Efficient Mining of Data Through Reuse in a Public Safety Network

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 study:
  - public safety wireless network: E-Comm
- Conclusions and references

Collaborators

M.A.Sc. and M.Eng. students at SFU:

- E-Comm data analysis:
  - Duncan Sharp
  - Hao (Leo) Chen
  - Bozidar Vujičić
  - Nikola Cackov
  - Svetlana Vujičić
  - Nenad Lasković
- ChinaSat data analysis:
  - Qing (Kenny) Shao
  - Savio Lau

Network traffic measurements

- Focus of networking research during:
  - mid to late 1980's
  - early 1990's
- Motivation for traffic measurements:
  - understand traffic characteristics in deployed networks
  - develop traffic models
  - evaluate performance of protocols and applications
  - perform trace driven simulations

Case studies: deployed networks

- Analysis of traffic from operational networks:
  - provides useful information about the user behavior patterns
  - enables network operators to better understand the behavior of network users
  - helps provide better quality of service
- Traffic prediction: important to assess future network capacity requirements and to plan future network developments

Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection
  - statistical analysis
  - clustering tools
  - prediction analysis
- Case study:
  - wireless network: E-Comm
- Conclusions and references

Traffic traces

- Most available traffic traces are from wired networks within research communities:
  - Bellcore, LBNL, Auckland University
- Few traces have been collected from wireless or satellite commercial networks
- Various factors affect Internet traffic patterns:
  - Web, Proxy, Napster, MP3, Web mail

Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection
  - statistical analysis
  - clustering tools
  - prediction analysis
- Case study:
  - wireless network: E-Comm
- Conclusions and references

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

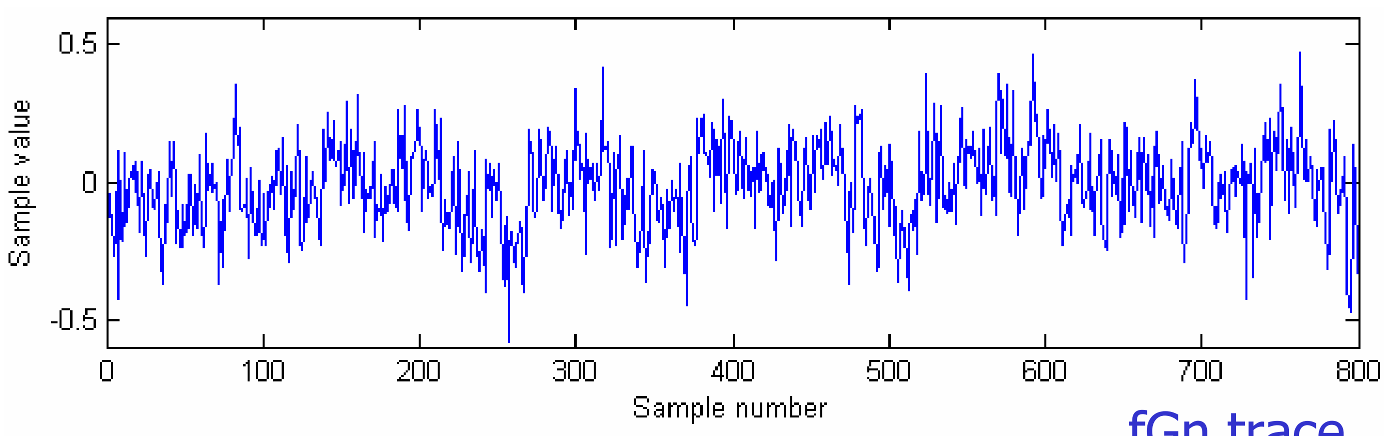
    where m is the level of aggregation
- Implications:
  - no natural length of bursts
  - bursts exist across many time scales
  - traffic does not become "smoother" when aggregated (unlike Poisson traffic)

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

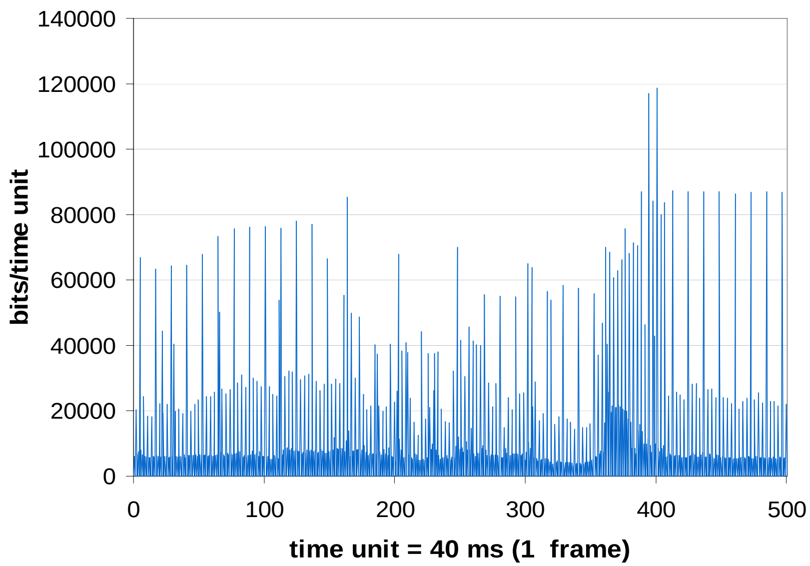
- High variability:
  - when the sample size increases, variance of the sample mean decays more slowly than expected
- Burstiness over a range of timescales:
  - long runs of large values followed by long runs of small values, repeated in aperiodic patterns



