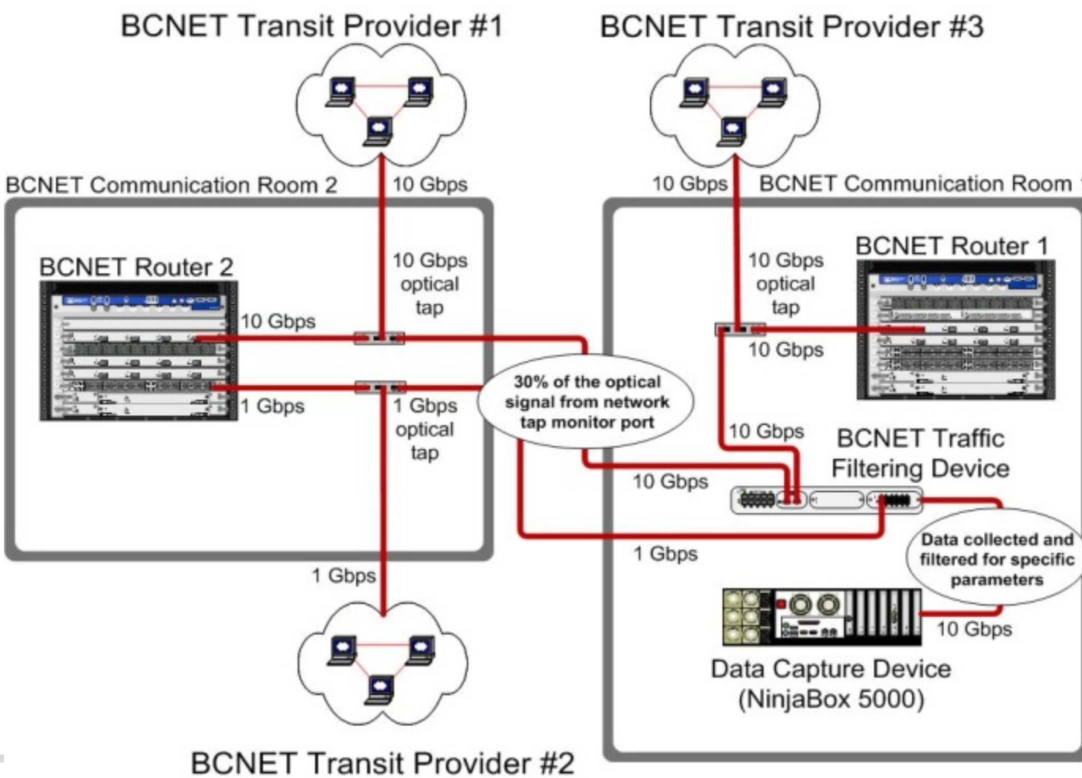
- * observation
- * prediction without clustering
- o prediction with clustering

Roadmap

- Introduction
- Traffic measurements and analysis tools
- Case studies:
  - public safety wireless network: E-Comm
- 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

