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

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

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:
  - $r^{(m)}(k) = r(k),\ k \geqslant 0,\ m = 1, 2, \ldots, n,$

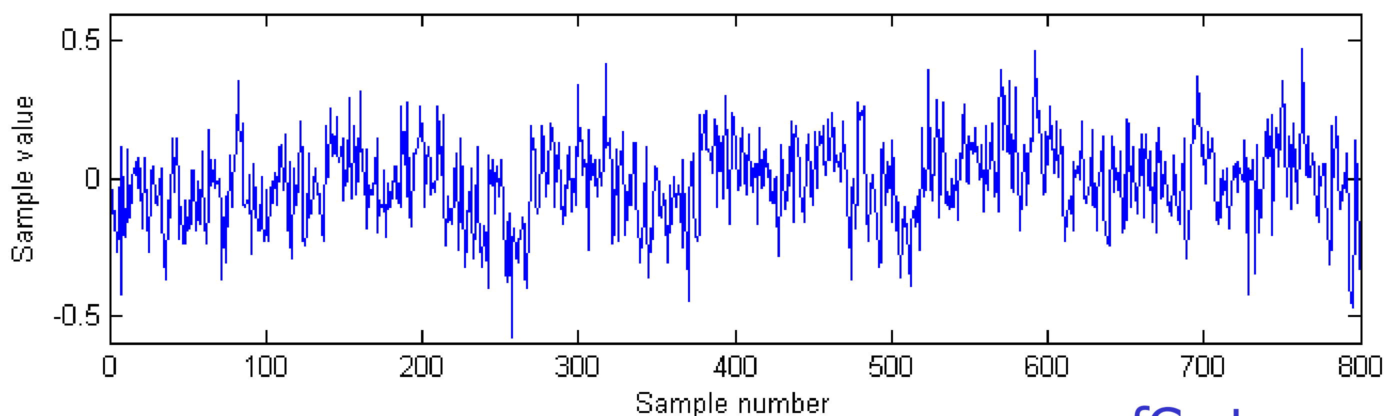
    where m is the level of aggregation
- Implications:
  - no natural length of bursts
  - bursts exist across many time scales
  - traffic does not become "smoother" when aggregated (unlike Poisson traffic)

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

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

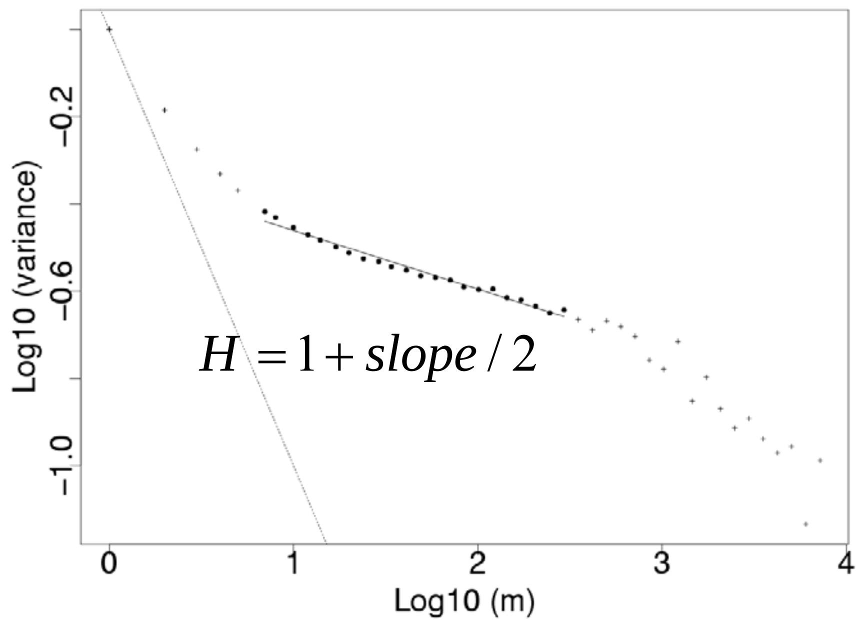


fGn trace

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 analysis

- Groups collection of objects into subsets (clusters):
  - resulting intra-cluster similarity is high while inter-cluster similarity is low
- The inter-cluster distance reflects dissimilarity between clusters:
  - Euclidean distance between two cluster centroids (mean value of objects in a cluster, viewed as cluster's center of gravity)
- The intra-cluster distance expresses coherent similarity of data in the same cluster:
  - average distance of objects from their cluster centroids
- Better clustering:
  - large inter-cluster and small intra-cluster distances

Clustering quality

- Overall clustering quality: defined as difference between minimum inter-cluster and maximum intra-cluster distances
  - larger indicator implies better overall clustering quality
- Silhouette coefficient (x):

  $$(b(x) - a(x)) / \max\{a(x), b(x)\}$$

  a(x) and b(x) are average distances between data point x and other data points in clusters A and B, respectively
  - independent of number of clusters K

Clustering algorithms

- Two approaches:
  - partitioning clustering (k-means)
  - hierarchical clustering
- Clustering tools:
  - k-means algorithm
  - AutoClass tool

- P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): theory and results," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., AAAI Press/MIT Press, 1996.
- L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

Clustering algorithms: k-means

- The k-means algorithm is commonly used for data clustering
- The algorithm is well-known for its simplicity and efficiency
- Based on the input parameter k, it partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high and the inter-cluster similarity is low
- Similarity of clusters is measured with respect to the mean value of the objects in a cluster (viewed as the cluster's center of gravity)

k-means: partitioning clustering

- Constructs k partitions of the data from n objects, where $k \leq n$
- Two constraints:
  - each cluster must contain at least one object
  - each object must belong to exactly one group
- Requires exhaustive enumeration of all possible combinations to find the optimal cluster solution

k-means clustering

- Generates k clusters from n objects
- Requires two inputs:
  - k: number of desired partitions
  - n objects
- Uses random placement of initial clusters
- Determines clustering results through an iteration technique to relocate objects to the most similar cluster:
  - similarity is defined as the distance between objects
  - objects that are closer to each other are more similar
- Computational complexity of O(nkt), where t is the maximum number of iterations

Finding number of clusters

- The number of clusters k is not known a priori
- k-means algorithm is repeated for different k values
- Number of clusters is found by comparing average SC value for various values of k:
  - average SC is calculated for all objects
  - the natural number of clusters k is found at the local maxima

SC: silhouette coefficient

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:
    - *AIC*
    - corrected *AICc*
  - Bayesian information criterion