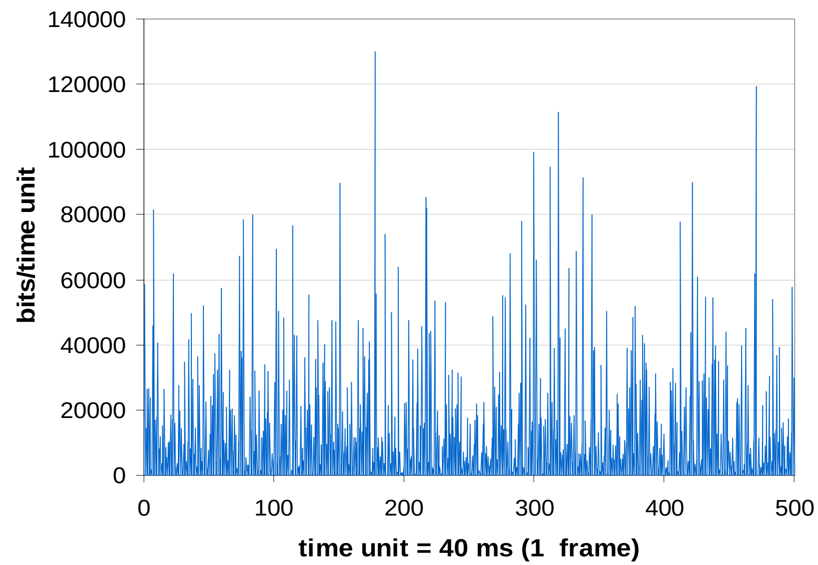
fGn trace

Self-similar traffic patterns

Genuine MPEG traffic trace



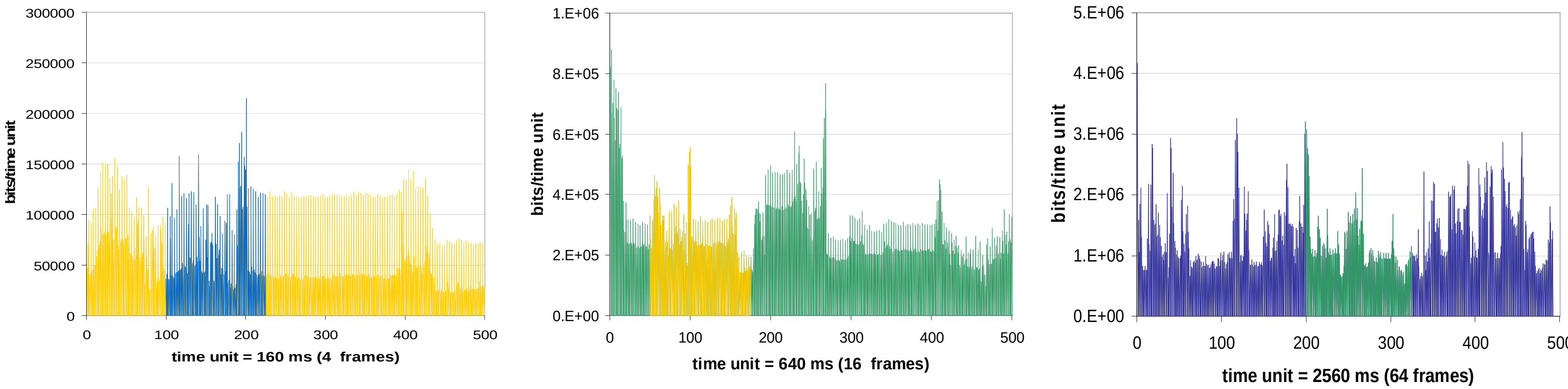
Poisson model



The two traces have identical mean.

Influence of time-scales

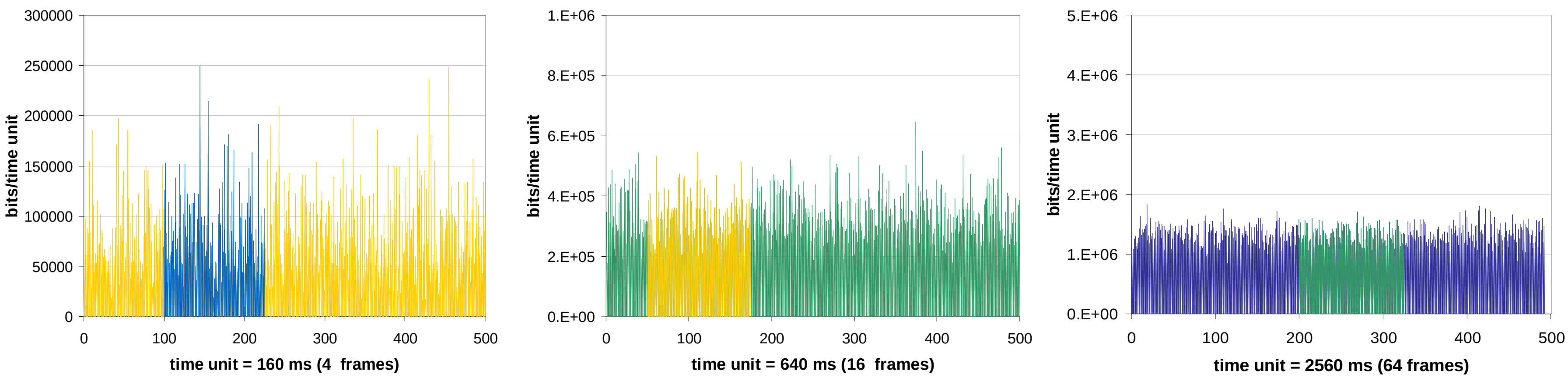
- Genuine MPEG traffic trace:



W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking,* vol. 2, pp. 1–15, 1994.

Influence of time-scales

- Synthetically generated Poisson model:



W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking,* vol. 2, pp. 1–15, 1994.