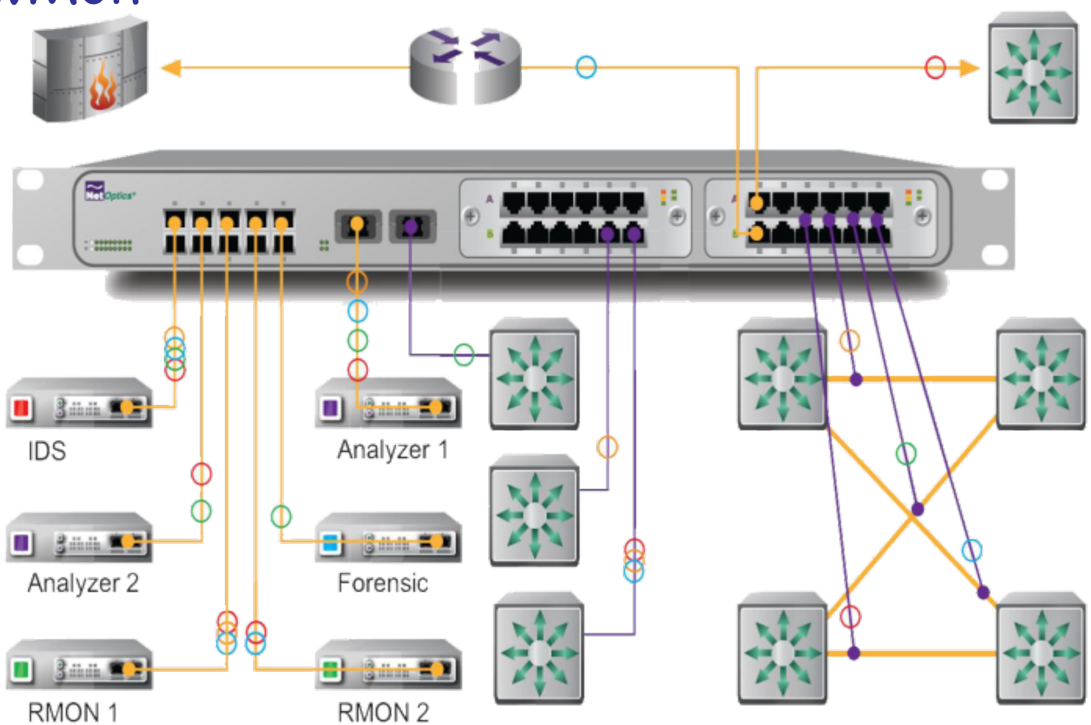


BCNET packet capture

- 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

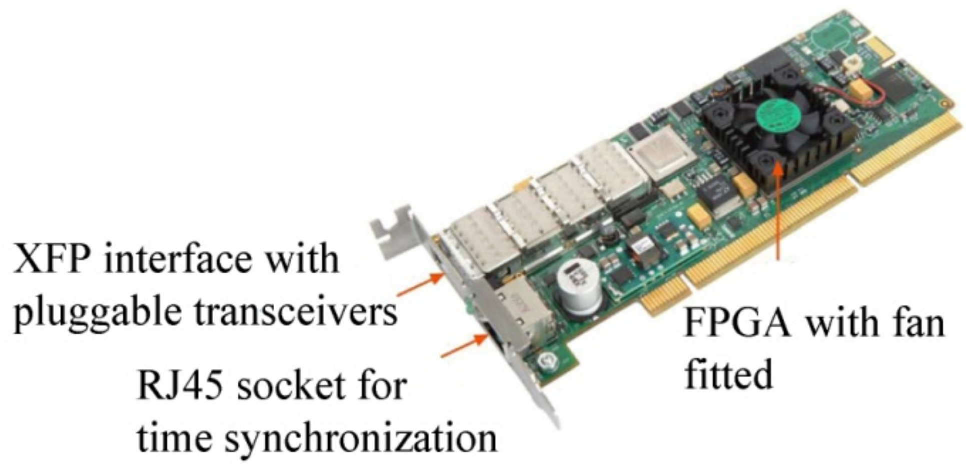
Net Optics Director 7400: application diagram

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



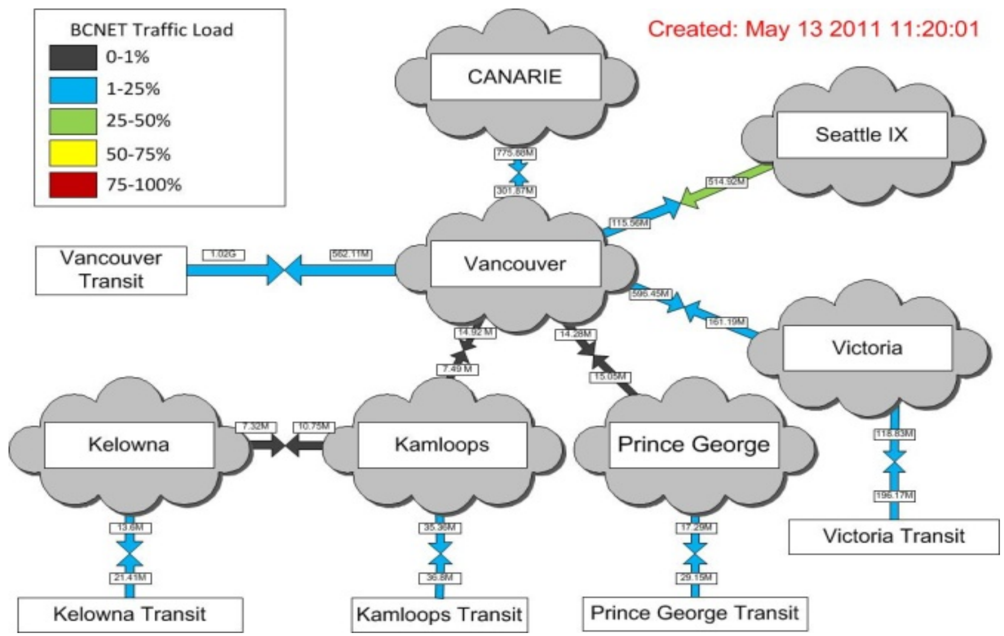
Network monitoring and analyzing: Endace card