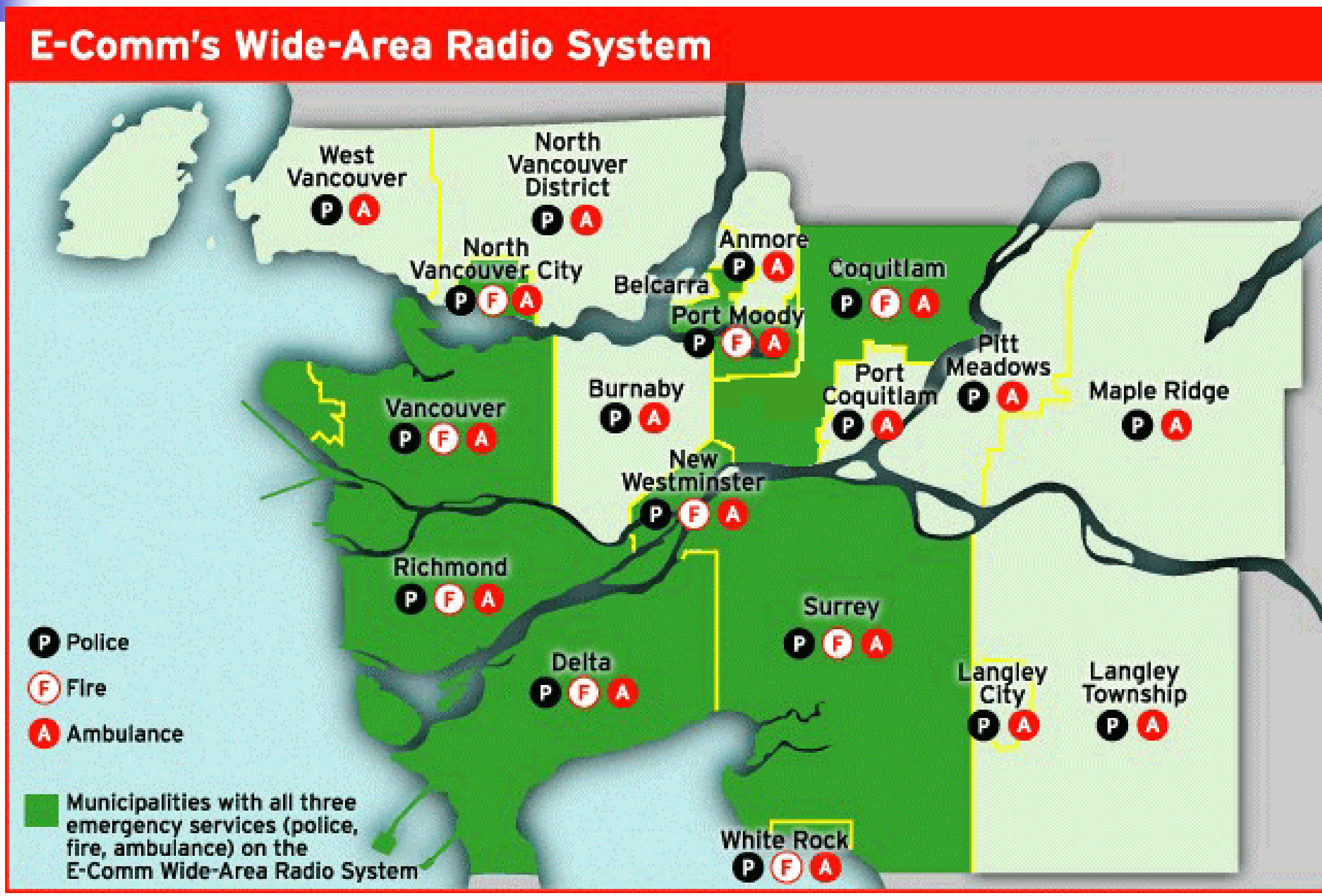
Roadmap

- 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

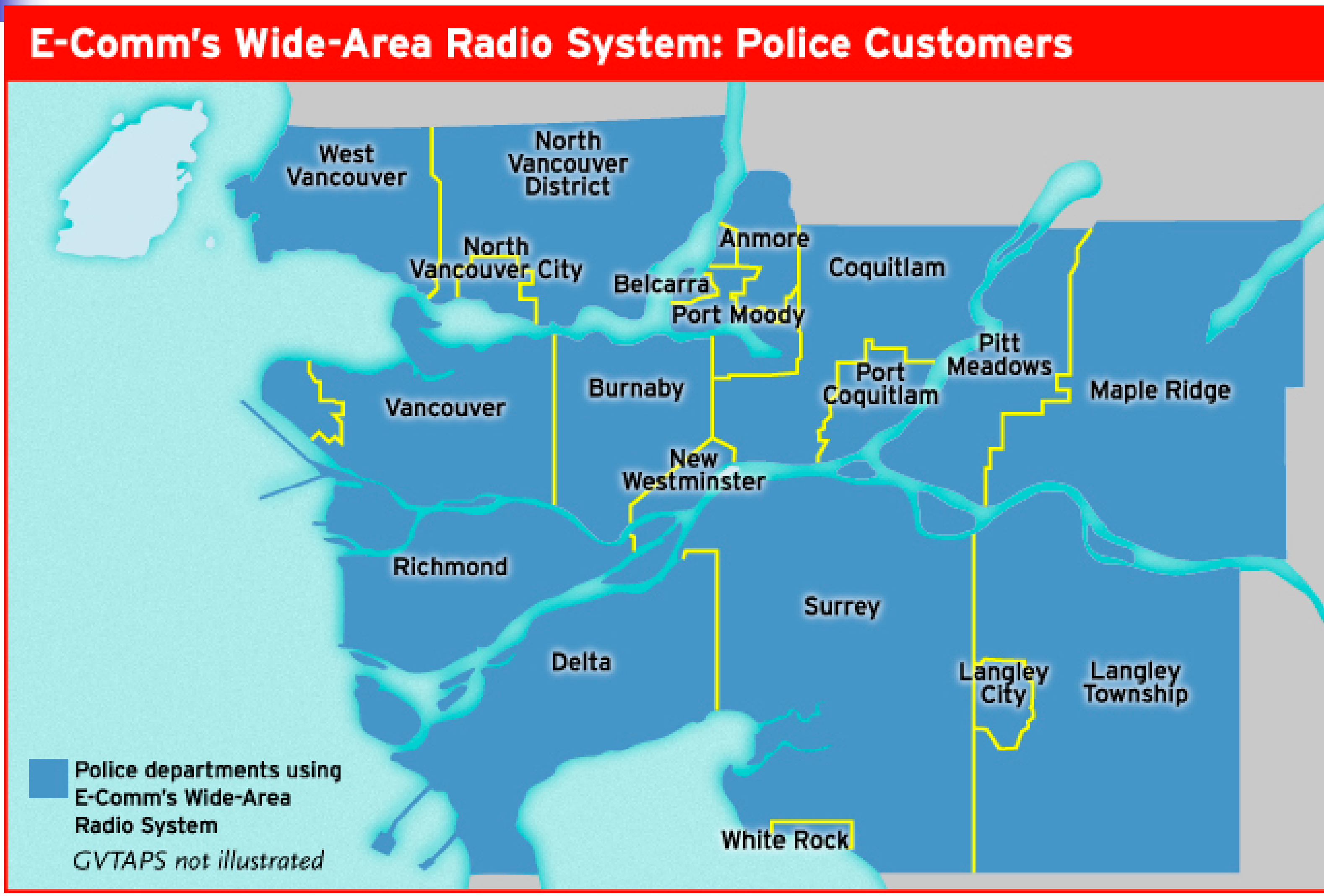
Case study: E-Comm network

- 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

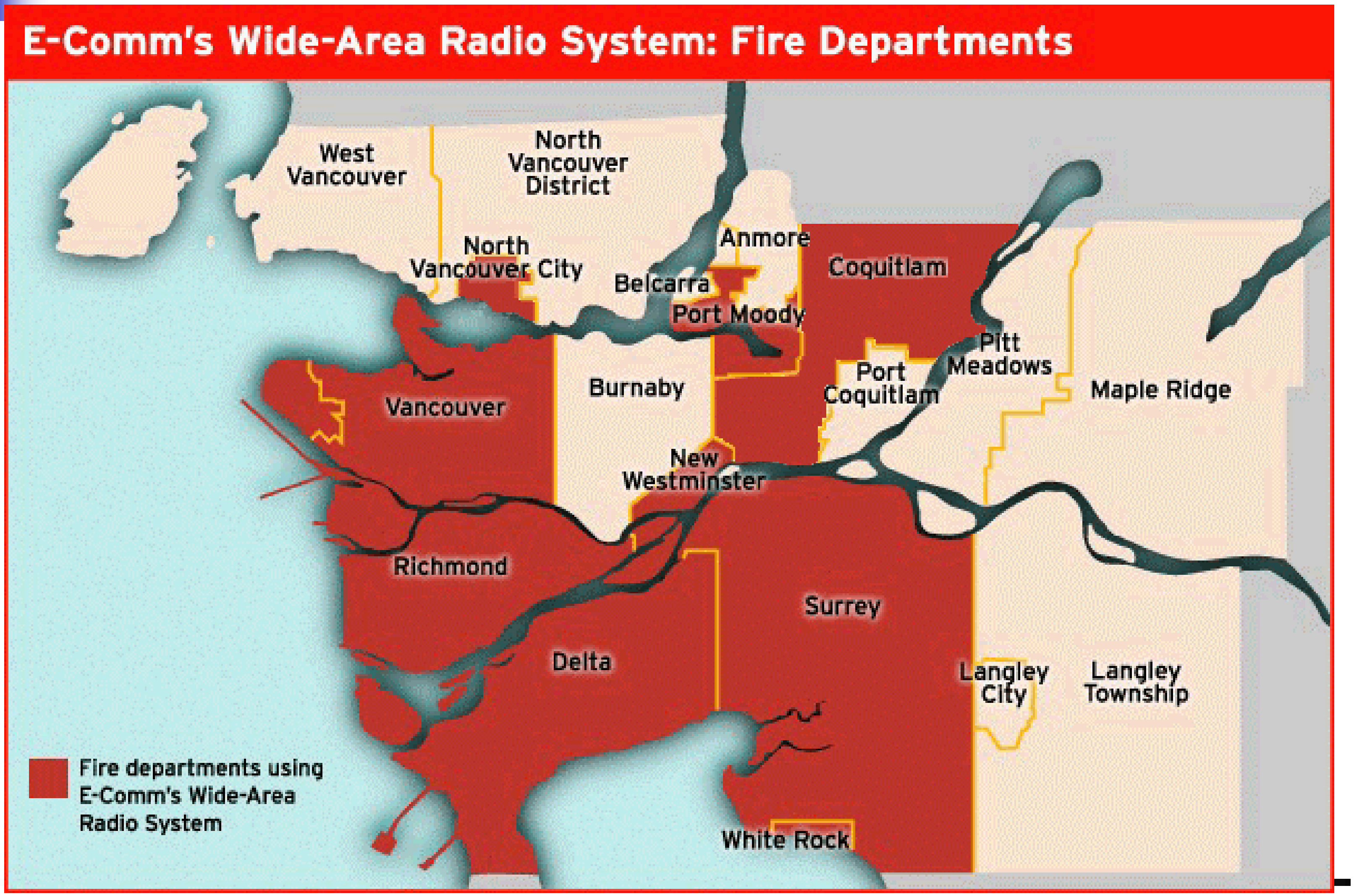
E-Comm network



E-Comm network

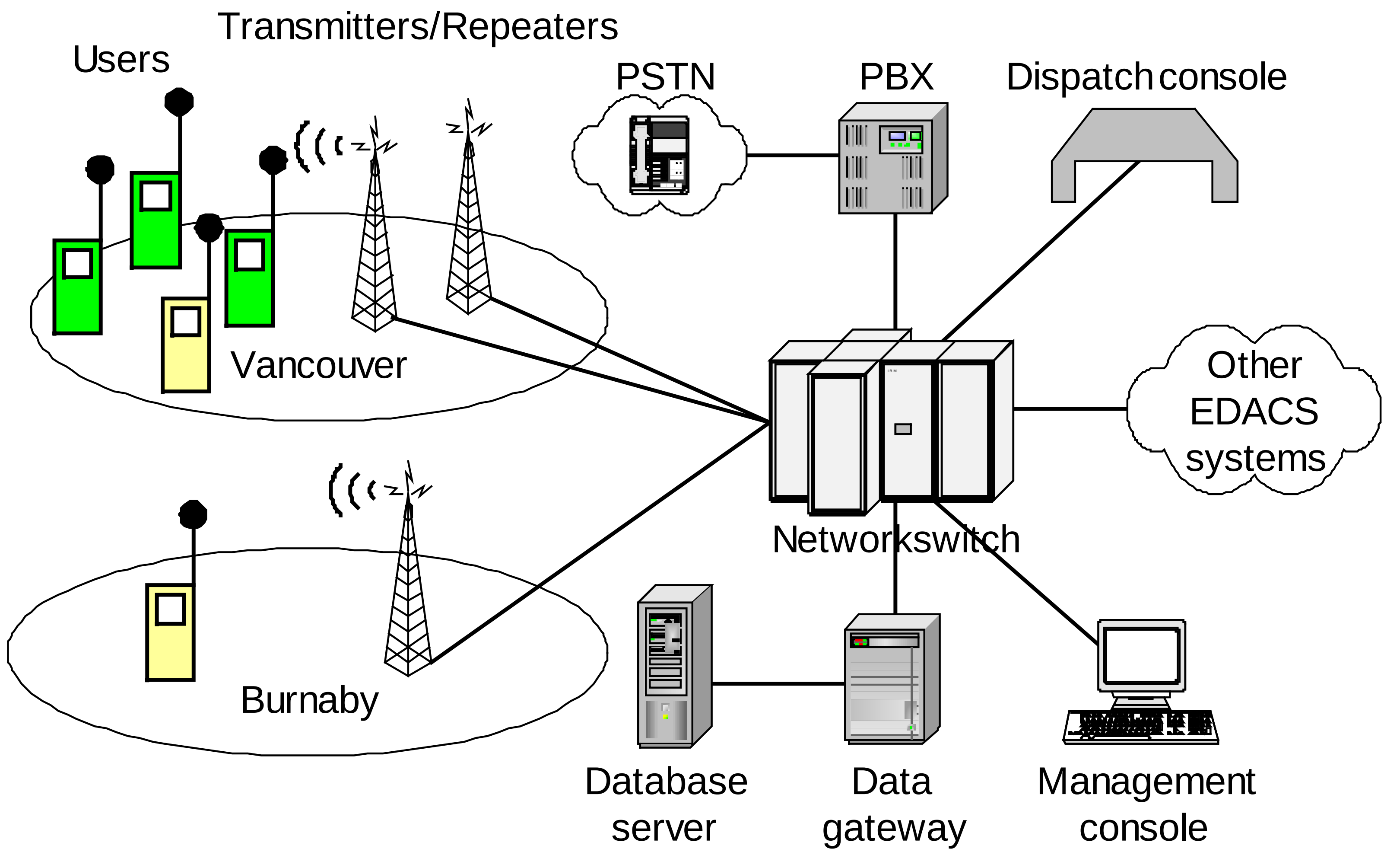


E-Comm network



E-Comm network

E-Comm network architecture



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)

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

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

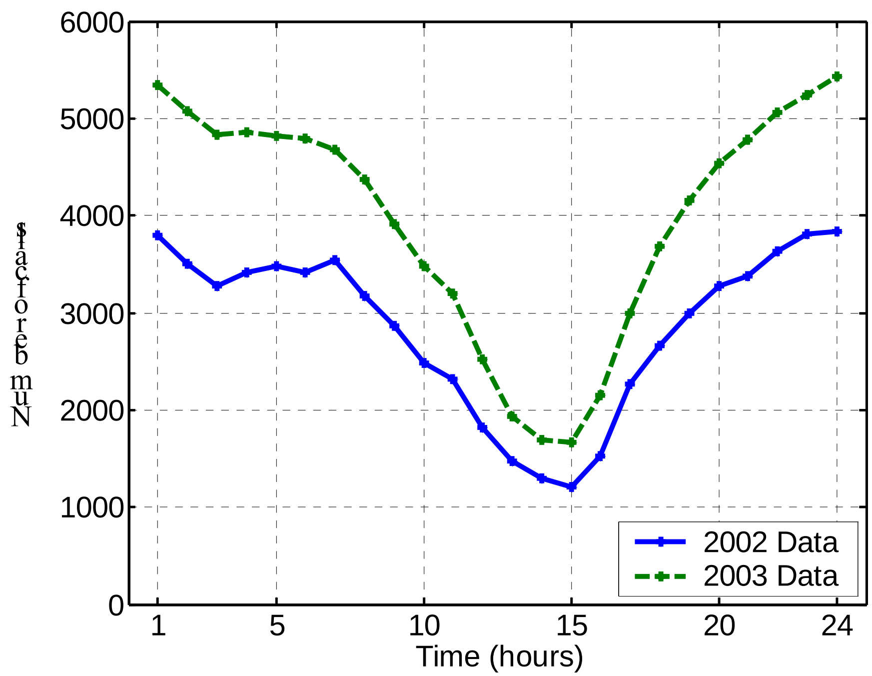
- Modeling and Performance Analysis of Public Safety Wireless Networks
- WarnSim: a simulator for public safety wireless networks (PSWN)
- Traffic data analysis
- Traffic modeling
- Simulation and prediction

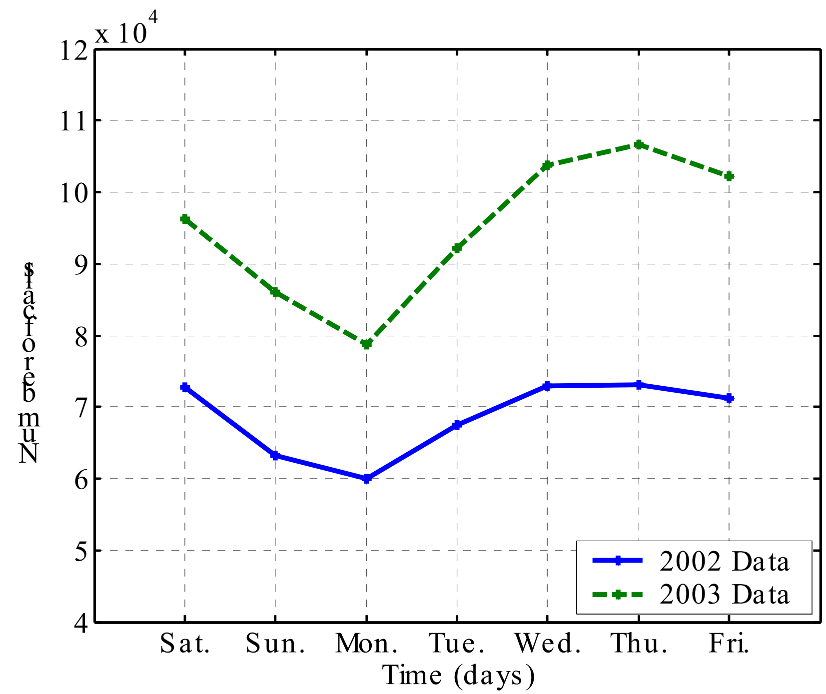
- N. Cackov, B. Vujičić, S. Vujičić, and Lj. Trajković, "Using network activity data to model the utilization of a trunked radio system," in *Proc. SPECTS2004*, San Jose, CA, July 2004, pp. 517-524.
- N. Cackov, J. Song, B. Vujičić, S. Vujičić, and Lj. Trajković, "Simulation of a public safety wireless networks: a case study," *Simulation*, vol. 81, no. 8, pp. 571-585, Aug. 2005.
- J. Song and Lj. Trajković, "Modeling and performance analysis of public safety wireless networks," in *Proc. IEEE IPCCC*, Phoenix, AZ, Apr. 2005, pp. 567-572.

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

- B. Vujičić, N. Cackov, S. Vujičić, and Lj. Trajković, "Modeling and characterization of traffic in public safety wireless networks," in *Proc. SPECTS 2005*, Philadelphia, PA, July 2005, pp. 214–223.

Erlang traffic models

Erlang B

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

Erlang C

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

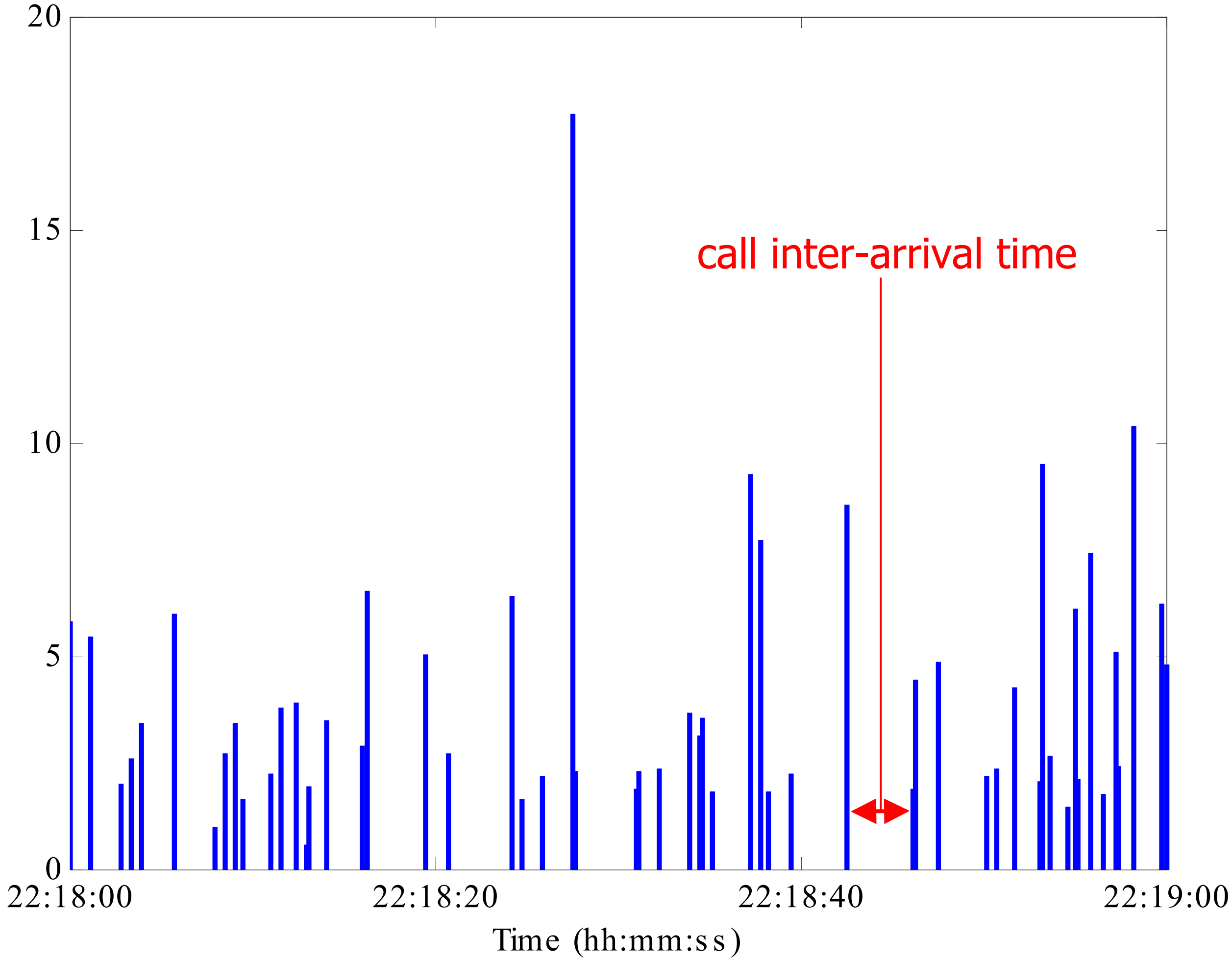
- P_B : probability of rejecting a call
- P_c : probability of delaying a call
- N : number of channels/lines
- A : total traffic volume