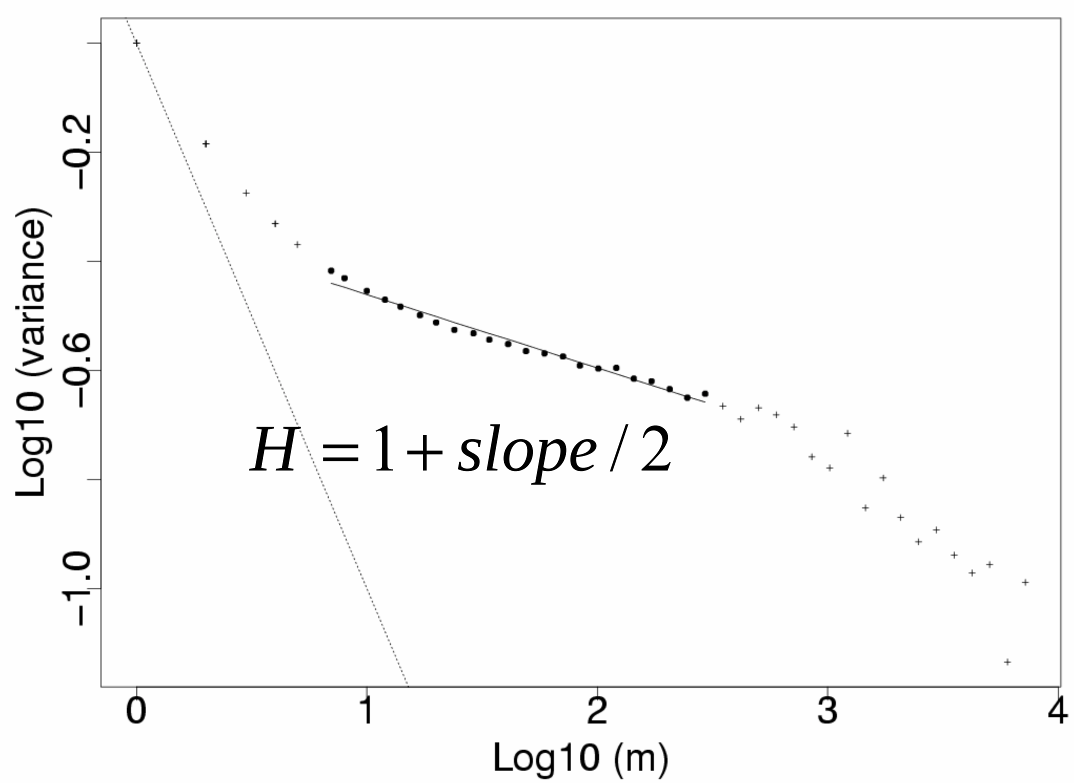
Estimation of H

Various estimators:

- variance-time plots
- R/S plots
- periodograms
- wavelets

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

Estimation of H: variance-time plot



Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection
  - statistical analysis
  - clustering tools
  - prediction analysis
- Case study:
  - wireless network: E-Comm
- Conclusions and references

Clustering analysis

- Clustering analysis groups or segments a collection of objects into subsets or clusters
- Objects within a cluster are more similar to each other than objects in distinct clusters
- An object can be described by a set of measurements or by its relations to other objects
- Network users are classified into clusters, according to the similarity of their behavior patterns

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

- We classify the calling patterns of talk groups by using two known clustering tools:
  - AutoClass tool
  - K-means algorithm

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: AutoClass

- AutoClass is an unsupervised classification tool based on the Bayesian approach
- The goal of an unsupervised classification is to find the most probable set of class descriptions given the data and the prior expectations
- AutoClass begins by creating a random classification and then manipulates it into a high probability classification through local changes
- It repeats the process until it converges to a local maximum
- It starts over again and continues for a specified number of tries
- Each new try begins with a certain number of classes and may conclude with a smaller number of classes

Clustering algorithms: K-means

- The K-means algorithm is commonly used for data clustering
- 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)
- The algorithm is well-known for its simplicity and efficiency

Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection
  - statistical analysis
  - clustering tools
  - traffic prediction
- Case study:
  - wireless network: E-Comm
- Conclusions and references

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

Traffic prediction: ARMA model

- Predicting traffic using AutoRegressive Moving-Average (ARMA) model:

$$X(t) = \phi_1 X(t-1) + \cdots + \phi_p X(t-p) + e(t) + \theta_1 e(t-1) \cdots + \theta_q e(t-q)$$

$$\phi_P(B)\omega_t = \theta_q(B)\varepsilon_t$$

where:

  - AR and MA parts: $\phi(B)$ and $\theta(B)$
  - B = the back-shift operator: $B^i X_t = X_{t-i}$
- Model:
  - past values: AutoRegressive (AR) structure
  - past random fluctuant effect: Moving Average (MA) process

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) selected based on:
  - time series plot of traffic data
  - autocorrelation and partial autocorrelation functions
- Validity of parameter selection:
  - Akaike's information criterion AIC
  - Akaike's information criterion corrected AICc
  - Bayesian information criterion BIC

Roadmap

- Introduction
- Traffic data and analysis tools:
  - data collection
  - statistical analysis
  - clustering tools
  - prediction analysis
- Case study:
  - wireless network: E-Comm
- 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