- 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 measurements and analysis tools
- Case study:
  - public safety wireless network: E-Comm
- 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
- 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

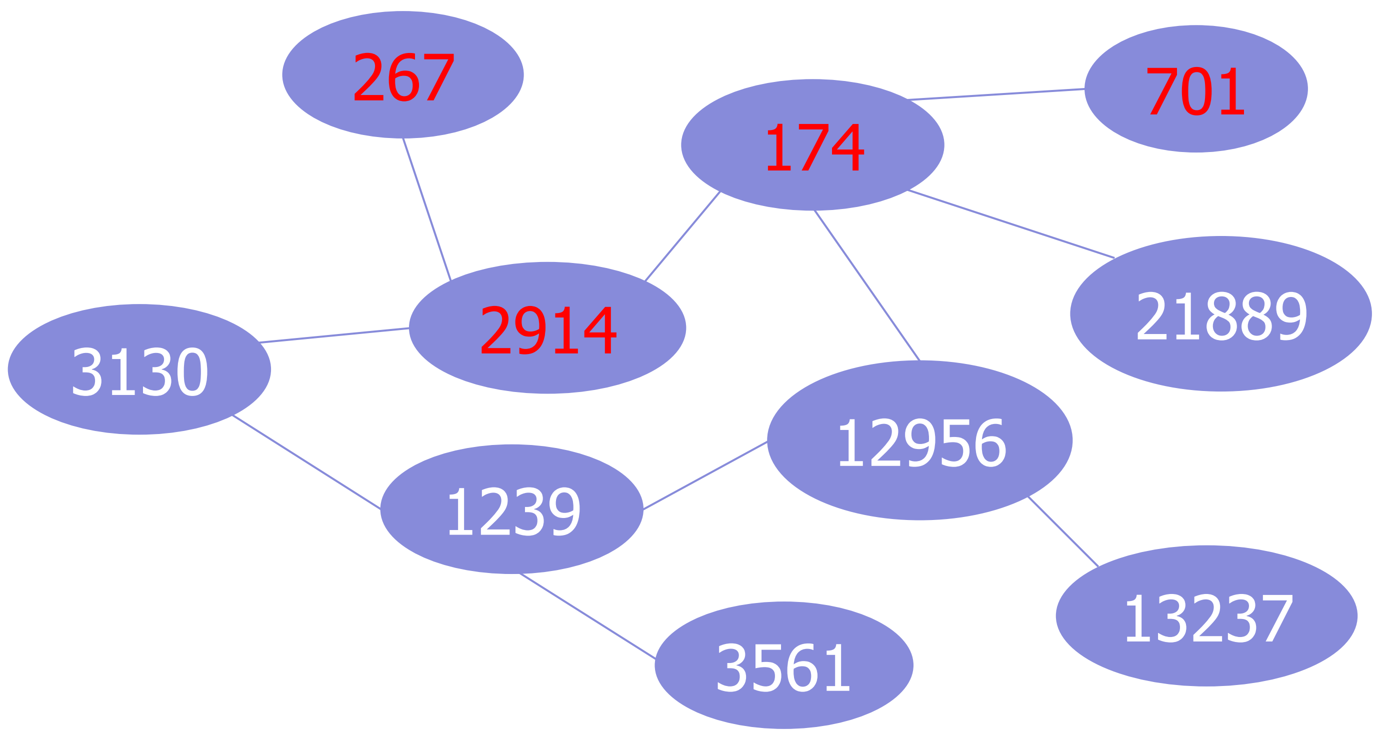
- Sample datasets:
  - Route Views:

    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



Internet topology

- 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 measurements and analysis tools
- Case study:
  - public safety wireless network: E-Comm
- 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
- BGP anomalies often occur
- Techniques for BGP anomalies detection have recently gained visible attention and importance

Sources of datasets

- The RIPE and Route Views BGP update message
- BGP traffic traces collected from the BCNET

	Class	Date	Duration (h)
Slammer	Anomaly	January 25, 2003	16
Nimda	Anomaly	September 18, 2001	59
Code Red I	Anomaly	July 19, 2001	10
RIPE	Regular	July 14, 2001	24
BCNET	Regular	December 20, 2011	24

Extracted features

Feature	Definition	Category
1	Number of announcements	Volume
2	Number of withdrawals	Volume
3	Number of announced NLRI prefixes	Volume
4	Number of withdrawn NLRI prefixes	Volume
5	Average AS-PATH length	AS-path
6	Maximum AS-PATH length	AS-path
7	Average unique AS-PATH length	AS-path
8	Number of duplicate announcements	Volume
9	Number of duplicate withdrawals	Volume
10	Number of implicit withdrawals	Volume

Feature selection algorithms

- Features scoring algorithms:
  - Fisher
  - Minimum Redundancy Maximum Relevance (mRMR)
  - Odds Ratio
- These algorithms measure the correlation and relevancy among features
- The top ten features were selected for the Fisher feature selection

Performance measures and indices

- Performance measures:
  - sensitivity = TP/(TP + FN)
  - precision = TP/(TP + FP)
- Performance indices:
  - accuracy = (TP + TN)/(TP + TN + FP + FN)
  - balanced accuracy = (sensitivity + precision)/2
  - F-score = 2 x (precision x sensitivity)/precision + sensitivity)

- TP = true positive FP = false positive
- TN = true negative FN = false negative

Classification tools

- Support Vector Machines
- Hidden Markov Models
- Naive Bayes

Support Vector Machine

- For each training dataset $X_{7200 \times 37}$, we target two classes:
  - anomaly (true) and regular (false)
- Dimension of feature matrix: 7,200×10
- Each row contains the top ten selected features within the one-minute interval

SVM two-way datasets

	Training dataset	Test dataset
SVMV1	Slammer and Nimda	Code Red I
SVM2	Slammer and Code Red I	Nimda
SVM3	Code Red I and Nimda	Slammer