Hourly traces

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

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

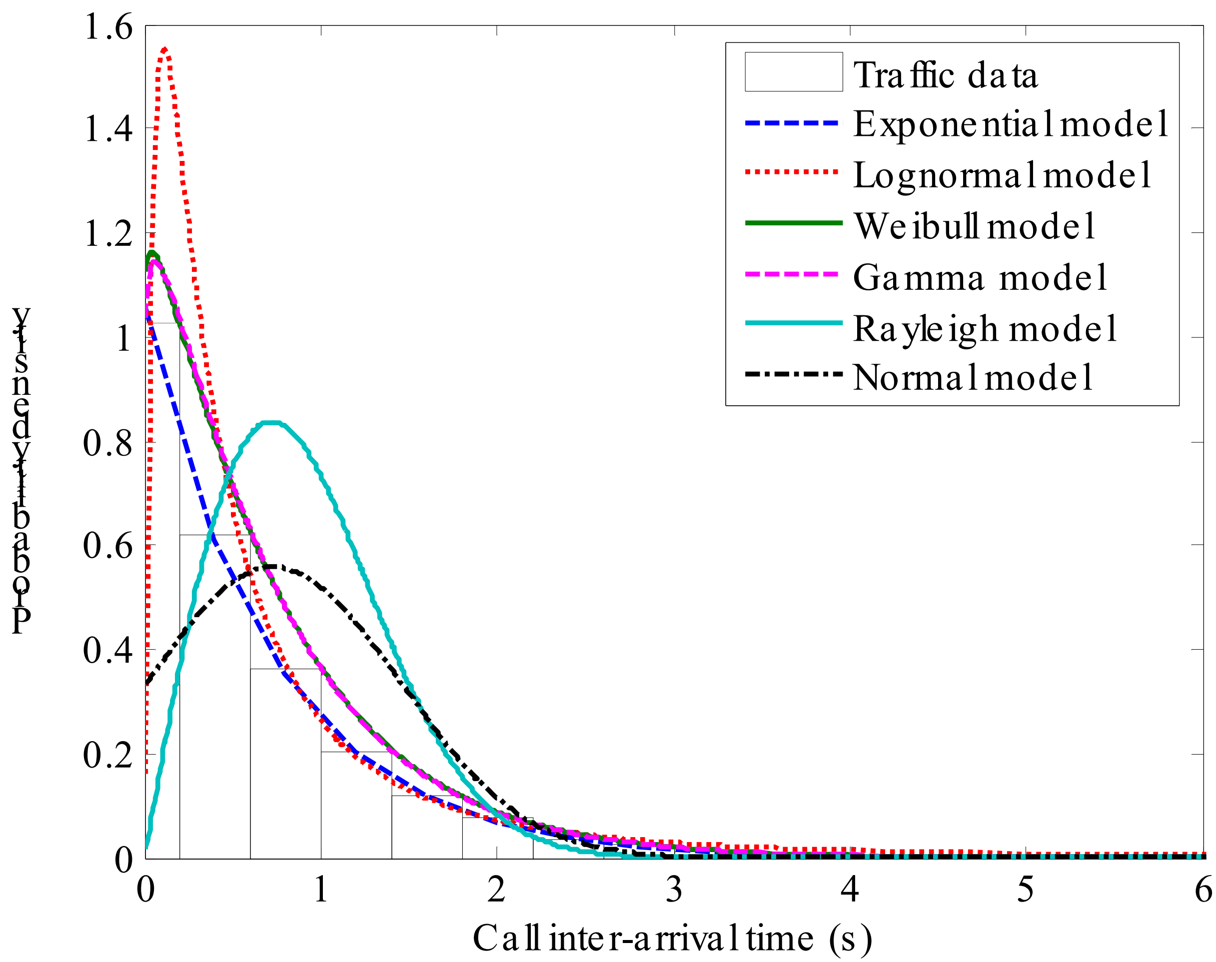
Example: March 26, 2003



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 times: K-S test results (2003 data)

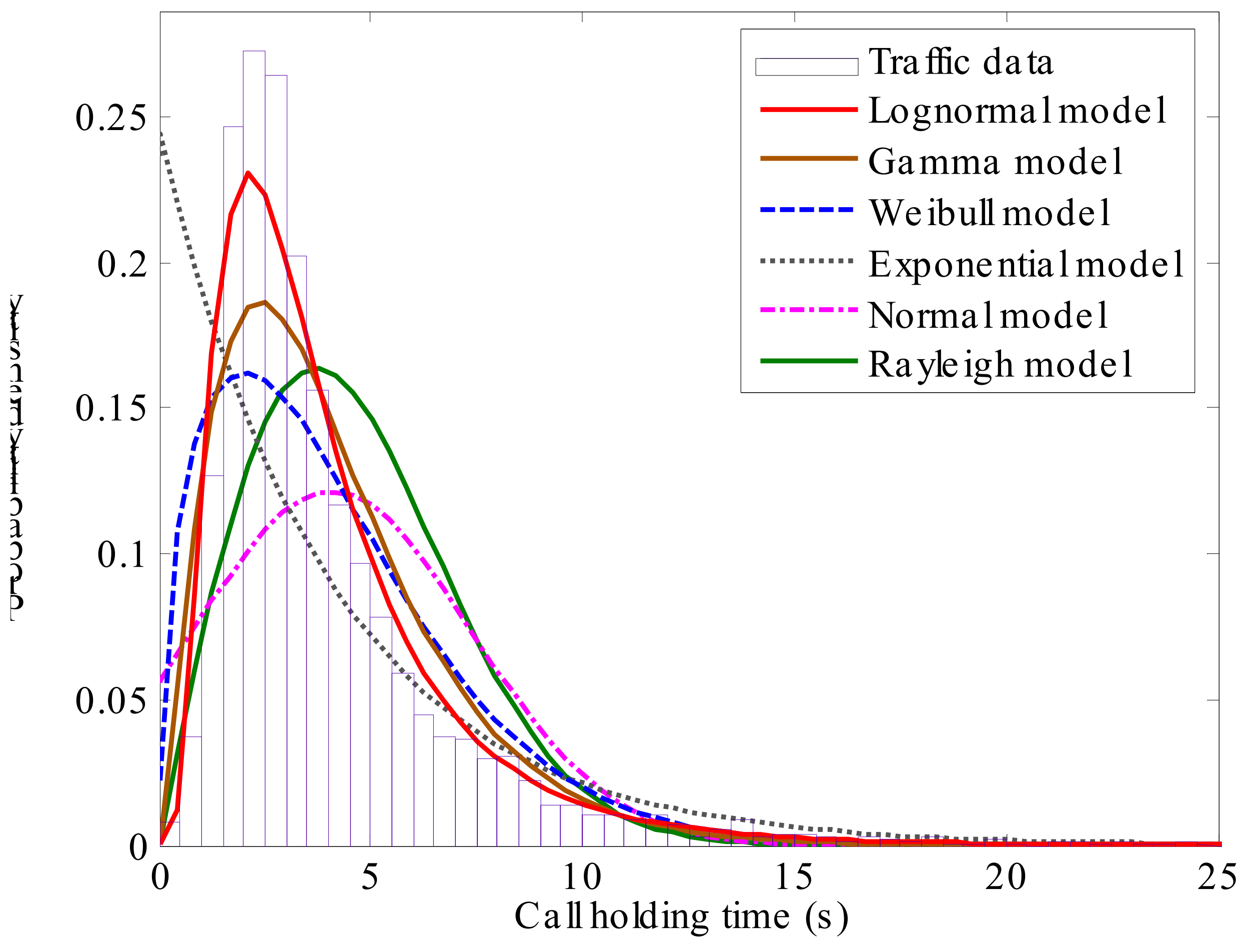
Distribution	Parameter	26.03.2003, 22:00–23:00	25.03.2003, 23:00–24:00	26.03.2003, 23:00–24:00	29.03.2003, 02:00–03:00	29.03.2003, 01:00–02:00
Exponential	h	1	1	0	1	1
	p	0.0027	0.0469	0.4049	0.0316	0.1101
	k	0.0283	0.0214	0.0137	0.0205	0.0185
Weibull	h	0	0	0	0	0
	p	0.4885	0.4662	0.2065	0.286	0.2337
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

Call inter-arrival times: estimates of H

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

2001		2002		2003	
Day/hour	*H*	Day/hour	*H*	Day/hour	*H*
02.11.2001 15:00–16:00	0.907	01.03.2002 04:00–05:00	0.679	26.03.2003 22:00–23:00	0.788
01.11.2001 00:00–01:00	0.802	01.03.2002 22:00–23:00	0.757	25.03.2003 23:00–24:00	0.832
02.11.2001 16:00–17:00	0.770	01.03.2002 23:00–24:00	0.780	26.03.2003 23:00–24:00	0.699
01.11.2001 19:00–20:00	0.774	01.03.2002 00:00–01:00	0.741	29.03.2003 02:00–03:00	0.696
02.11.2001 20:00–21:00	0.663	02.03.2002 00:00–01:00	0.747	29.03.2003 01:00–02:00	0.705

Call holding times: pdf candidates



Call holding times: estimates of H

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

2001		2002		2003	
Day/hour	*H*	Day/hour	*H*	Day/hour	*H*
02.11.2001 15:00–16:00	0.493	01.03.2002 04:00–05:00	0.490	26.03.2003 22:00–23:00	0.483
01.11.2001 00:00–01:00	0.471	01.03.2002 22:00–23:00	0.460	25.03.2003 23:00–24:00	0.483
02.11.2001 16:00–17:00	0.462	01.03.2002 23:00–24:00	0.489	26.03.2003 23:00–24:00	0.463 *
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

	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

Traffic prediction

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

- H. Chen and Lj. Trajković, "Trunked radio systems: traffic prediction based on user clusters," in *Proc. IEEE ISWCS 2004*, Mauritius, Sept. 2004, pp. 76-80.
- B. Vujičić, L. Chen, and Lj. Trajković, "Prediction of traffic in a public safety network," in *Proc. ISCAS 2006*, Kos, Greece, May 2006, pp. 2637-2640.

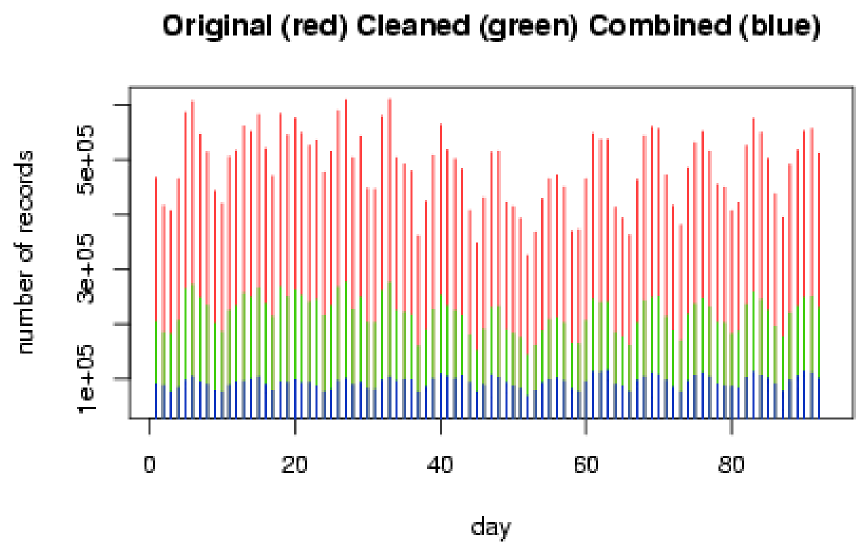
Traffic data: preprocessing

- 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

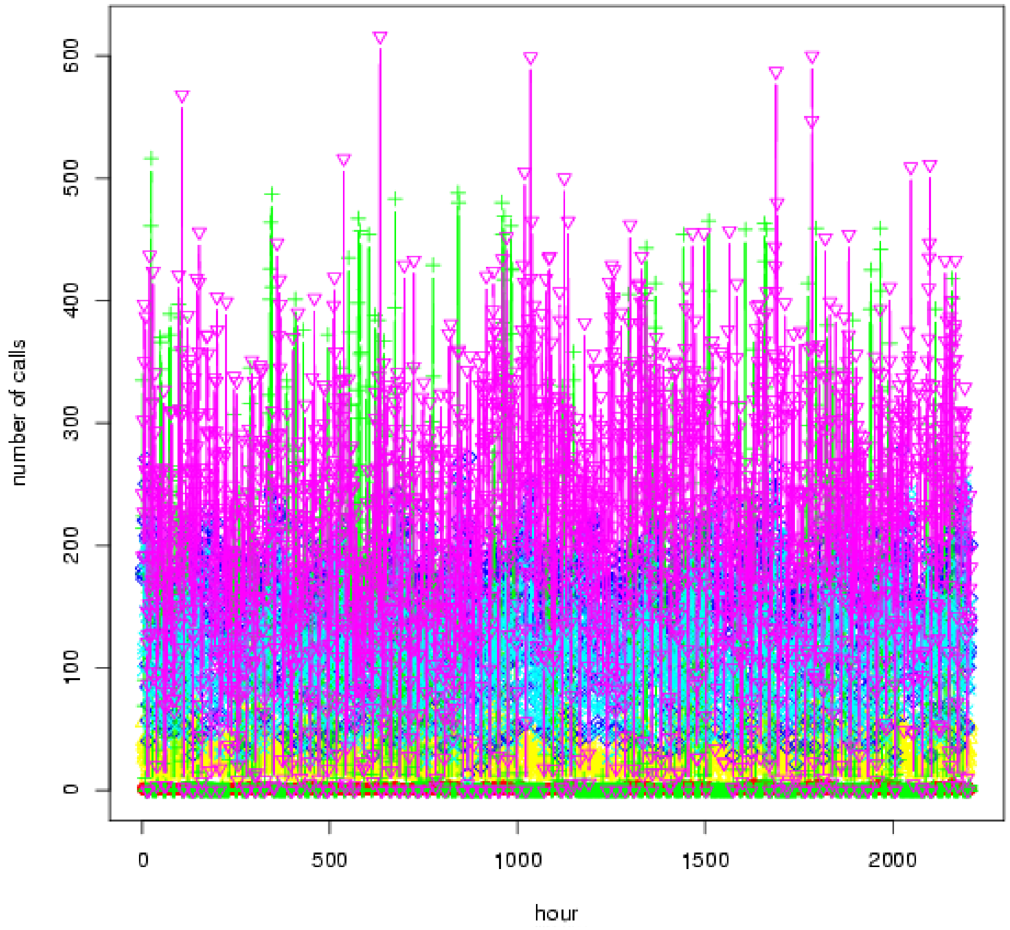
Data preparation

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

- 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

Cluster size	Minimum number of calls	Maximum number of calls	Average number of calls	Total number of calls	Total number of calls (%)
17	0-6	352-700	94-208	5,091,695	59
31	0-3	135-641	17-66	2,261,055	26
569	0	1-1613	0-16	1,310,836	15