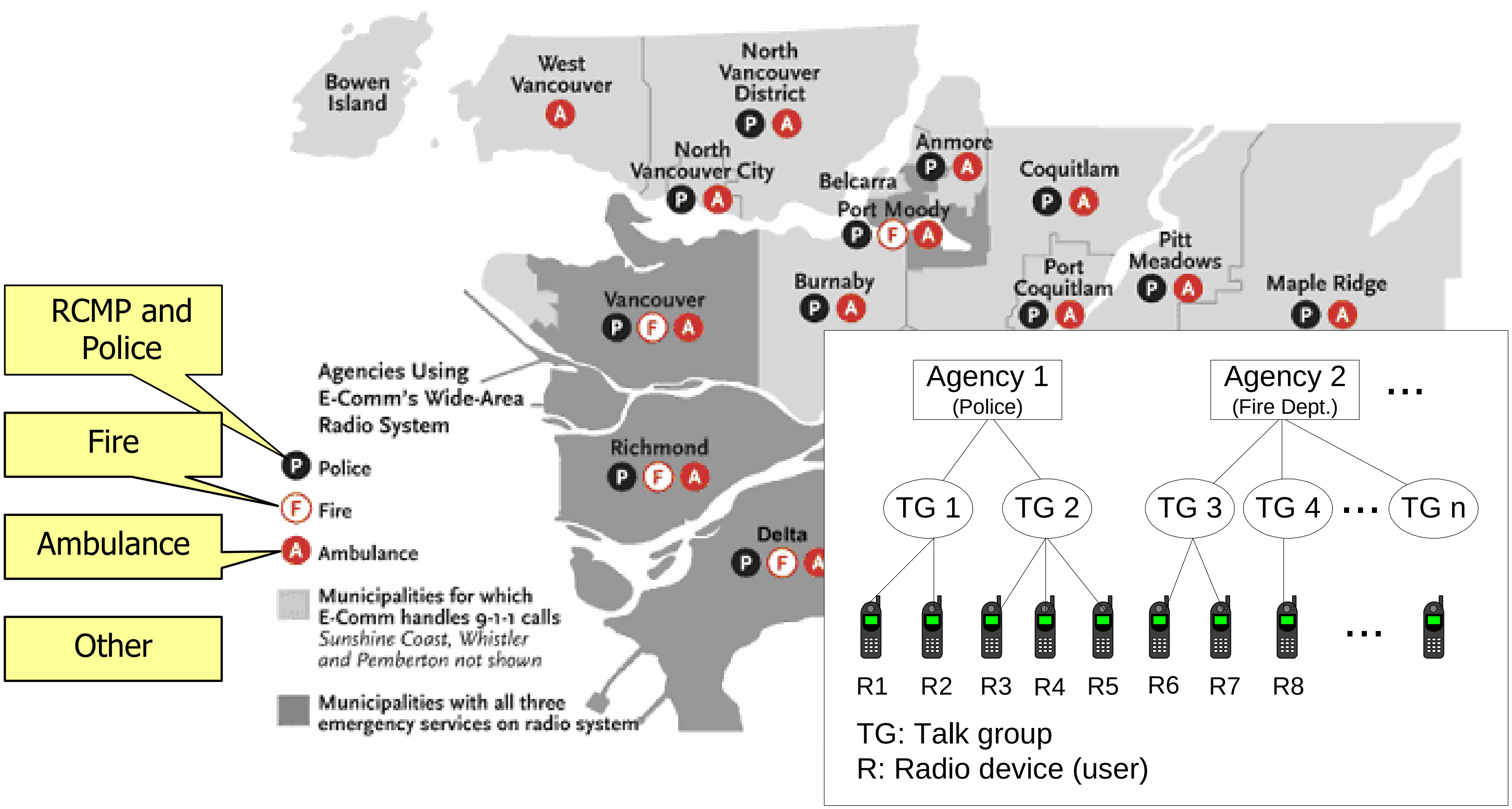
E-Comm network: utilization

- Using network activity data to model the utilization of a trunked radio system
- Data and network models
- OPNET simulation results

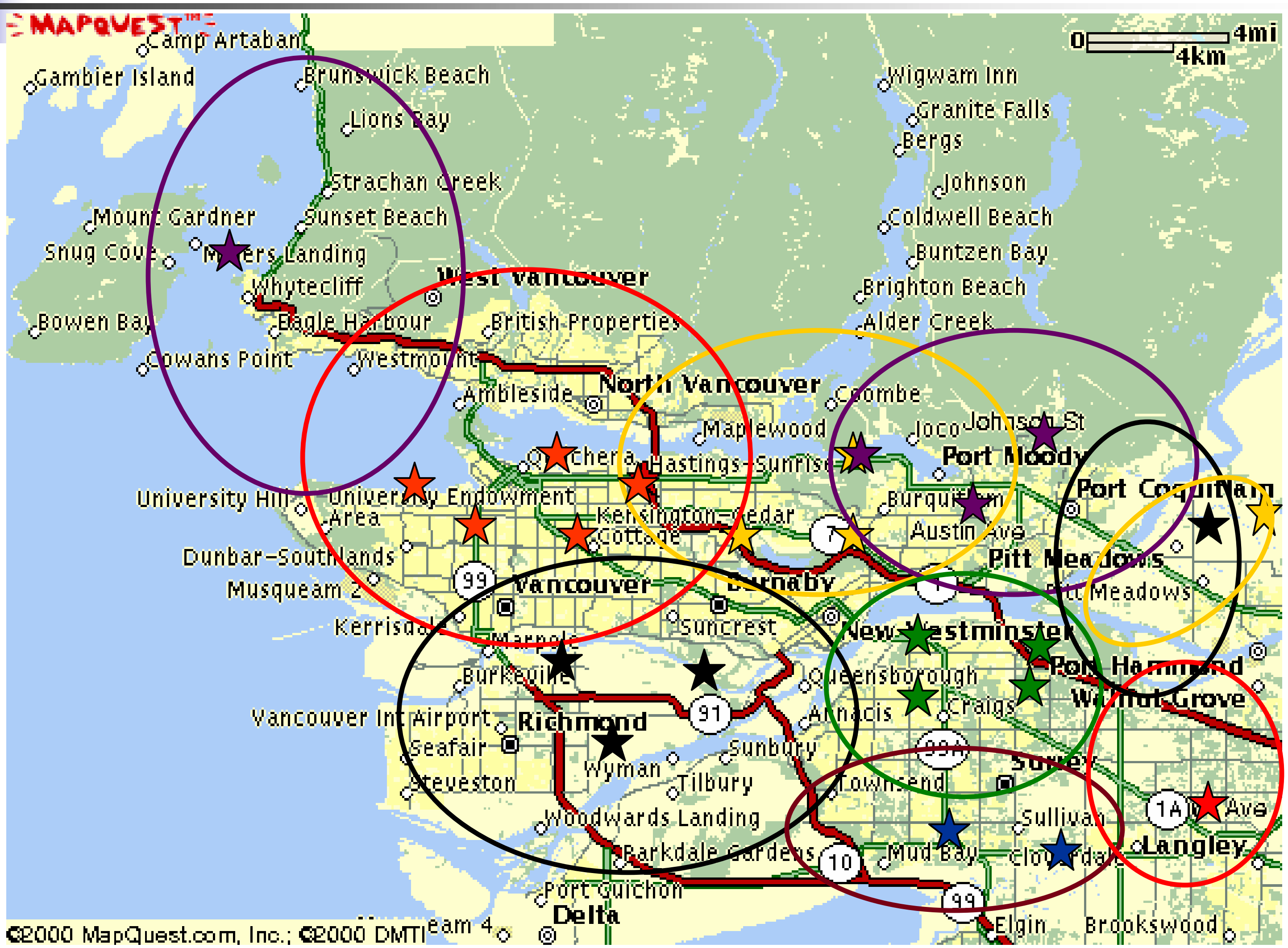
N. Cackov, B. Vujičić, S. Vujičić, and Lj. Trajković, "Using network activity data to model the utilization of a trunked radio system," in *Proc. SPECTS 2004*, San Jose, CA, July 2004, pp. 517–524.

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.