Two-way classification: performance

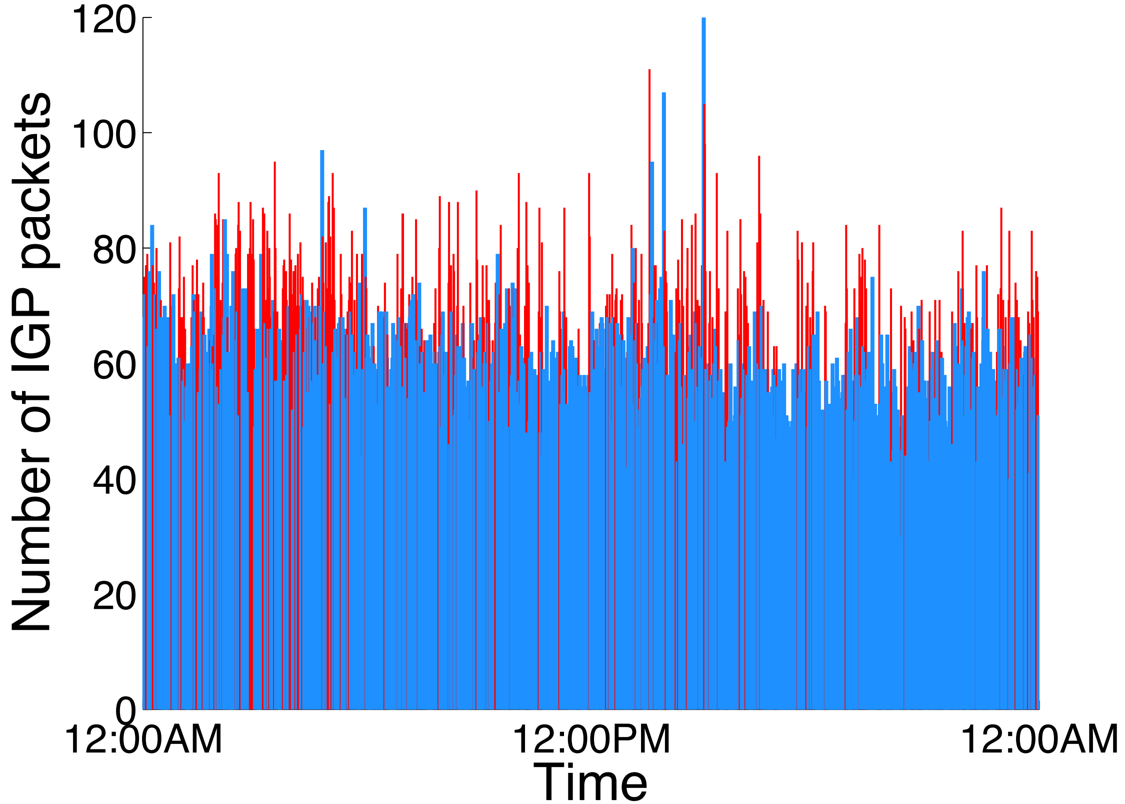
All anomalies are treated as one class

SVM	Feature	Performance index			
		Accuracy (%)			F-score (%)
		Test dataset (anomaly)	RIPE (regular)	BCNET (regular)	Test dataset (anomaly)
SMV3	All features	81.95	92.0	69.2	84.6
SMV3	Fisher	89.3	93.8	68.4	75.2
SMV3	MID	75.4	92.8	71.7	79.2
SMV3	MIQ	85.1	92.2	73.2	86.1
SMV3	MIBASE	89.3	89.7	69.7	80.1

MID: Mutual Information Deference
MIQ: Mutual Information Quotient
MIBASE: Mutual Information Base

Classification results

- Incorrectly classified (anomaly) BCNET traffic collected on December 20, 2011 (red):