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 (nmse) is used to measure prediction quality by comparing deviation between predicted and observed data
- The nmse of forecast is equal to ratio of normalized sum of variance of forecast to squared bias of forecast
- Smaller values of nmse indicate better prediction model

Prediction: based on the aggregate traffic

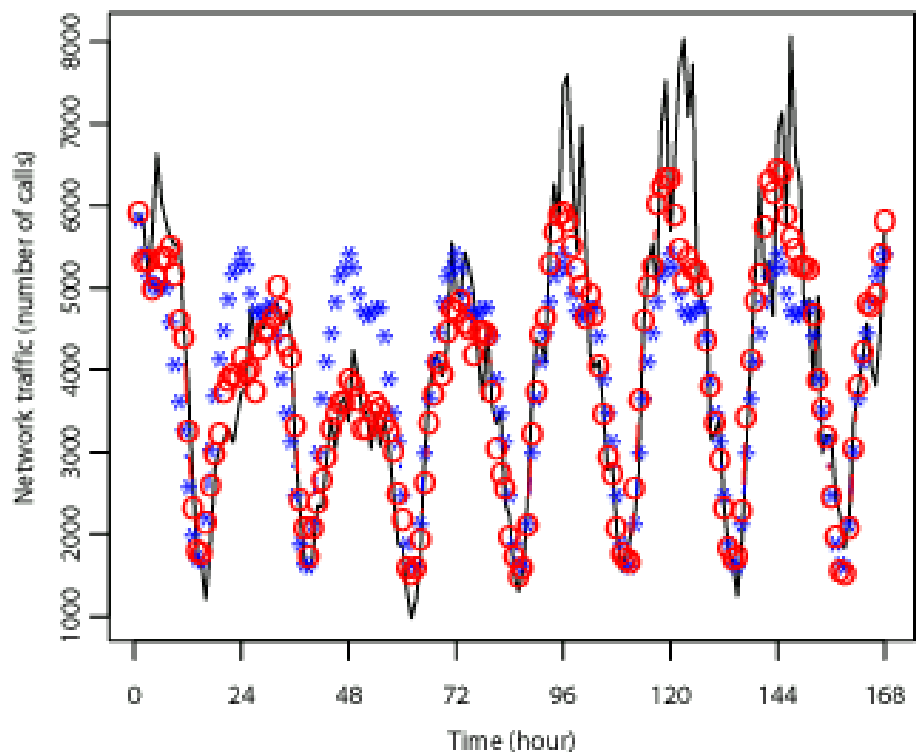
No.	p	d	q	P	D	Q	S	m	n	nmse
A1	2	0	9	0	1	1	24	1512	672	0.3790
A2	2	0	1	0	1	1	24	1512	672	0.3803
A3	2	0	9	0	1	1	168	1512	672	0.1742
A4	2	0	1	0	1	1	168	1512	672	0.1732
B1	2	0	9	0	1	1	24	1680	168	0.3790
B2	2	0	1	0	1	1	24	1680	168	0.4079
B3	2	0	9	0	1	1	168	1680	168	0.1736
B4	2	0	1	0	1	1	168	1680	168	0.1745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

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 $(2, 0, 9) \times (0, 1, 1)_{24 \text{ and } 168}$
  - SARIMA $(2, 0, 1) \times (0, 1, 1)_{24 \text{ and } 168}$
- 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

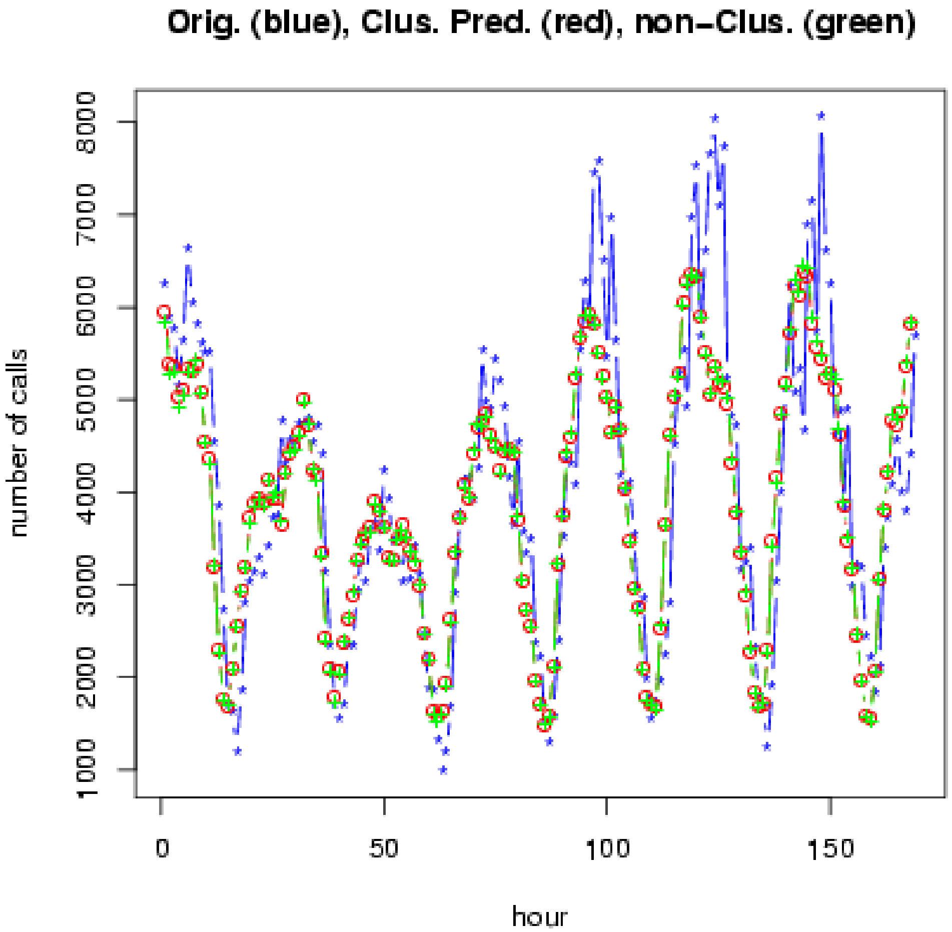
Prediction of 168 hours of traffic based on 1,680 past hours: sample



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 $(1,0,1)x(0,1,1)_{168}$

* observation

* prediction without clustering

o 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

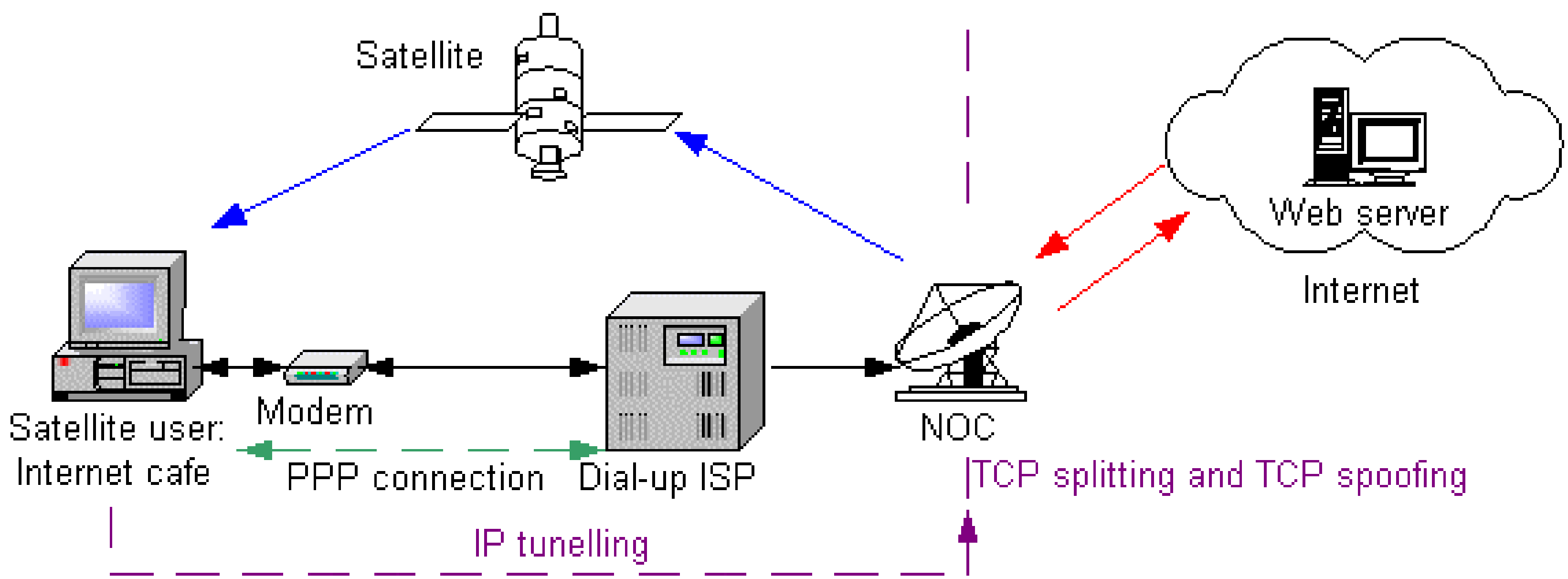
- Analysis of network traffic:
  - characteristics of TCP connections
  - network traffic patterns
  - statistical and cluster analysis of traffic
  - anomaly detection:
    - statistical methods
    - wavelets
    - principle component analysis

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 geosynchrous 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

- Q. Shao and Lj. Trajkovic, "Measurement and analysis of traffic in a hybrid satellite-terrestrial network," *Proc. SPECTS 2004*, San Jose, CA, July 2004, pp. 329–336.

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

- Scans and worms:
  - packets are sent to probe network hosts
  - used to discover and exploit resources
- Denial of service:
  - large number of packets is directed to a single destination
  - makes a host incapable of handling incoming connections or exhausts available bandwidth along paths to the destination

Network anomalies

- Flash crowd:
  - high volume of traffic is destined to a single destination
  - caused by breaking news or availability of new software
- Traffic shift:
  - redirection of traffic from one set of paths to another
  - caused by route changes, link unavailability, or network congestion

Network anomalies

- Alpha traffic:
  - unusually high volume of traffic between two endpoints
  - caused by file transfers or bandwidth measurements
- Traffic volume anomalies:
  - significant deviation of traffic volume from usual daily or weekly patterns
  - classified as:
    - outages: caused by unavailable links, crashed servers, or routing problems
    - short term increases in demand: caused by short term events such as holiday traffic
  - involve multiple sources and destinations

Billing records

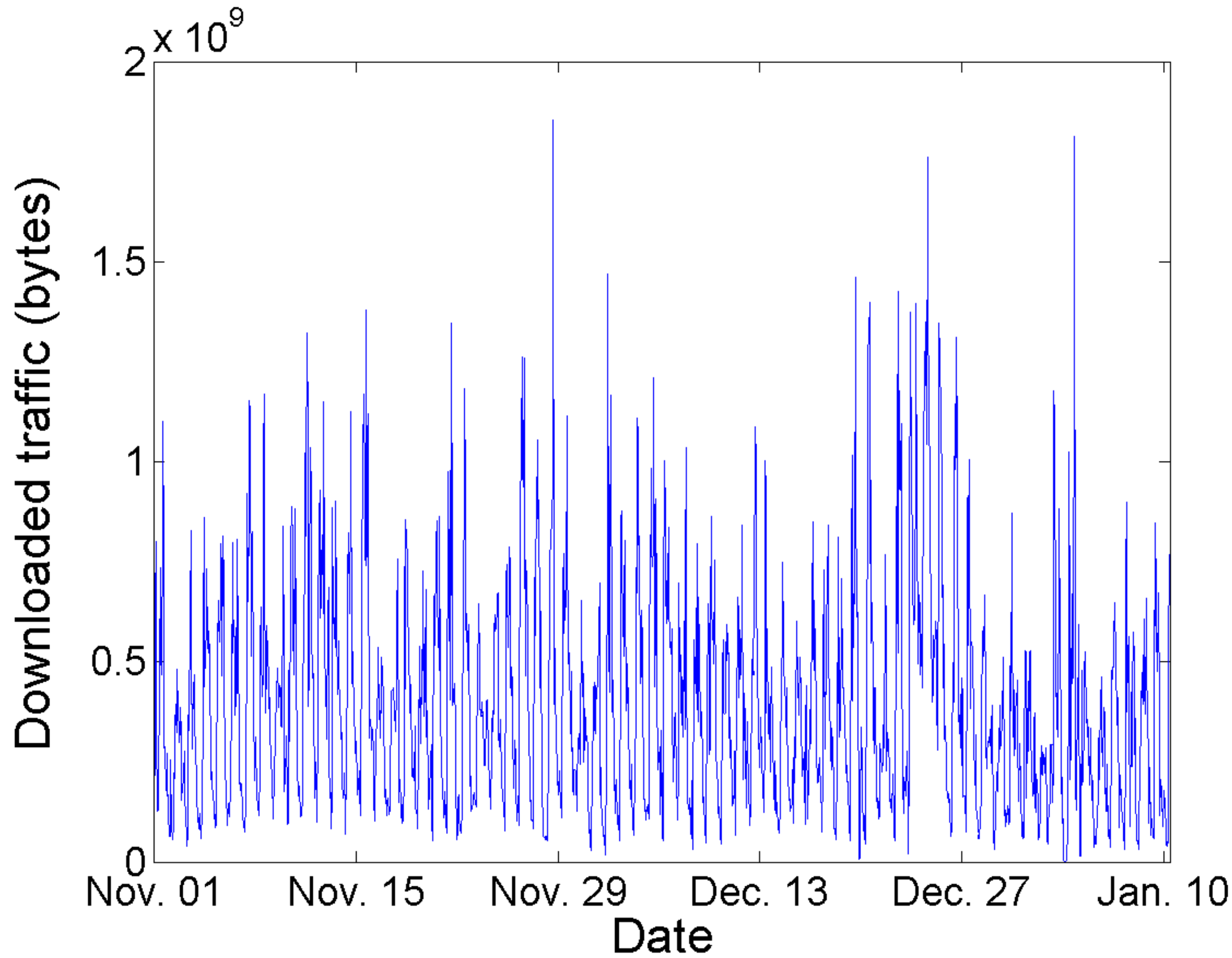
- Records were collected during the continuous period from 23:00 on Oct. 31, 2002 to 11:00 on Jan. 10, 2003
- Each file contains the hourly traffic summary for each user
- Fields of interests:
  - SiteID (user identification)
  - Start (record start time)
  - CTxByt (number of bytes downloaded by a user)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