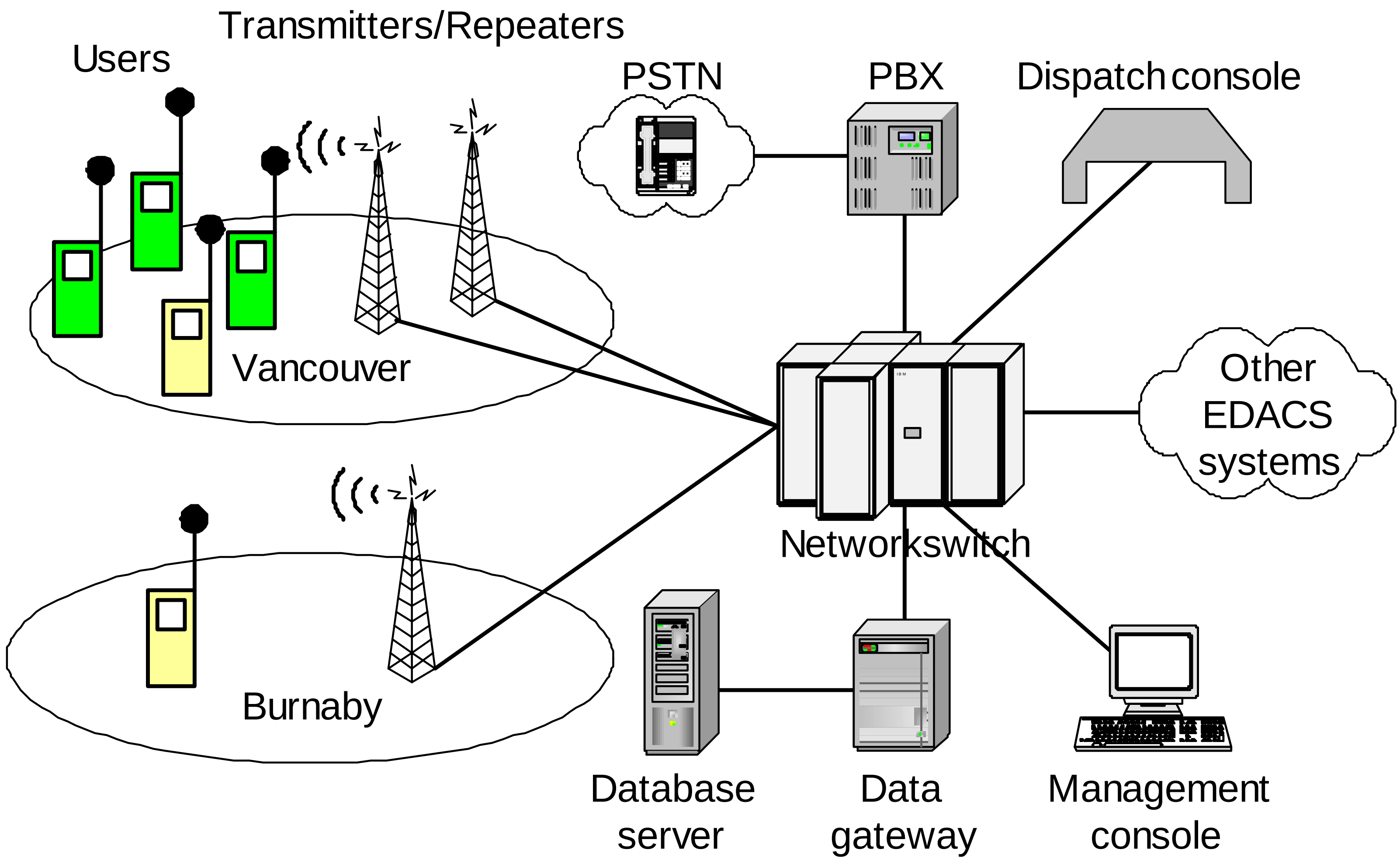
E-Comm network: coverage and user agencies



E-Comm network coverage



E-Comm network architecture



Call establishment

- Users are organized in 617 talk groups:
  - one-to-many type of conversations (group call)
  - multi-system call represents single group call involving more than one system/cell
- Push-to-talk (PTT) mechanism for network access:
  - user presses the PTT button
  - system locates other members of the talk group
  - system checks for availability of channels:
    - channel available: call established
    - all channels busy: call queued/dropped
  - user releases PTT:
    - call terminates

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

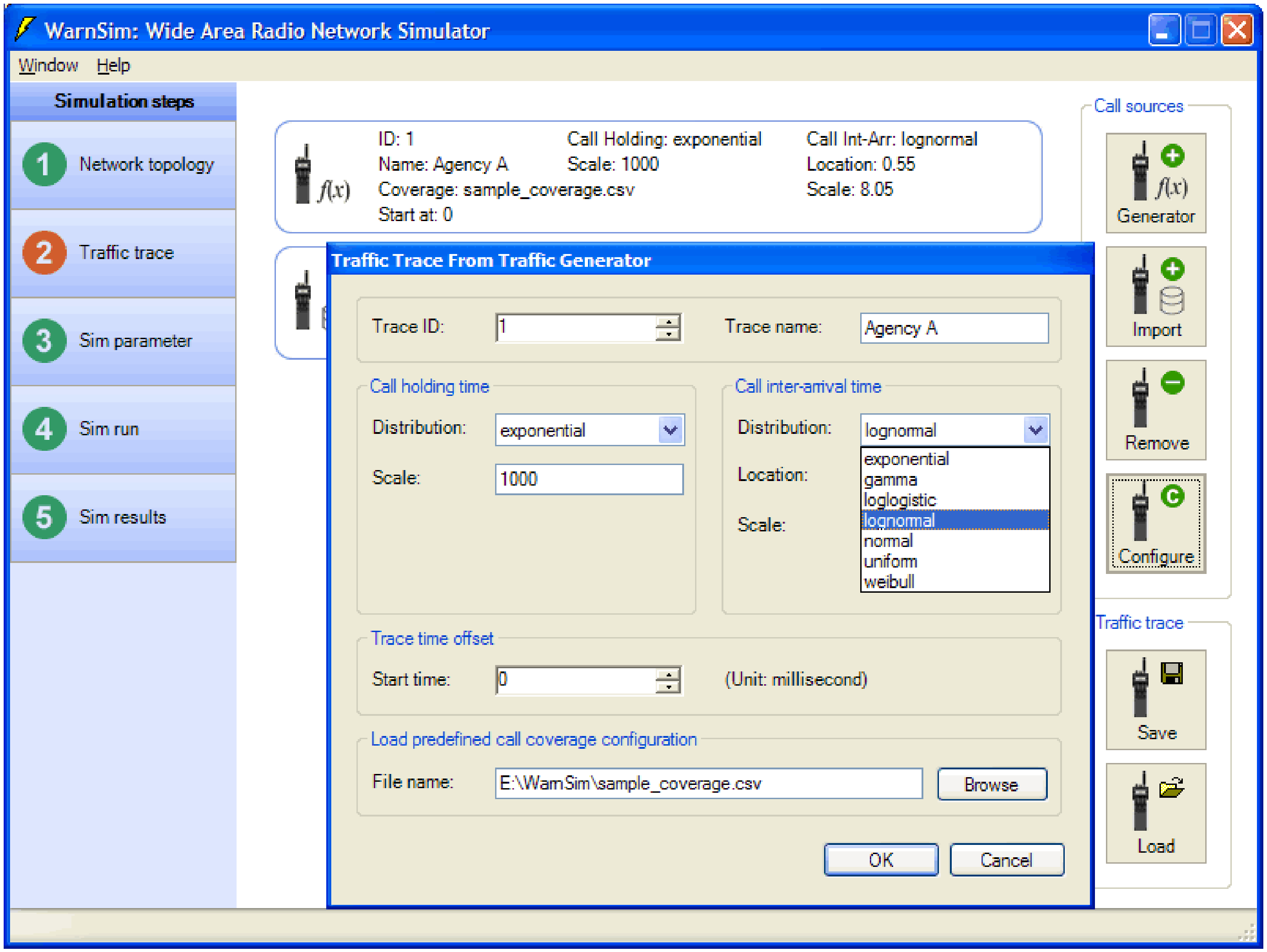
J. Song and Lj. Trajković, "Modeling and performance analysis of public safety wireless networks," in *Proc. IEEE IPCCC*, Phoenix, AZ, Apr. 2005, pp. 567–572.