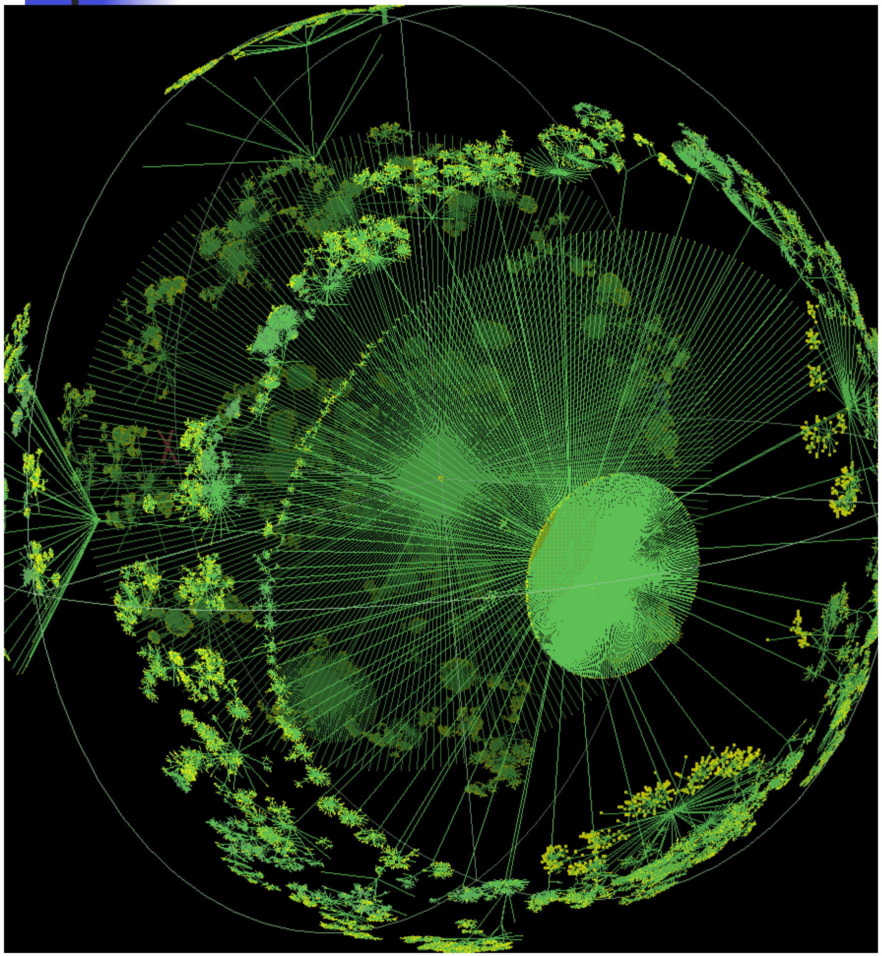
Roadmap

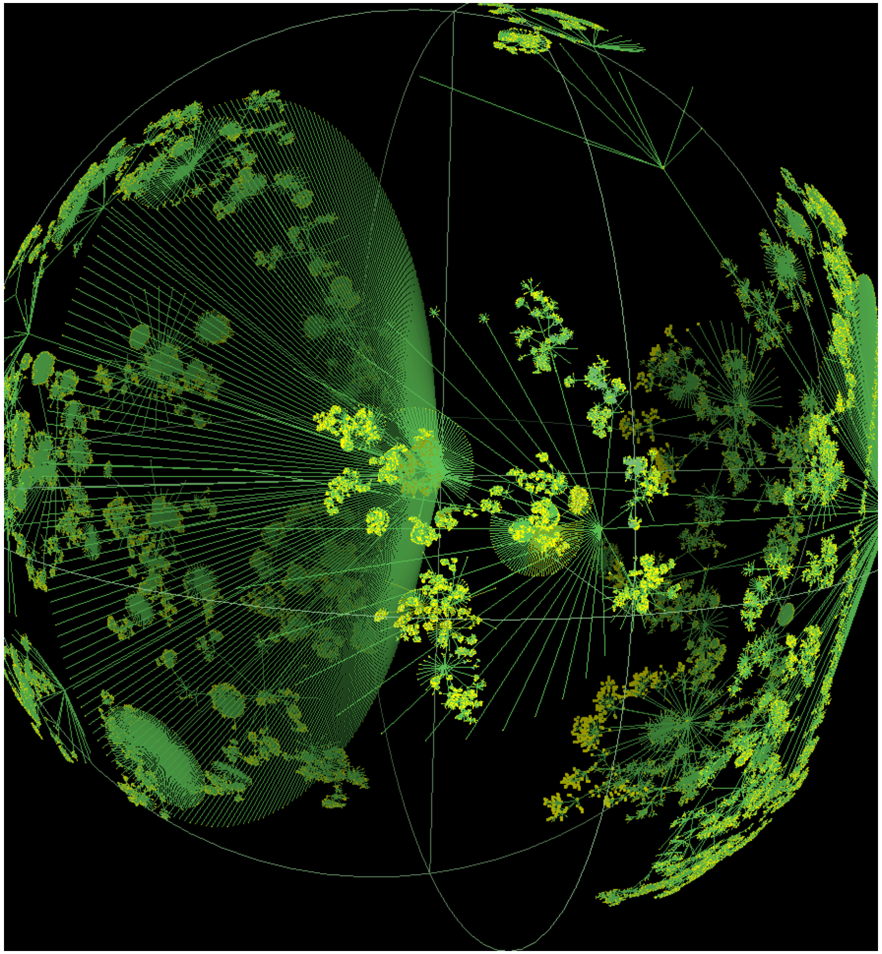
- Introduction
- Traffic measurements and analysis tools
- Case study:
  - public safety wireless network: E-Comm
- Collection of BCNET traffic
- Internet topology and spectral analysis of Internet graphs
- Machine learning models for feature selection and classification of traffic anomalies
- Conclusions

Conclusions

- Data collected from deployed networks can be 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 models based on aggregate user traffic and user clusters
  - identify trends in the evolution of the Internet topology
  - classify traffic and network anomalies