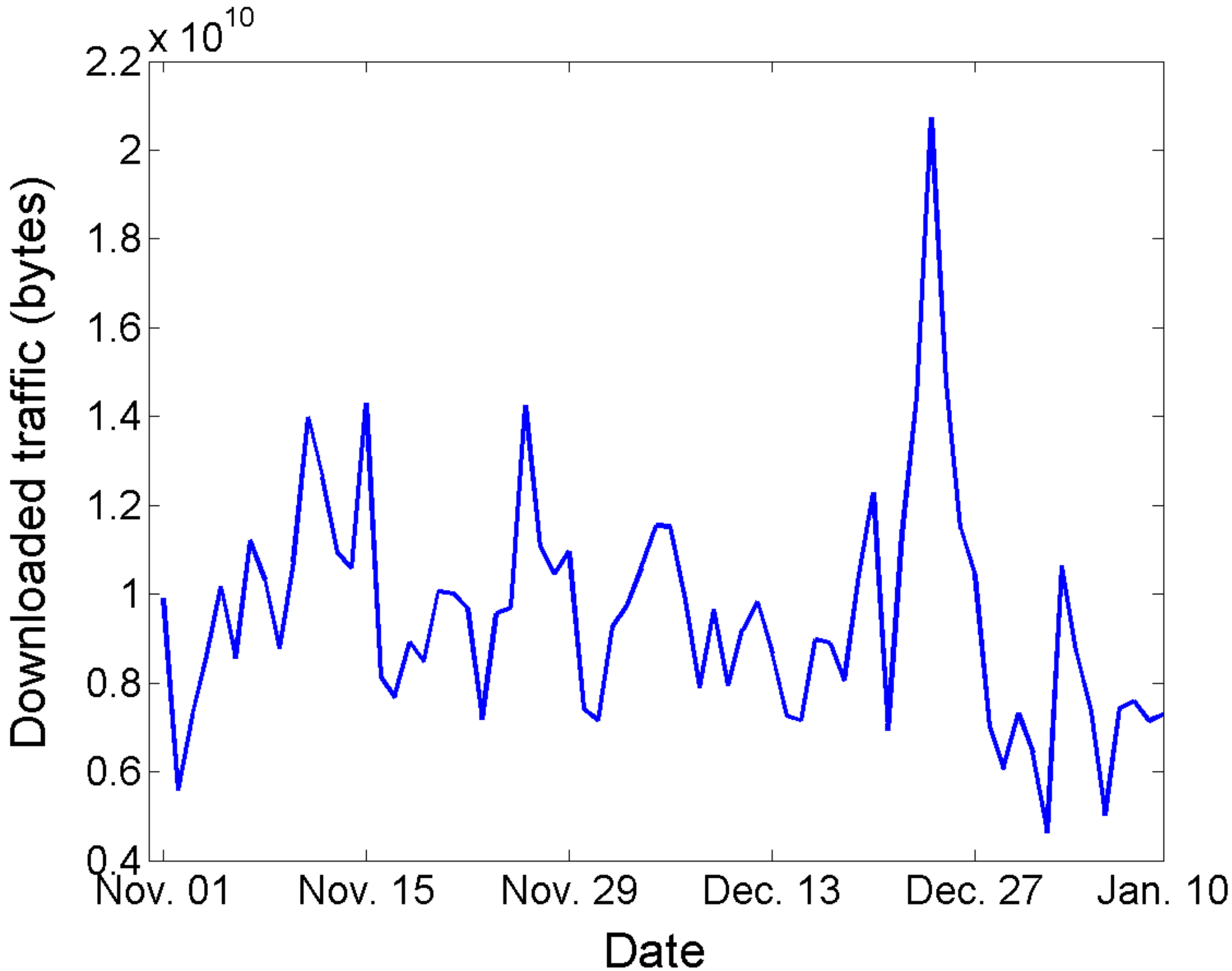
Billing records: characteristics

- 186 unique SiteIDs
- Daily and weekly cycles:
  - lower traffic volume on weekends
  - daily cycle starts at 7 AM, rises to three daily maxima at 11 AM, 3 PM, and 7 PM, then decrease monotonically until 7 AM
- Highest daily traffic recorded on Dec. 24, 2002
- Outage occurred on Jan. 3, 2003

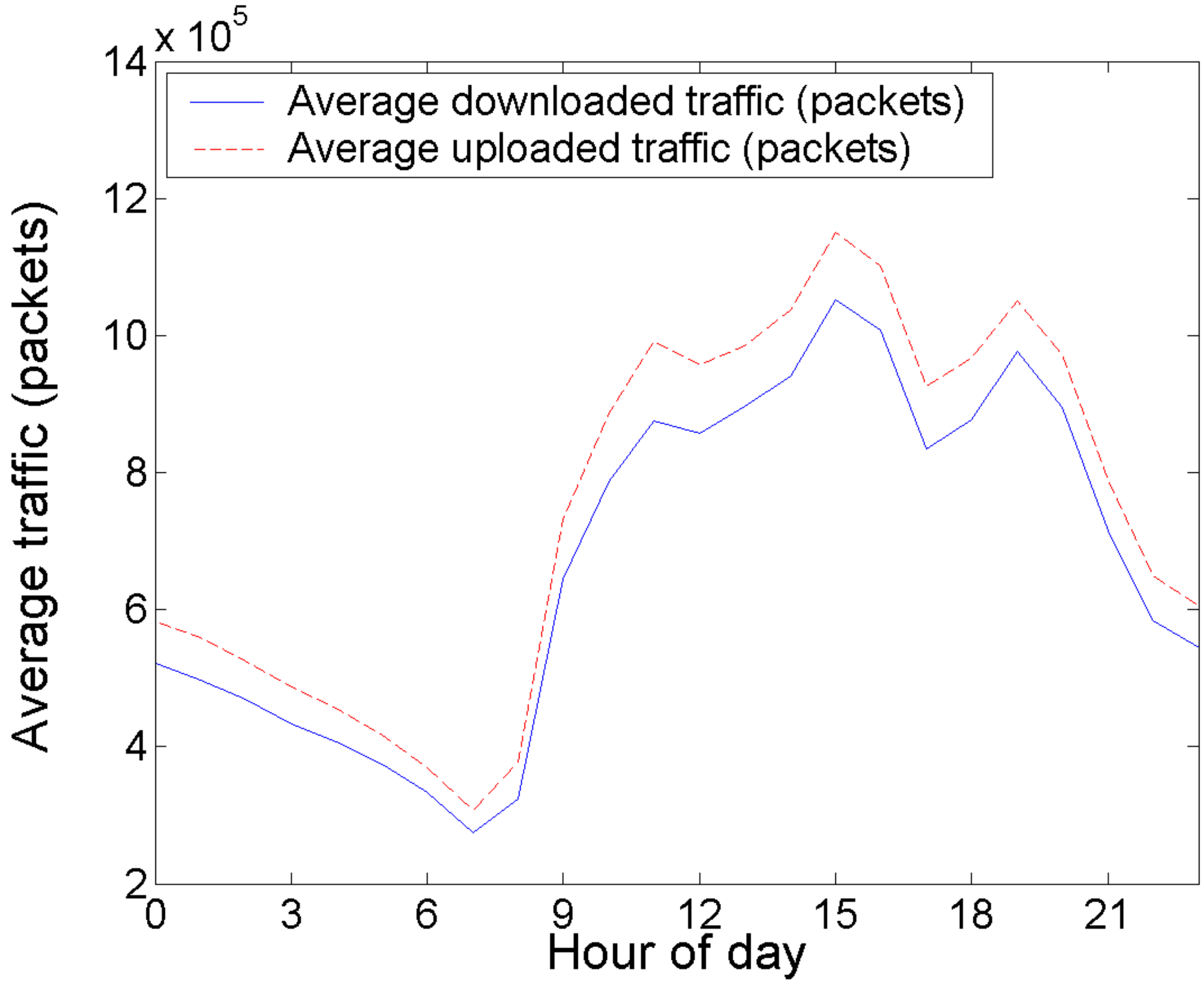
Aggregated hourly traffic



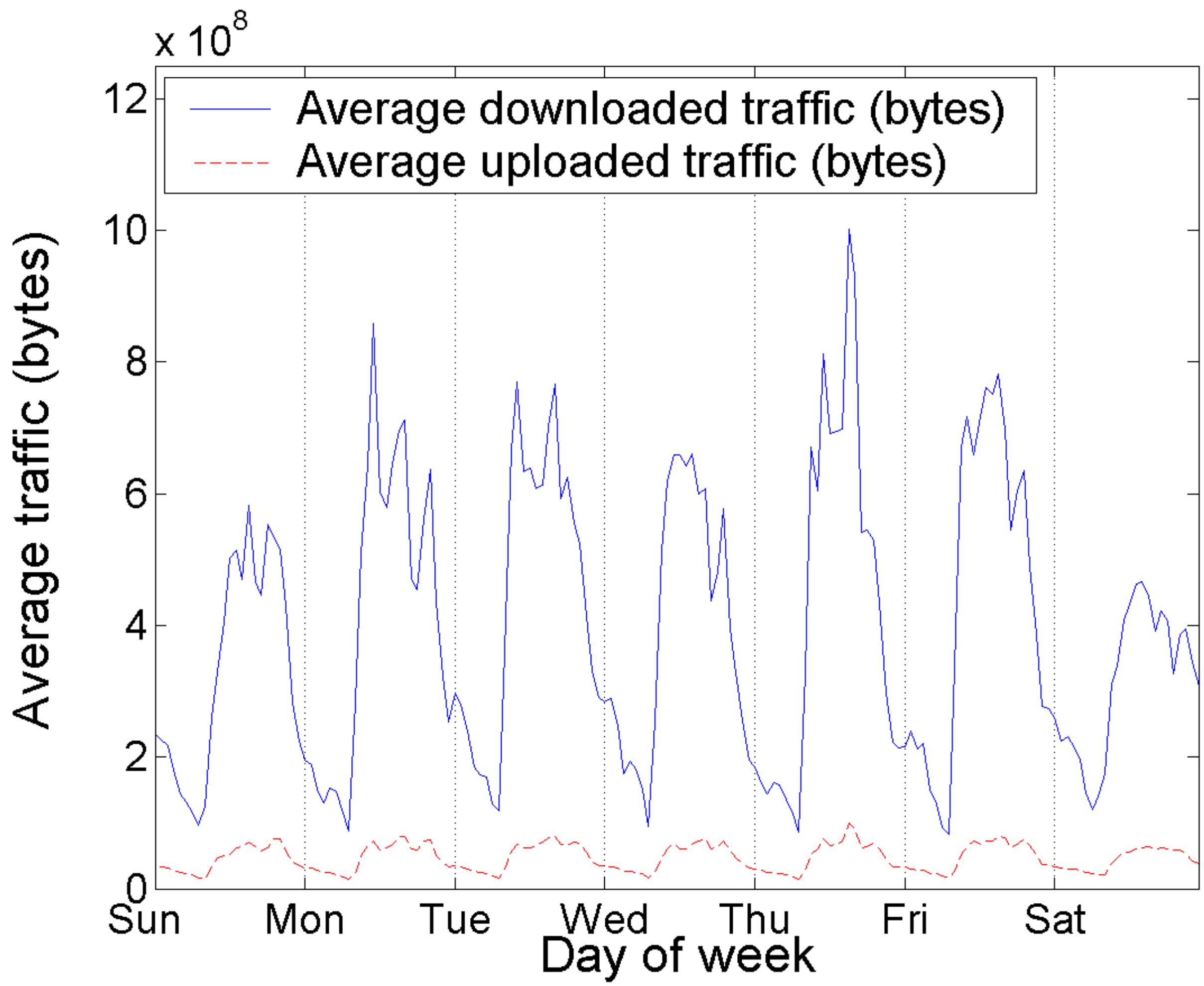
Aggregated daily traffic



Daily diurnal traffic: average downloaded bytes



Weekly traffic: average downloaded bytes



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)

k-means: clustering results

- Natural number of clusters is k=3 for downloaded and uploaded bytes
- Most users belong to the group with small traffic volume
- For k=3:
  - 159 users in group 1 (average 0.0–16.8 MB downloaded per hour)
  - 24 users in group 2 (average 16.8–70.6 MB downloaded per hour)
  - 3 users in group 3 (average 70.6–110.7 MB downloaded per hour)

Refinement: three most common traffic patterns

- Idle users:
  - rarely download/upload traffic
  - represented by zero traffic
- Active users:
  - download/upload traffic for more than 18 hours a day
  - represented by traffic over 24 hours each day
- Semi-active users:
  - download/upload traffic for 8–12 hours a day
  - represented by a cycle of 10 hours ACTIVE/14 hours IDLE cycle for each day

Refinement: clustering results

Traffic pattern	Number of users
Idle	162
Active	16
Semi-active	8
Total number of users	186

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 a 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

TCP flag	Packet count	% of Total
SYN only	19,050,849	48.500
RST only	7,440,418	18.900
FIN only	12,679,619	32.300
*SYN+FIN	408	0.001
*RST+FIN (no PSH)	85,571	0.200
*RST+PSH (no FIN)	18,111	0.050
*RST+FIN+PSH	8,329	0.020
*Total number of packets with invalid TCP flag combinations	112,419	0.300
Total packet count	39,283,305	100.000

Large number of TCP resets

- Connections are terminated by either TCP FIN or TCP RST:
  - 12,679,619 connections were terminated by FIN (63%)
  - 7,440,418 connections were terminated by RST (37%)
- Large number of TCP RST indicates that connections are terminated in error conditions
- TCP RST is employed by Microsoft Internet Explorer to terminate connections instead of TCP FIN

M. Arlitt and C. Williamson, "An analysis of TCP reset behaviour on the Internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 1, pp. 37–44, Jan. 2005.