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.

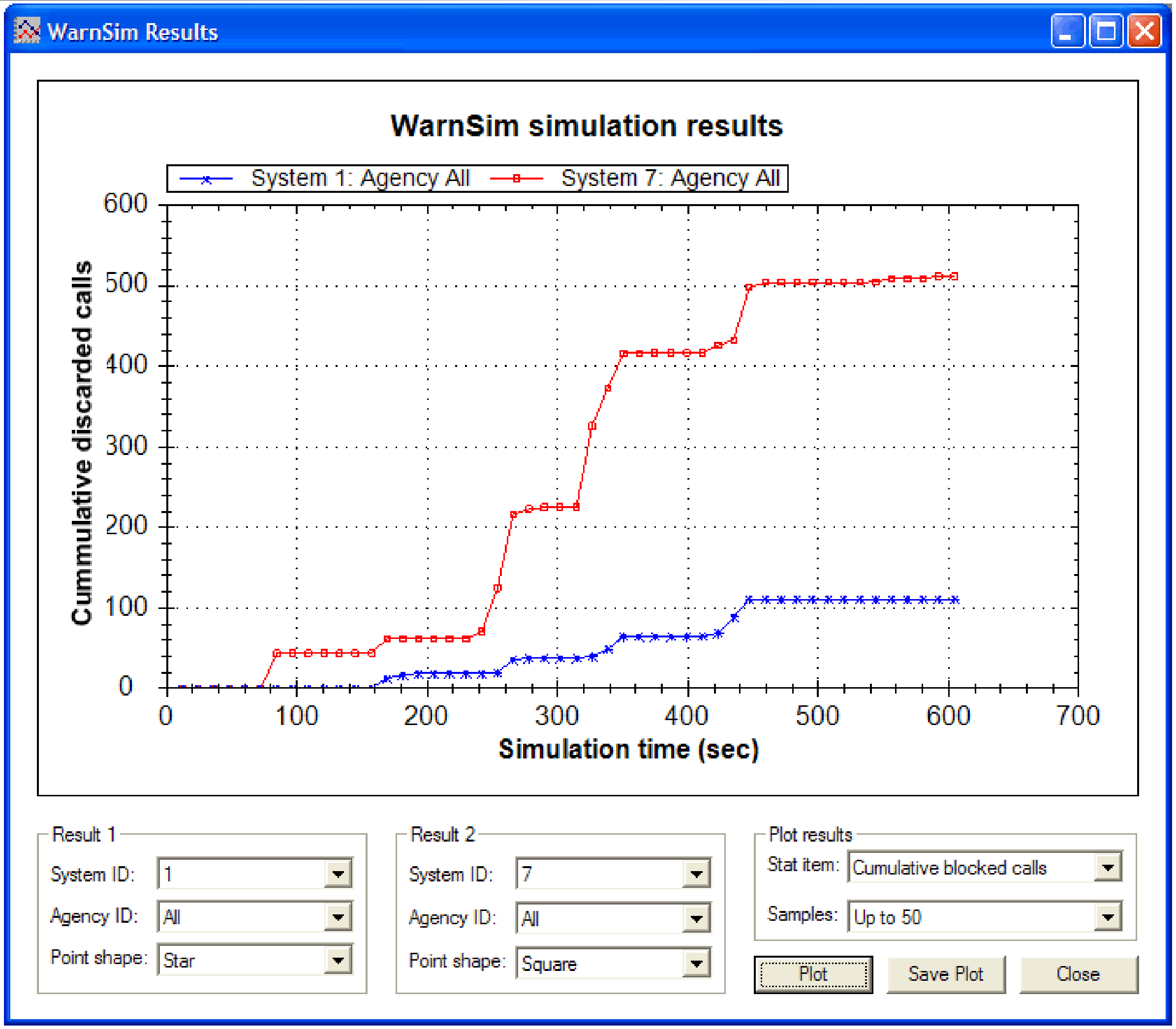
WarnSim overview

- 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.vannet.ca/warnsim
  - effective, flexible, and easy to use
  - developed using Microsoft Visual C# .NET
  - operates on Windows platforms

Traffic trace generator



WarnSim results



Modeling and characterization of traffic

- Statistical concepts and analysis tools
- Analysis of traffic data:
  - call inter-arrival times
  - call holding times
- Traffic modeling and characterization

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

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

Erlang C

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

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

Erlang models

- Erlang B model assumes:
  - call holding time follows exponential distribution
  - blocked call will be rejected immediately
- Erlang C model assumes:
  - call holding time follows exponential distribution
  - blocked call will be put into a FIFO queue with infinite size

Kolmogorov-Smirnov test

- Goodness-of-fit test: quantitative decision whether the empirical cumulative distribution function (ECDF) of a set of observations is consistent with a random sample from an assumed theoretical distribution
- ECDF is a step function (step size 1/N) of N ordered data points $Y_1, Y_2, \ldots, Y_N$:

$$E_N = \frac{n(i)}{N}$$

$n(i)$: the number of data samples with values smaller than Y_i

Parameters

- Hypothesis:
  - null: the candidate distribution fits the empirical data
  - alternative: the candidate distribution does not fit the empirical data
- Input parameters: significance level σ and tail
- Output parameters:
  - p-value
  - k: test statistic
  - cv: critical (cut-off) value

Input parameters

- Significance level σ: determines if the null hypothesis is wrongly rejected σ percent of times, if it is in fact true
  - default value $\sigma = 0.05$
- σ defines sensitivity of the test:
  - smaller σ implies larger critical value (larger tolerance)
- tail: specifies whether the K-S performs two sided test (default) or tests from one or other side of the candidate distribution

Output parameters

- Test statistic k is the maximum difference over all data points: $$k = \max_{1 \le i \le N} \left| F(Y_i) - \frac{i}{N} \right|$$ where F is the CDF of the assumed distribution
- The null hypothesis is accepted if the value of the test statistic is smaller than the critical value
- p-value is probability level when the difference between distributions (test statistics) becomes significant:
  - if p-value $\leqslant \sigma$: test rejects the null hypothesis
- If test returns critical value = NaN, the decision to accept or reject null hypothesis is based only on p-value