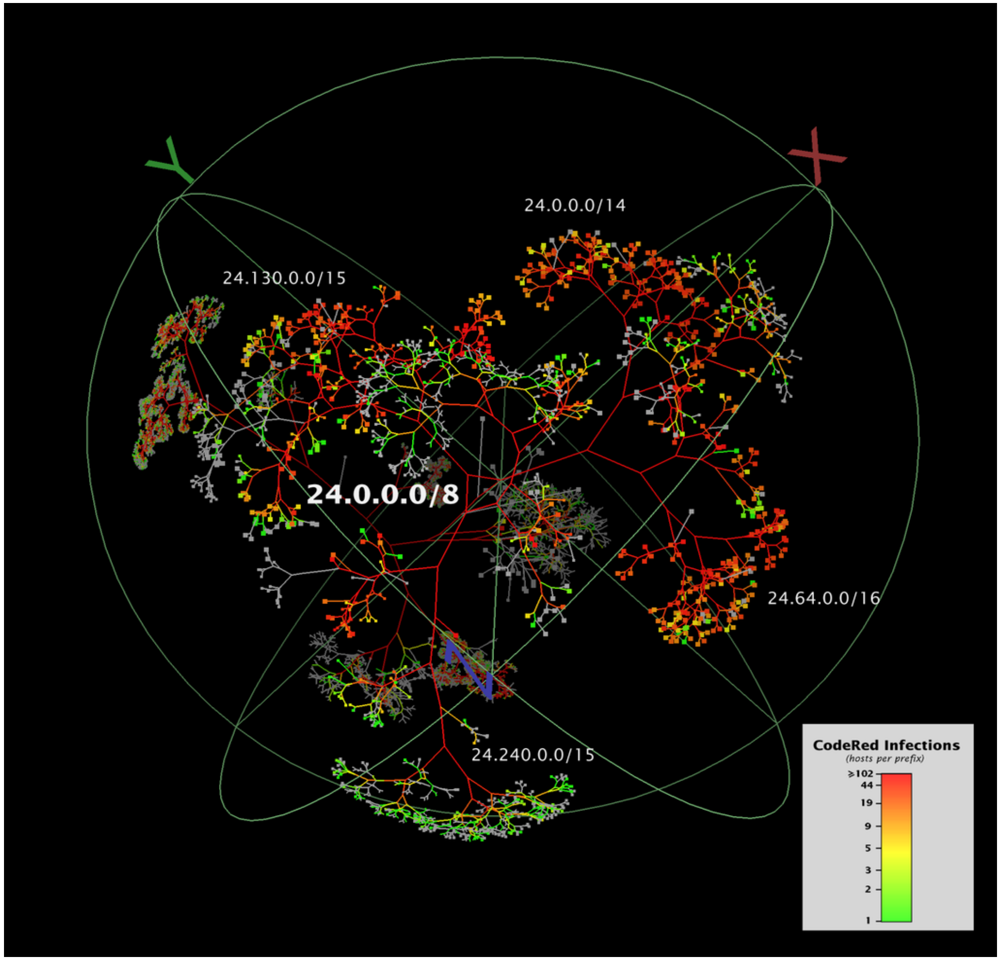
lhr: 535,102 nodes and 601,678 links





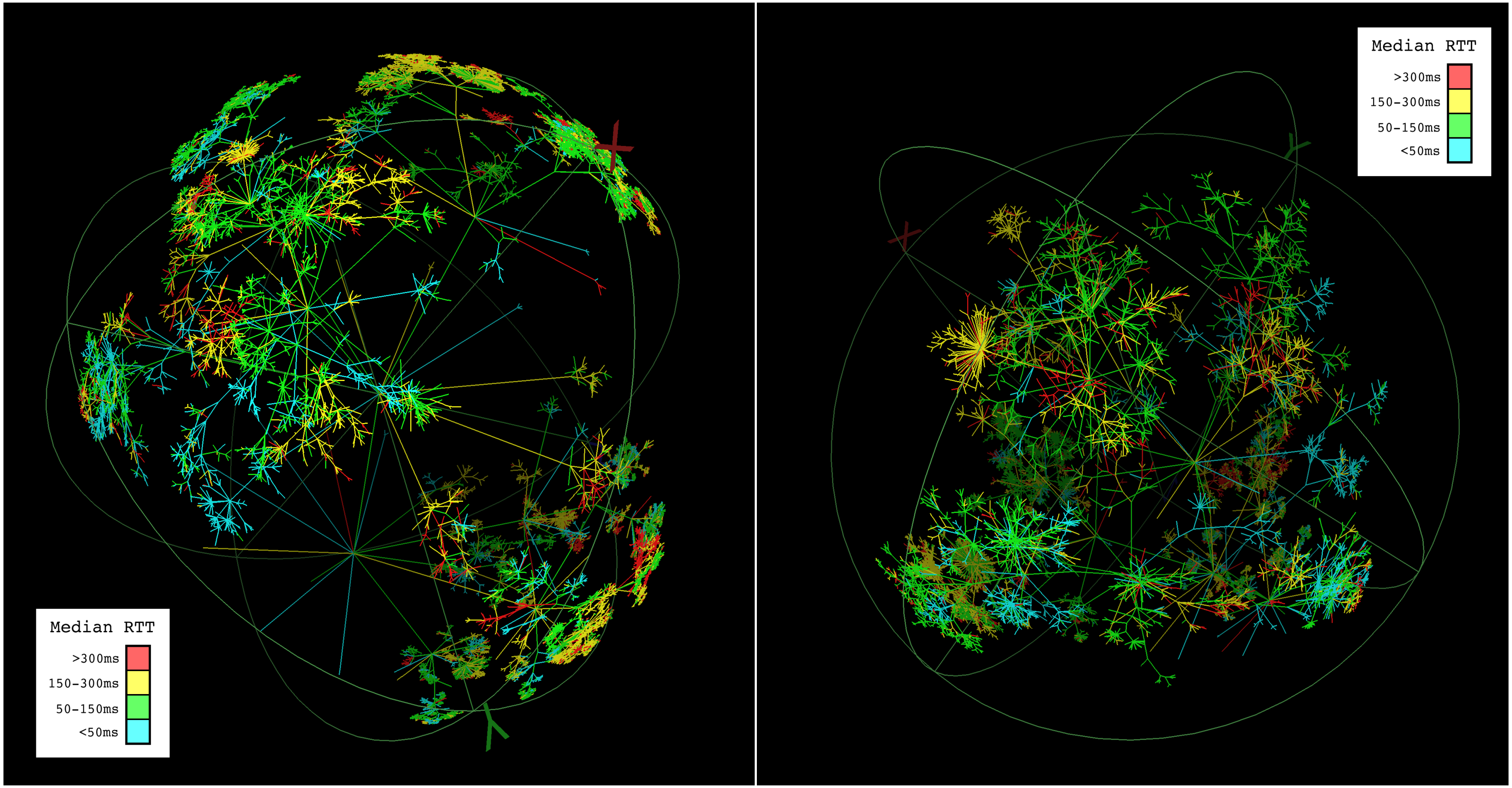
http://www.caida.org/home

Code Red infection



http://www.caida.org/home

Round-trip time measurements: 63,631 nodes and 63,630 links



http://www.caida.org/home

References

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

- Y. Li, H. J. Xing, Q. Hua, X.-Z. Wang, P. Batta, S. Haeri, and Lj. Trajkovic, "Classification of BGP anomalies using decision trees and fuzzy rough sets," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC 2013)*, San Diego, CA, October 2014, pp. 1331-1336.
- N. Al-Rousan, S. Haeri, and Lj. Trajkovic, "Feature selection for classification of BGP anomalies using Bayesian models," in *Proc. ICMLC 2012*, Xi'an, China, July 2012, pp. 140-147.
- N. Al-Rousan and Lj. Trajkovic, "Machine learning models for classification of BGP anomalies," in *Proc. IEEE Conf. High Performance Switching and Routing, HPSR 2012*, Belgrade, Serbia, June 2012, pp. 103-108.
- T. Farah, S. Lally, R. Gill, N. Al-Rousan, R. Paul, D. Xu, and Lj. Trajkovic, "Collection of BCNET BGP traffic," in *Proc. 23rd ITC*, San Francisco, CA, USA, Sept. 2011, pp. 322-323.
- S. Lally, T. Farah, R. Gill, R. Paul, N. Al-Rousan, and Lj. Trajkovic, "Collection and characterization of BCNET BGP traffic," in *Proc. 2011 IEEE Pacific Rim Conf. Communications, Computers and Signal Processing*, Victoria, BC, Canada, Aug. 2011, pp. 830-835.