UDP and TCP port scans

- UDP port scans are found on UDP port 137 (NETBEUI)
- TCP port scans are found on these TCP ports:
  - 80 Hypertext transfer protocol (HTTP)
  - 139 NETBIOS extended user interface (NETBEUI)
  - 434 HTTP over secure socket layer (HTTPS)
  - 1433 Microsoft structured query language (MS SQL)
  - 27374 Subseven trojan
- No HTTP(S) servers were active in the ChinaSat network
- MSSQL vulnerability was discovered on Oct. 2002, which may be the cause of scans on TCP port 1433
- The Subseven trojan is a backdoor program used with malicious intents

TCP: transport control protocol
UDP: user defined protocol

UDP port scans originating from the ChinaSat network

192.168.2.30:137 - 195.x.x.98:1025
192.168.2.30:137 - 202.x.x.153:1027
192.168.2.30:137 - 210.x.x.23:1035
192.168.2.30:137 - 195.x.x.42:1026
192.168.2.30:137 - 202.y.y.226:1026
192.168.2.30:137 - 218.x.x.238:1025
192.168.2.30:137 - 202.y.y.226:1025
192.168.2.30:137 - 202.y.y.226:1027
192.168.2.30:137 - 202.y.y.226:1028
192.168.2.30:137 - 202.y.y.226:1029
192.168.2.30:137 - 202.y.y.242:1026
192.168.2.30:137 - 61.x.x.5:1028
192.168.2.30:137 - 219.x.x.226:1025
192.168.2.30:137 - 213.x.x.189:1028
192.168.2.30:137 - 61.x.x.193:1025
192.168.2.30:137 - 202.y.y.207:1028
192.168.2.30:137 - 202.y.y.207:1025
192.168.2.30:137 - 202.y.y.207:1026
192.168.2.30:137 - 202.y.y.207:1027
192.168.2.30:137 - 64.x.x.148:1027

- Client (192.168.2.30) source port (137) scans external network addresses at destination ports (1025-1040):
  - > 100 are recorded within a three-hour period
  - targeted IP addresses are variable
  - multiple ports are scanned per IP
  - may correspond to Bugbear, OpaSoft, or other worms

UDP port scans direct to the ChinaSat network

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 - 192.168.1.5:137
210.x.x.23:1035 - 192.168.1.210:137
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

- 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

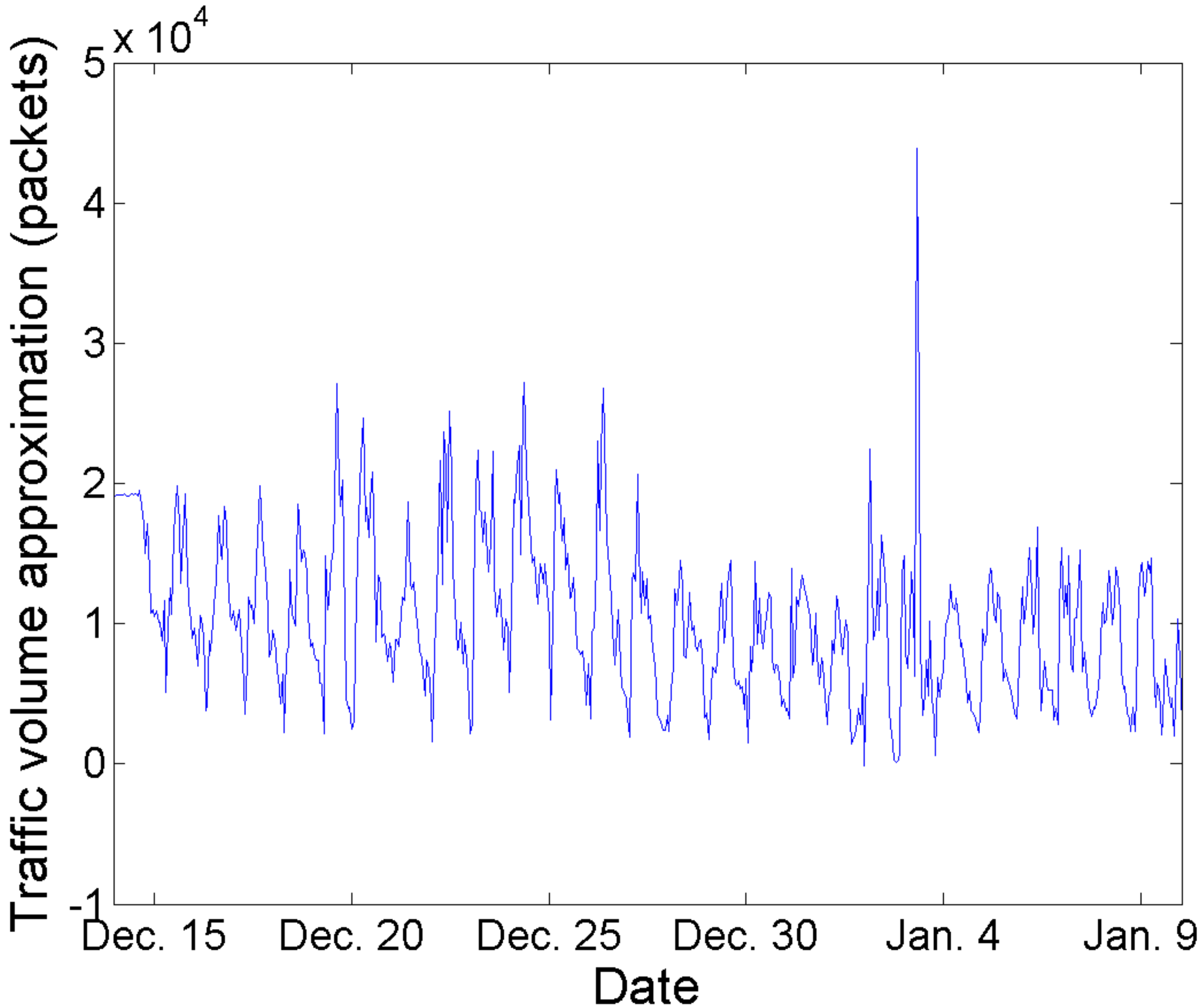
Detection of traffic volume anomalies using wavelets

- 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

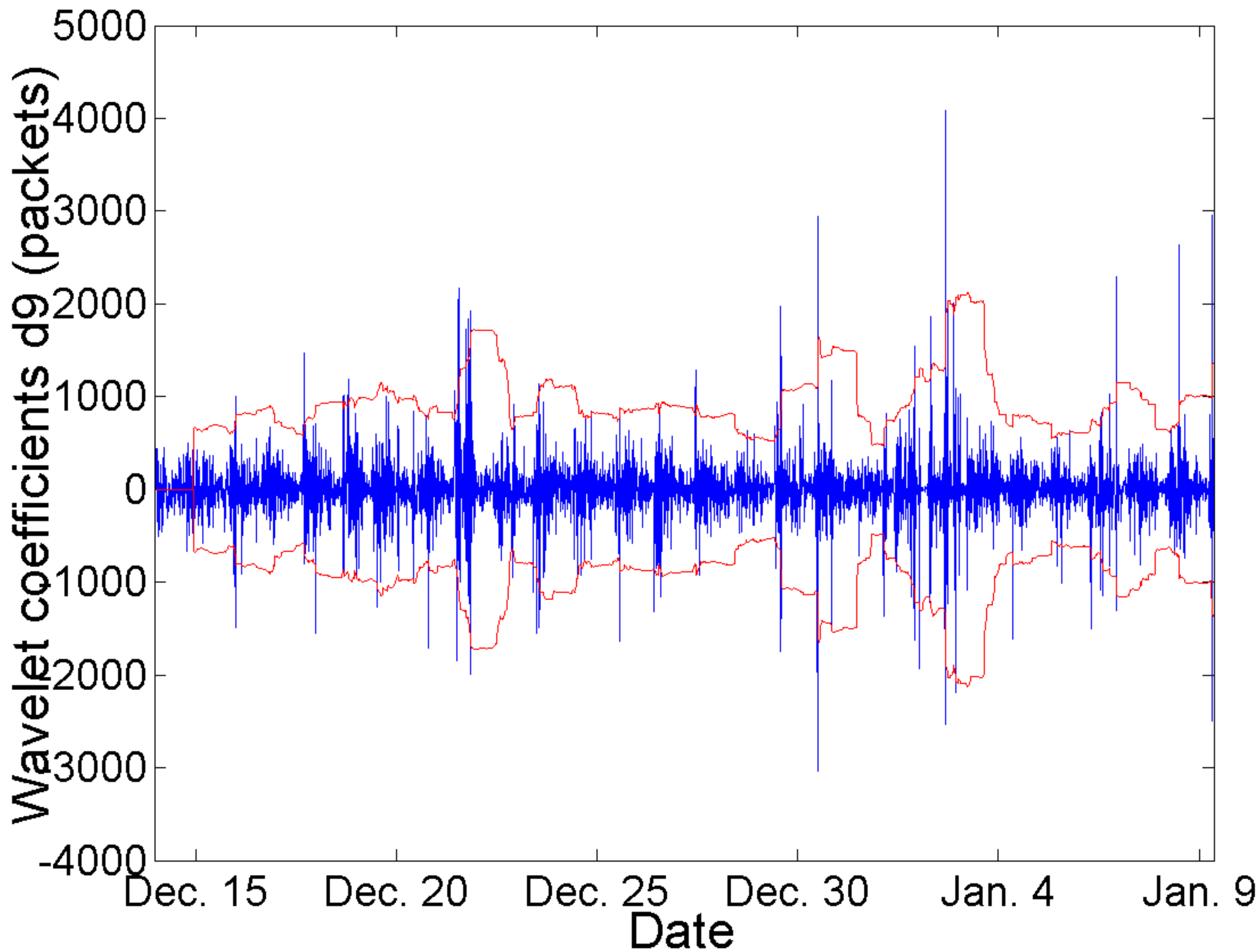
Detection of traffic volume anomalies using wavelets

- 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 (σ) of the wavelet coefficients in each window
  - identified when wavelet coefficients lie outside the $\pm 3\sigma$ of the mean value

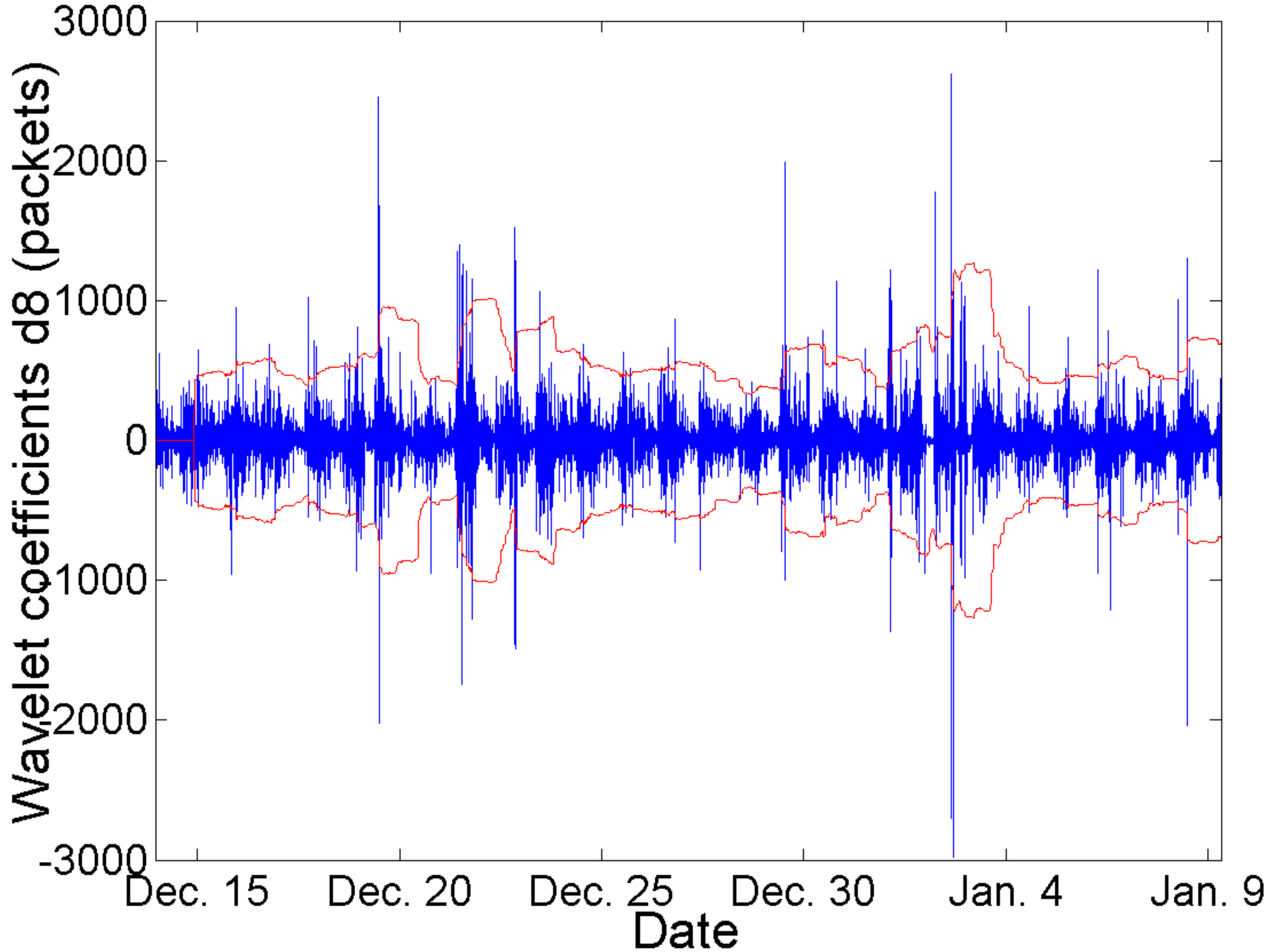
Wavelet approximation coefficients



Wavelet detail coefficients: d_9



Wavelet detail coefficients: d_8



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
- 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):

	Route Views	RIPE
Faloutsos, 1999	Yes	No
Chang, 2001	Yes	Yes
Vukadinovic, 2001	Yes	No
Mihail, 2003	Yes	Yes

- 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]:

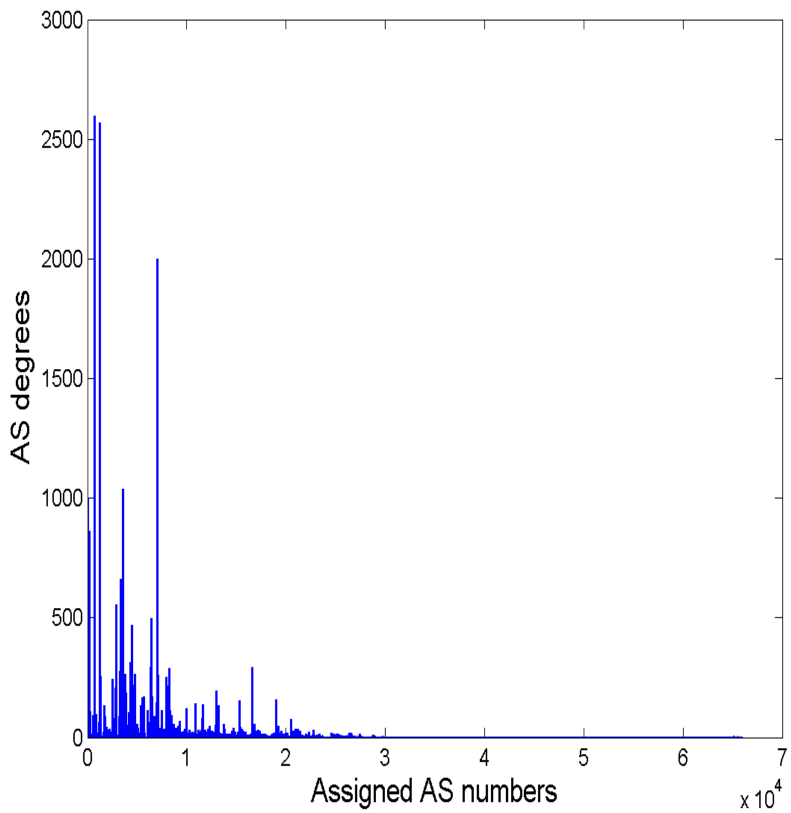
$$N(i,j) = \begin{cases} 1 & \text{if } i = j \text{ and } d_i \neq 0 \\ -\dfrac{1}{\sqrt{d_i d_j}} & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

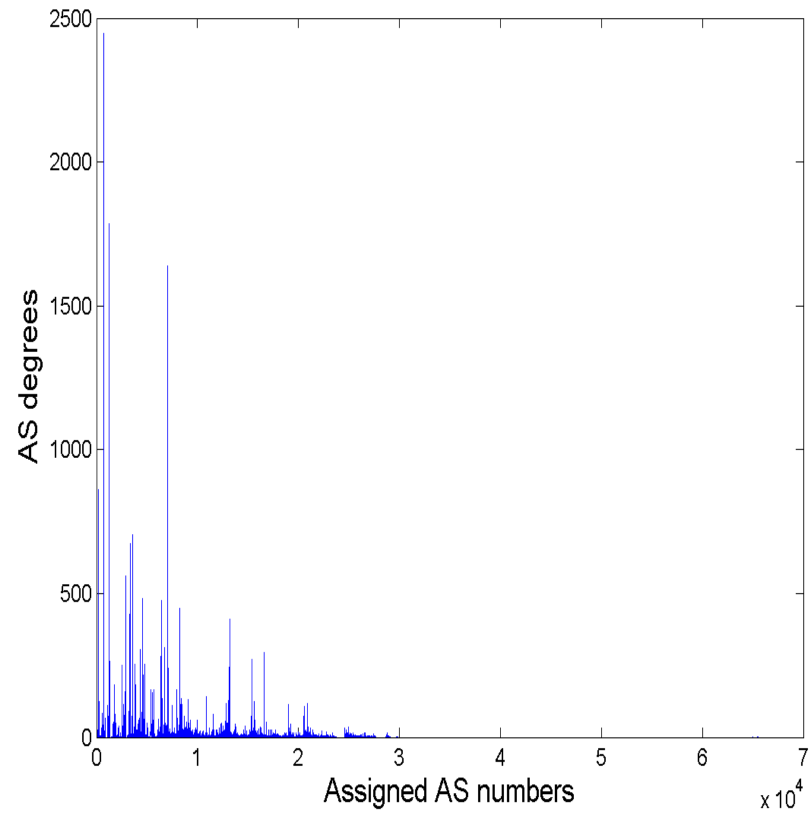
d_i and d_j are degrees of node i and j, respectively

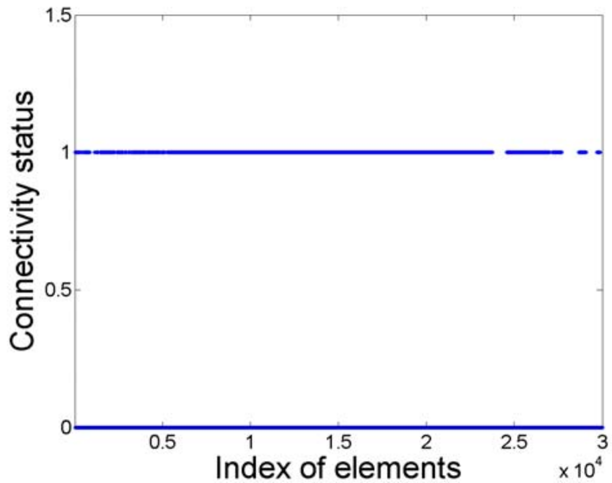
- The second smallest eigenvalue [Fiedler, 1973]
- The largest eigenvalue [Chung, 1997]
- Characteristic valuation [Fiedler, 1975]

Spectral analysis of topology data

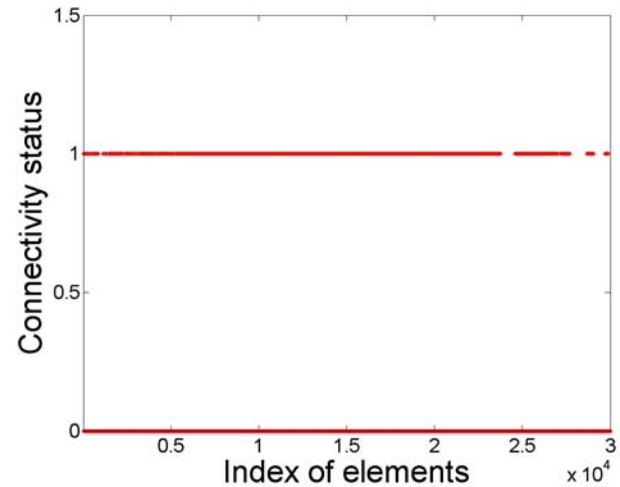
- Consider only ASs with the first 30,000 assigned AS numbers
- AS degree distribution in Route Views and RIPE datasets:





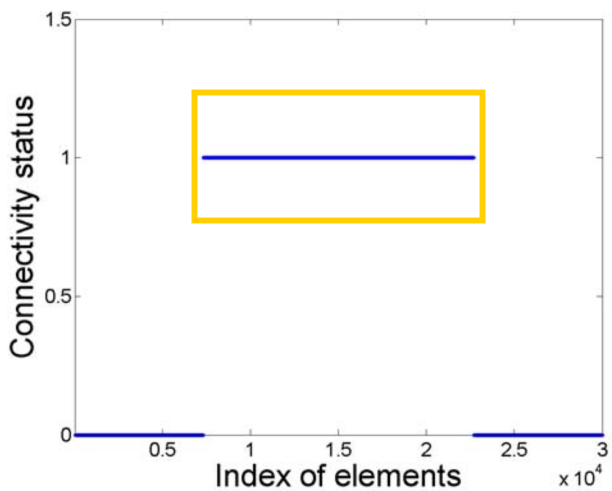


(a) RouteViews_original

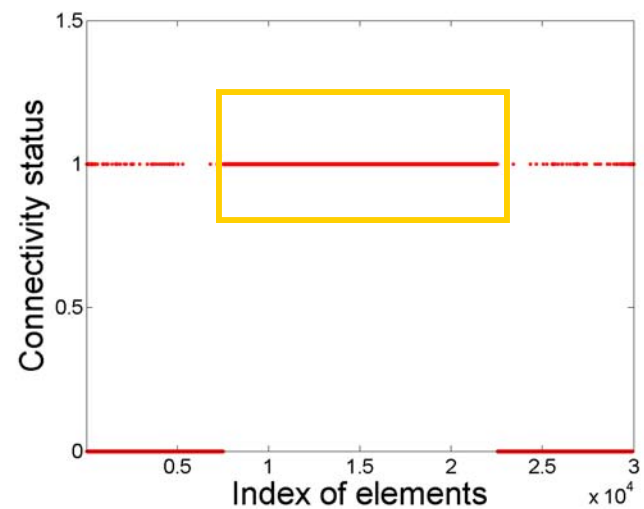


(b) RIPE_original

Before the sort

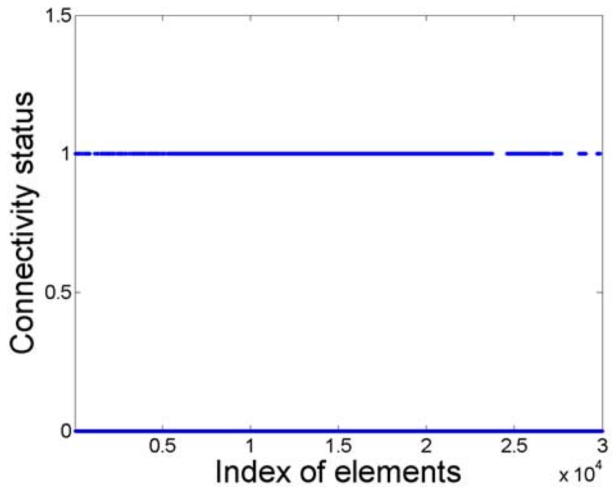


(c) RouteViews_min

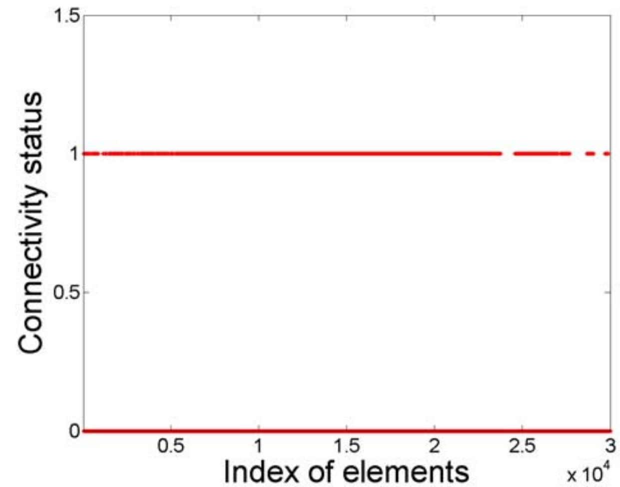


(d) RIPE_min

After the sort

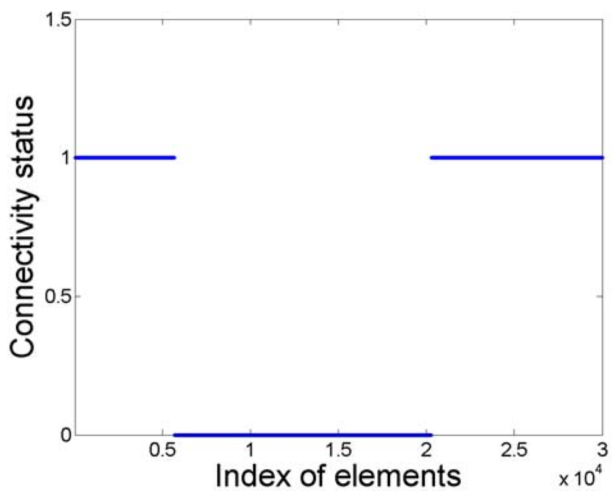


(a) RouteViews_original

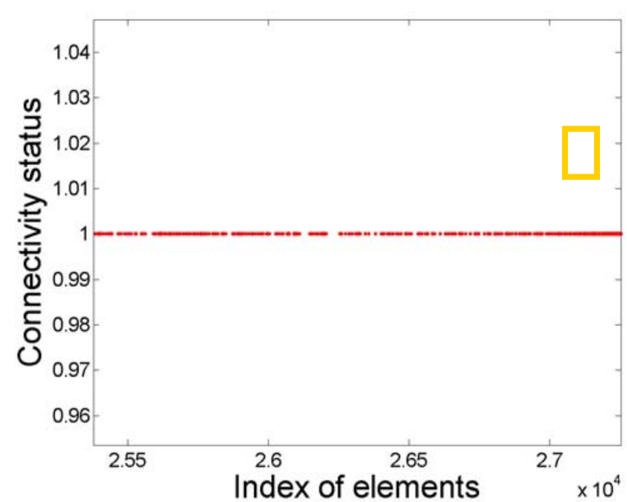


(b) RIPE_original

Before the sort



(c) RouteViews_max



(d) RIPE_max

After the sort

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

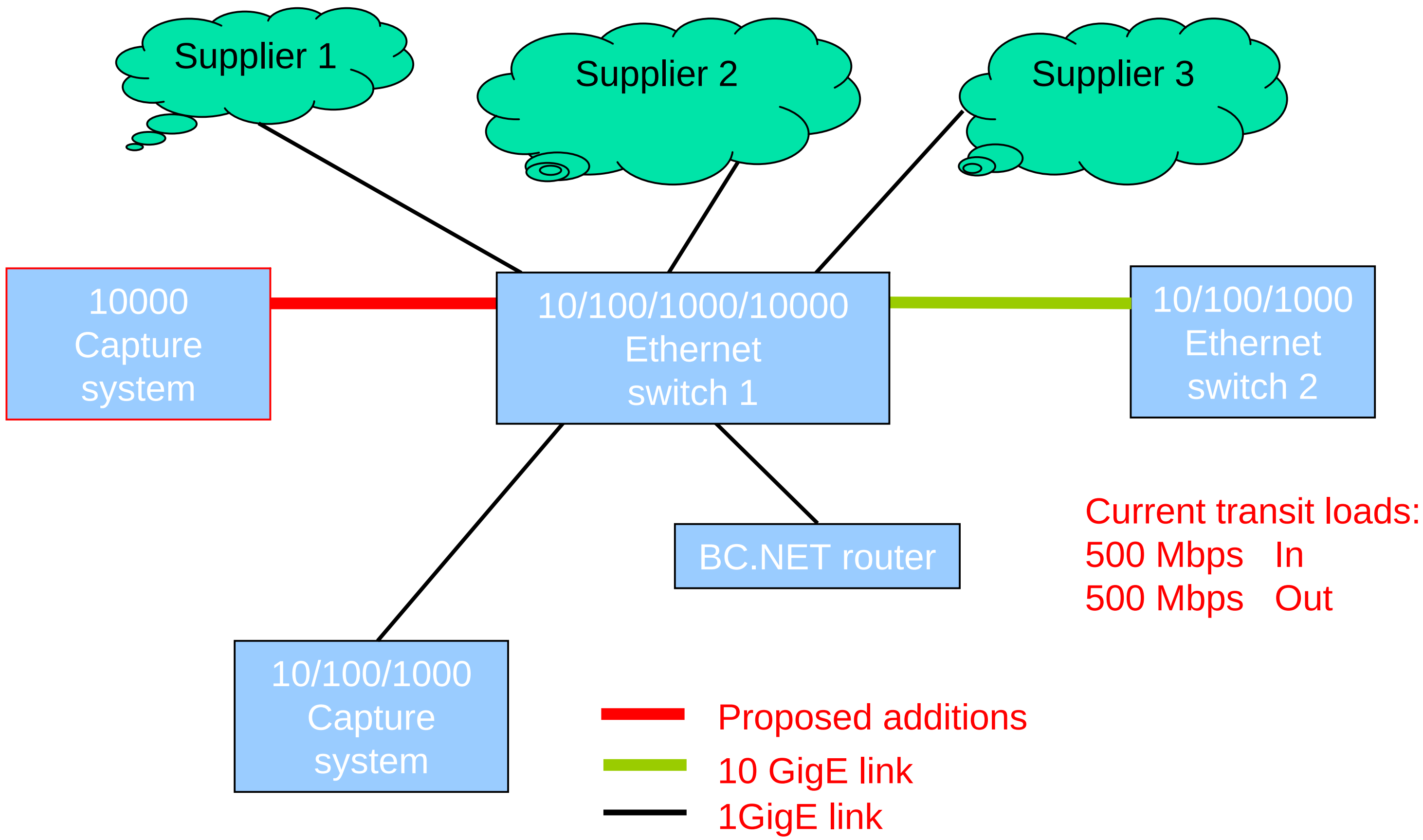
- Traffic data from deployed networks (Telus Mobility, E-Comm, ChinaSat, the Internet) were used to:
- evaluate network performance
- characterize and model traffic (inter-arrival and call holding times)
- classify network users using clustering algorithms
- predict network traffic by employing SARIMA 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 form 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 Gpbs): 8 GB RAM, 8 TB RAID storage with write-to-disk performance of 2 Gbps