Traffic data

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

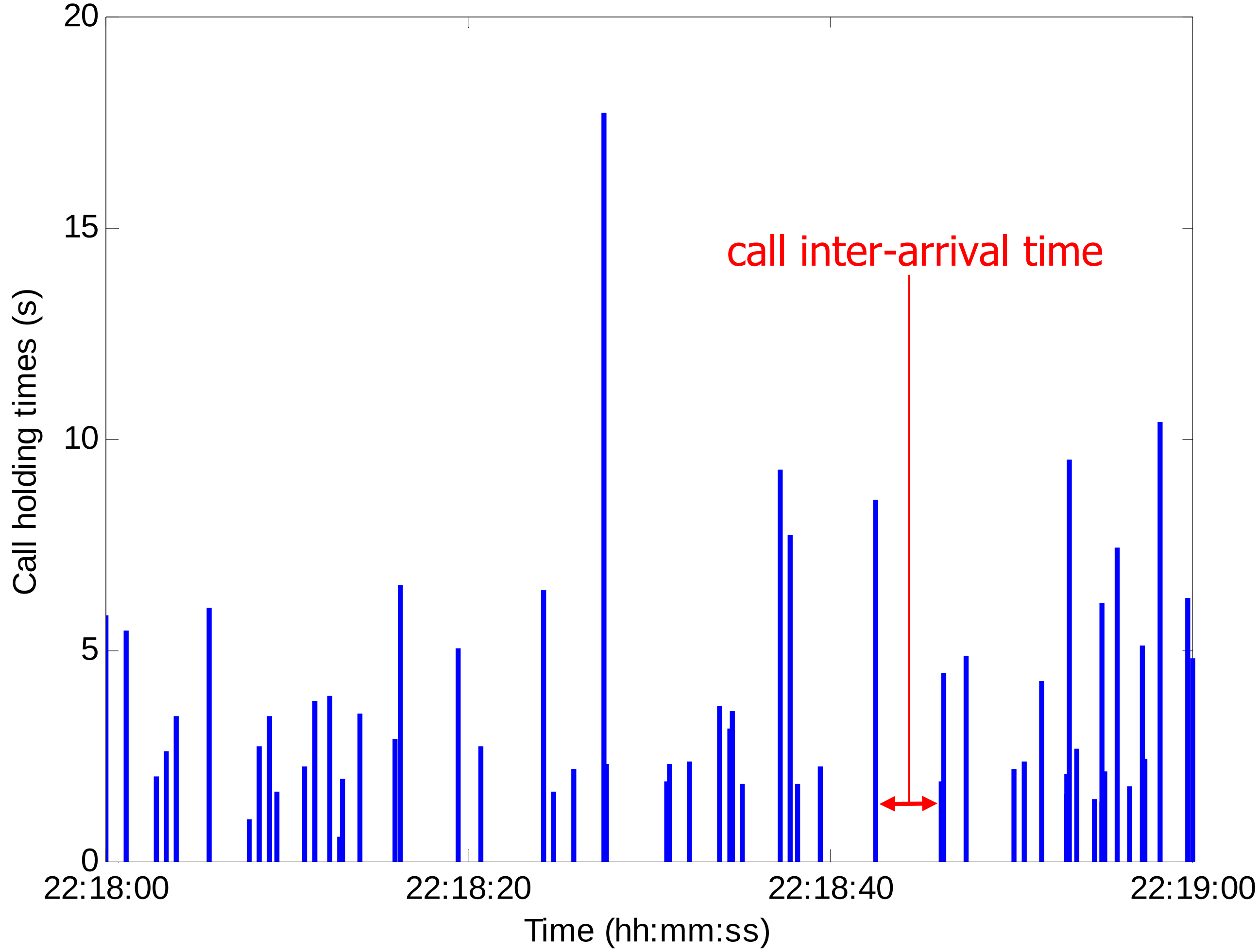
Trace (dataset)	Time span	No. of established calls
2001	November 1–2, 2001	110,348
2002	March 1–7, 2002	370,510
2003	March 24–30, 2003	387,340

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:
  - exponetial, 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
  - K-S test on potential candidates

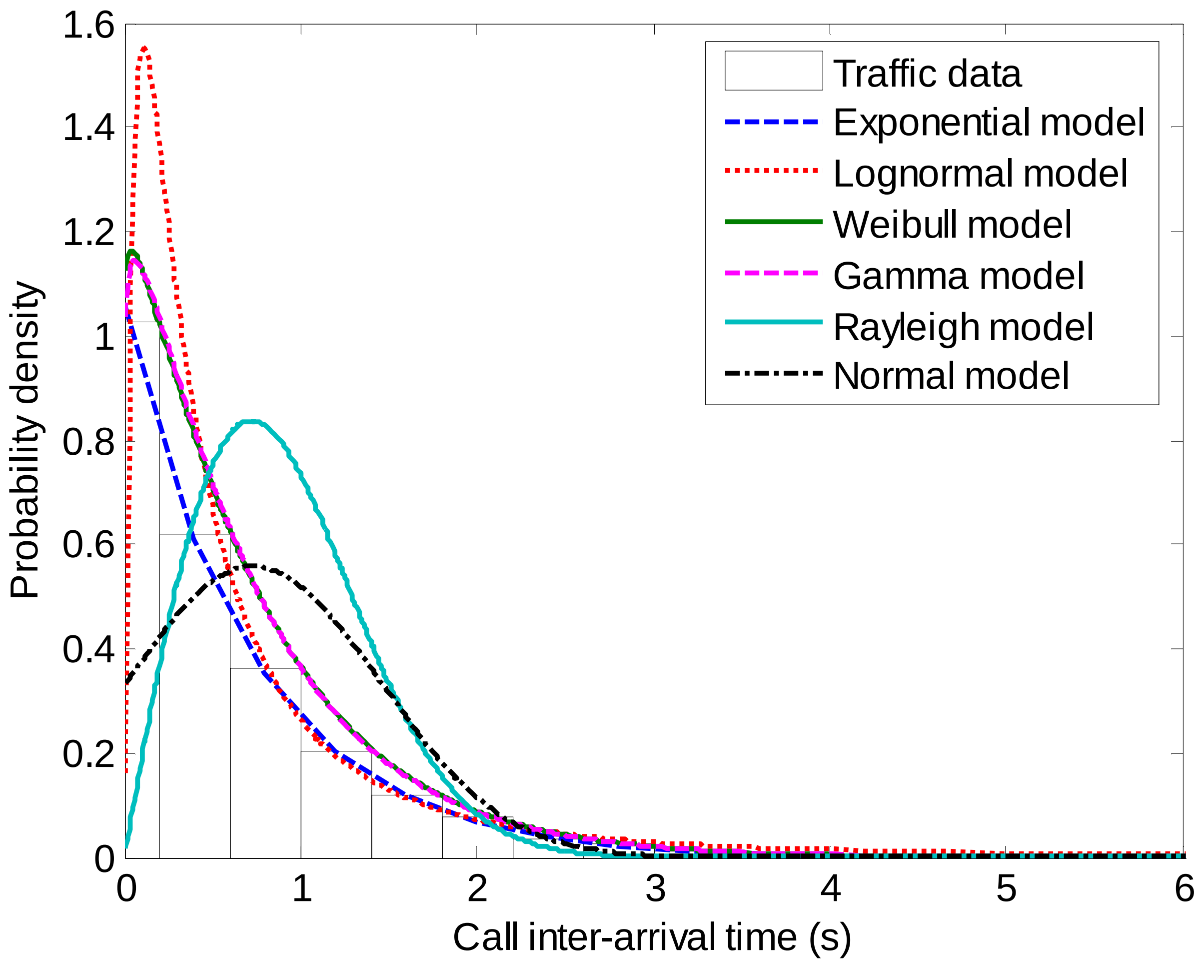
Maximum likelihood estimation (MLE)