BGP: border gateway protocol

BC.NET traffic measurements



References: downloads

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

- M. Najiminaini, L. Subedi, and Lj. Trajkovic, "Analysis of Internet topologies: a historical view," presented at *IEEE Int. Symp. Circuits and Systems*, Taipei, Taiwan, May 2009.
- S. Lau and Lj. Trajkovic, "Analysis of traffic data from a hybrid satellite-terrestrial network," in *Proc. QShine 2007*, Vancouver, BC, Canada, Aug. 2007.
- B. Vujičić, L. Chen, and Lj. Trajković, "Prediction of traffic in a public safety network," in *Proc. ISCAS 2006*, Kos, Greece, May 2006, pp. 2637–2640.
- N. Cackov, J. Song, B. Vujičić, S. Vujičić, and Lj. Trajković, "Simulation of a public safety wireless networks: a case study," *Simulation*, vol. 81, no. 8, pp. 571–585, Aug. 2005.
- B. Vujičić, N. Cackov, S. Vujičić, and Lj. Trajković, "Modeling and characterization of traffic in public safety wireless networks," in *Proc. SPECTS 2005*, Philadelphia, PA, July 2005, pp. 214–223.
- J. Song and Lj. Trajković, "Modeling and performance analysis of public safety wireless networks," in *Proc. IEEE IPCCC*, Phoenix, AZ, Apr. 2005, pp. 567–572.
- H. Chen and Lj. Trajković, "Trunked radio systems: traffic prediction based on user clusters," in *Proc. IEEE ISWCS 2004*, Mauritius, Sept. 2004, pp. 76–80.
- D. Sharp, N. Cackov, N. Lasković, Q. Shao, and Lj. Trajković, "Analysis of public safety traffic on trunked land mobile radio systems," *IEEE J. Select. Areas Commun.*, vol. 22, no. 7, pp. 1197–1205, Sept. 2004.
- Q. Shao and Lj. Trajković, "Measurement and analysis of traffic in a hybrid satellite-terrestrial network," in *Proc. SPECTS 2004*, San Jose, CA, July 2004, pp. 329–336.
- N. Cackov, B. Vujičić, S. Vujičić, and Lj. Trajković, "Using network activity data to model the utilization of a trunked radio system," in *Proc. SPECTS 2004*, San Jose, CA, July 2004, pp. 517–524.
- J. Chen and Lj. Trajkovic, "Analysis of Internet topology data," *Proc. IEEE Int. Symp. Circuits and Systems*, Vancouver, British Columbia, Canada, May 2004, vol. IV, pp. 629-632.

References: self-similarity

- A. Feldmann, "Characteristics of TCP connection arrivals," in *Self-similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds., New York: Wiley, 2000, pp. 367–399.
- T. Karagiannis, M. Faloutsos, and R. H. Riedi, "Long-range dependence: now you see it, now you don't!," in *Proc. GLOBECOM '02*, Taipei, Taiwan, Nov. 2002, pp. 2165–2169.
- W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.
- M. S. Taqqu and V. Teverovsky, "On estimating the intensity of long-range dependence in finite and infinite variance time series," in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications.* Boston, MA: Birkhauser, 1998, pp. 177–217.

References: self-similarity

- P. Abry and D. Veitch, "Wavelet analysis of long-range dependence traffic," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 2–15, Jan. 1998.
- P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data," in *Self-similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds. New York: Wiley, 2000, pp. 39–88.
- P. Barford, A. Bestavros, A. Bradley, and M. Crovella, "Changes in Web client access patterns: characteristics and caching implications in world wide web," *World Wide Web*, Special Issue on Characterization and Performance Evaluation, vol. 2, pp. 15–28, 1999.
- Z. Bi, C. Faloutsos, and F. Korn, "The 'DGX' distribution for mining massive, skewed data," in *Proc. of ACM SIGCOMM Internet Measurement Workshop*, San Francisco, CA, Aug. 2001, pp. 17–26.
- M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

References: time series

- G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, 2nd edition. San Francisco, CA: Holden-Day, 1976, pp. 208–329.
- P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd Edition. New York: Springer-Verlag, 2002.
- N. H. Chan, *Time Series: Applications to Finance*. New York: Wiley-Interscience, 2002.
- K. Burnham and D. Anderson, *Model Selection and Multimodel Inference*, 2nd ed. New York, NY: Springer-Verlag, 2002.
- G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

References: cluster analysis

- P. Cheeseman and J. Stutz, “Bayesian classification (AutoClass): theory and results,” in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., AAAI Press/MIT Press, 1996.
- J. W. Han and M. Kamber, *Data Mining: Concepts And Techniques*. San Francisco: Morgan Kaufmann, 2001.
- T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.
- L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

References: data mining

- J. Han and M. Kamber, *Data Mining: concept and techniques*. San Diego, CA:Academic Press, 2001.
- W. Wu, H. Xiong, and S. Shekhar, *Clustering and Information Retrieval*. Norwell,MA: Kluwer Academic Publishers, 2004.
- Z. Chen, *Data Mining and Uncertainty Reasoning: and integrated approach*. New York, NY: John Wiley & Sons, 2001.
- T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, July. 2002.
- P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Reading,MA: Addison-Wesley, 2006, pp. 487–568.
- L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an introduction to cluster analysis*. New York, NY: John Wiley & Sons, 1990.
- M. Last, A. Kandel, and H. Bunke, Eds., *Data Mining in Time Series Databases*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2004.
- W.-K. Ching and M. K.-P. Ng, Eds., *Advances in Data Mining and Modeling*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2003.

References: protocols

- D. E. Comer, *Internetworking with TCP/IP, Vol 1: Principles, Protocols, and Architecture*, 4th ed. Upper Saddle River, NJ: Prentice-Hall, 2000.
- W. R. Stevens, *TCP/IP Illustrated (vol. 1): The Protocols*. Reading, MA: Addison-Wesley, 1994.
- J. Postel, Ed., "Transmission Control Protocol," RFC 793, Sep. 1981.
- J. Postel, "TCP and IP bake off," RFC 1025, Sep. 1987.
- J. Mogul and S. Deering, "Path MTU discovery," RFC 1191, Nov. 1990.
- V. Jacobson, R. Braden, and D. Borman, "TCP extensions for high performance," RFC 1323, May 1992.
- M. Allman, S. Floyd, and C. Partridge, "Increasing TCP's initial window," RFC 2414, Sep. 1998.
- M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, "TCP selective acknowledgment options," RFC 2018, Oct. 1996.
- M. Allman, D. Glover, and L. Sanchez, "Enhancing TCP over satellite channels using standard mechanisms," RFC 2488, Jan. 1999.
- M. Allman, S. Dawkins, D. Glover, J. Griner, D. Tran, T. Henderson, J. Heidemann, J. Touch, H. Kruse, S. Ostermann, K. Scott, and J. Semke, "Ongoing TCP research related to satellites," RFC 2760, Feb. 2000.
- J. Border, M. Kojo, J. Griner, G. Montenegro, and Z. Shelby, "Performance enhancing proxies intended to mitigate link-related degradations," RFC 3135, June 2001.
- S. Floyd, "Inappropriate TCP resets considered harmful," RFC 3360, Aug. 2002.

References: fingerprinting

- R. Beverly, "A Robust Classifier for Passive TCP/IP Fingerprinting," in *Proc. Passive and Active Meas. Workshop 2004*, Antibes Juan-les-Pins, France, Apr. 2004, pp. 158–167.
- C. Smith and P. Grundl, "Know your enemy: passive fingerprinting," The Honeynet Project, Mar. 2002. [Online]. Available: http://www.honeynet.org/papers/finger/.
- Passive OS fingerprinting tool ver. 2 (p0f v2). [Online]. Available: http://lcamtuf.coredump.cx/p0f.shtml/.
- B. Petersen, "Intrusion detection FAQ: What is p0f and what does it do?" The SysAdmin, Audit, Network, Security (SANS) Institute. [Online]. Available: http://www.sans.org/resources/idfaq/p0f.php.
- T. Miller, "Passive OS fingerprinting: details and techniques," The SysAdmin, Audit, Network, Security (SANS) Institute. [Online]. Available: http://www.sans.org/reading room/special.php/.

References: anomalies

- P. Barford and D. Plonka, "Characteristics of network traffic flow anomalies," in *Proc. ACM SIGCOMM Internet Meas. Workshop 2001*, Nov. 2001, pp. 69–73.
- P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in *Proc. ACM SIGCOMM Internet Meas. Workshop 2002*, Marseille, France, Nov. 2002, pp. 71–82.
- Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *Proc. ACM SIGCOMM Internet Meas. Conf. 2005*, Berkeley, CA, Oct. 2005, pp. 317–330.
- A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *Proc. ACM SIGCOMM Internet Meas. Conf. 2005*, Berkeley, CA, Oct. 2005, pp. 331–344.
- P. Huang, A. Feldmann, and W. Willinger, "A non-instrusive, wavelet-based approach to detecting network performance problems," in *Proc. ACM SIGCOMM Internet Meas. Workshop 2001*, San Francisco, CA, Nov. 2001, pp. 213–227.
- A. Lakhina, M. Crovella, and C. Diot, "Characterization of network-wide anomalies in traffic flows," in *Proc. ACM SIGCOMM Internet Meas. Conf. 2004*, Taormina, Italy, Oct. 2004, pp. 201–206.
- A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 4, pp. 219–230, Oct. 2004.
- M. Arlitt and C. Williamson, "An analysis of TCP reset behaviour on the Internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 1, pp. 37–44, Jan. 2005.

References: spectral analysis

- M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the Internet topology,” *Proc. ACM SIGCOMM, Computer Communication Review*, vol. 29, no. 4, pp. 251-262, Sept. 1999.
- G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos, "Power-laws and the AS-level Internet topology," *IEEE/ACM Trans. Networking*, vol. 11, no. 4, pp. 514–524, Aug. 2003.
- A. Medina, I. Matta, and J. Byers, "On the origin of power laws in Internet topologies," *Proc. ACM SIGCOMM 2000*, Computer Communication Review, vol. 30, no. 2, pp. 18–28, Apr. 2000.
- L. Gao, "On inferring autonomous system relationships in the Internet," *IEEE/ACM Trans. Networking*, vol. 9, no. 6, pp. 733–745, Dec. 2001.
- D. Vukadinovic, P. Huang, and T. Erlebach, “On the Spectrum and Structure of Internet Topology Graphs,” in H. Unger et al., editors, *Innovative Internet Computing Systems*, LNCS2346, pp. 83–96. Springer, Berlin, Germany, 2002.
- Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "The origin of power laws in Internet topologies revisited," *Proc. INFOCOM*, New York, NY, USA, Apr. 2002, pp. 608–617.
- H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Towards capturing representative AS-level Internet topologies," *Proc. of ACM SIGMETRICS 2002*, New York, NY, June 2002, pp. 280–281.

References: spectral analysis

- H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Network topology generators: degree-based vs. structural," *Proc. ACM SIGCOMM, Computer Communication Review*, vol. 32, no. 4, pp. 147–159, Oct. 2002.
- M. Mihail, C. Gkantsidis, and E. Zegura, "Spectral analysis of Internet topologies," *Proc. of Infocom 2003,* San Francisco, CA, Mar. 2003, vol. 1, pp. 364–374.
- S. Jaiswal, A. Rosenberg, and D. Towsley, "Comparing the structure of power-law graphs and the Internet AS graph," *Proc. 12th IEEE International Conference on Network Protocols*, Washington DC, Aug. 2004, pp. 294–303.
- F. R. K. Chung, *Spectral Graph Theory*. Providence, Rhode Island: Conference Board of the Mathematical Sciences, 1997, pp. 2–6.
- M. Fiedler, "Algebraic connectivity of graphs," *Czech. Math. J.*, vol. 23, no. 2, pp. 298–305, 1973.

References: traffic analysis

- Y. W. Chen, "Traffic behavior analysis and modeling sub-networks," *International Journal of Network Management*, John Wiley & Sons, vol. 12, pp. 323–330, 2002.
- Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling of PCS networks," *IEEE Trans. on Communications*, vol. 47, no. 7, pp. 1062–1072, July 1999.
- N. K. Groschwitz and G. C. Polyzos, "A time series model of long-term NSFNET backbone traffic," in *Proc. IEEE International Conference on Communications* (ICC'94), New Orleans, LA, May 1994, vol. 3, pp. 1400–1404.
- D. Papagiannaki, N. Taft, Z.-L. Zhang, and C. Diot, "Long-term forecasting of Internet backbone traffic: observations and initial models," in *Proc. IEEE INFOCOM 2003,* San Francisco, CA, April 2003, pp. 1178–1188.
- D. Tang and M. Baker, "Analysis of a metropolitan-area wireless network," *Wireless Networks*, vol. 8, no. 2/3, pp. 107–120, Mar.-May 2002.
- R. B. D'Agostino and M. A. Stephens, Eds., *Goodness-of-Fit Techniques*. New York: Marcel Dekker, 1986. pp. 63–93, pp. 97–145, pp. 421–457.
- F. Barceló and J. I. Sánchez, "Probability distribution of the inter-arrival time to cellular telephony channels," in *Proc. of the 49th Vehicular Technology Conference*, May 1999, vol. 1, pp. 762–766.
- F. Barcelo and J. Jordan, "Channel holding time distribution in public telephony systems (PAMR and PCS)," *IEEE Trans. Vehicular Technology*, vol. 49, no. 5, pp. 1615–1625, Sept. 2000.