- Introduced by R. A. Fisher in 1920s
- The most popular method for parameter estimation
- Goal: to find the distribution parameters that make the given distribution that follow the most closely underlying data set
- Conduct an experiment and obtain *N* independent observations
- $\theta_1, \theta_2, ..., \theta_k$ are *k* unknown constant parameters which

$$L(x_1, x_2, ..., x_N | \theta_1, \theta_2, ..., \theta_k) = L = \prod_{i=1}^{N} f(x_i; \theta_1, \theta_2, ..., \theta_k)$$

$$i = 1, 2, ..., N$$

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

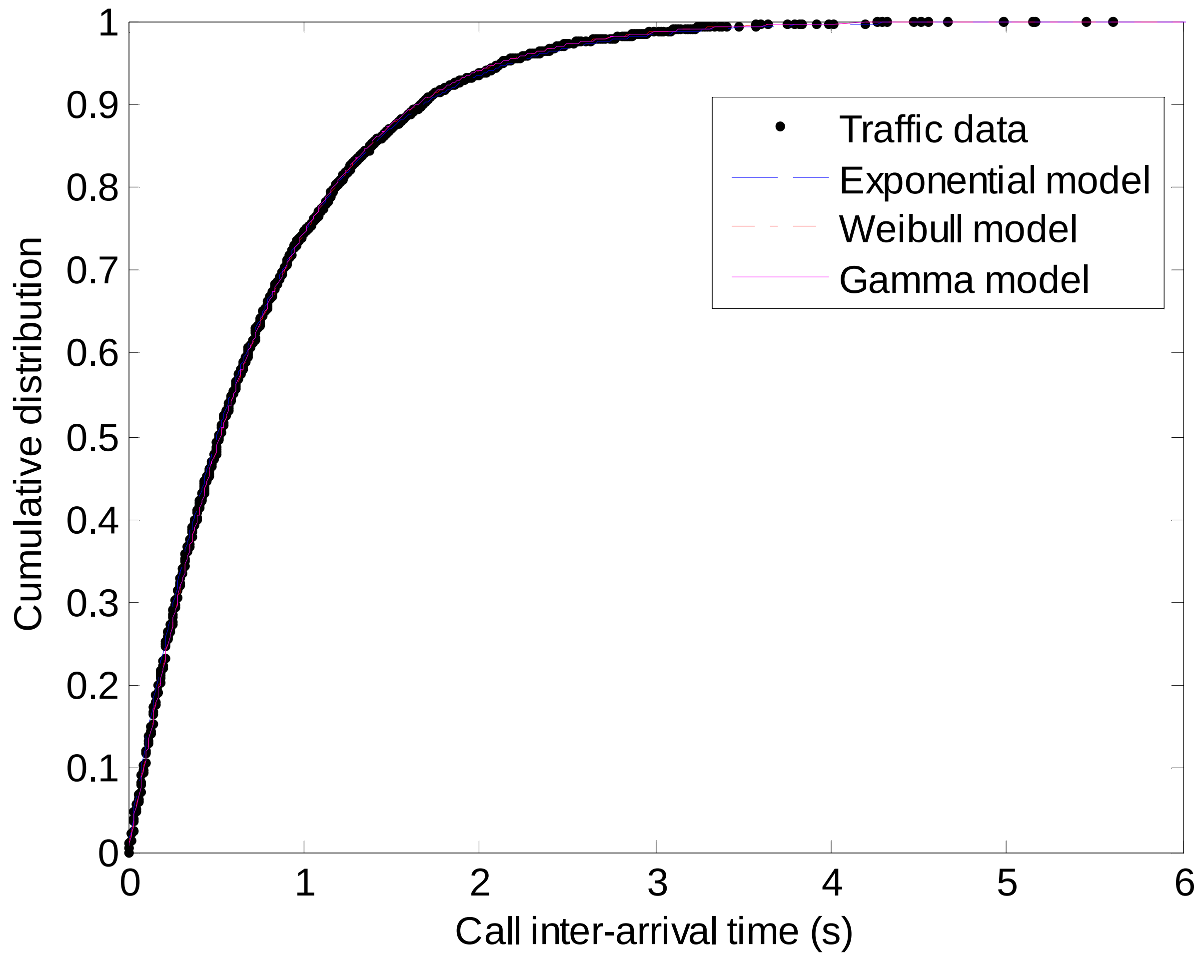
K-S test: call inter-arrival times 2001

Significance level σ = 0.1

Distribution	Parameter	02.11.2001, 20:00–21:00	02.11.2001, 16:00–17:00	02.11.2001, 15:00–16:00	01.11.2001, 19:00–20:00	01.11.2001, 00:00–01:00
exponential	h	1	1	0	1	1
	p	0.0384	0.0001	0.5416	0.0122	0.0135
	k	0.0247	0.0369	0.0131	0.0277	0.0259
Weibull	h	0	1	0	0	1
	p	0.3036	0.0409	0.4994	0.1574	0.0837
	k	0.0171	0.0236	0.0136	0.0195	0.0206
gamma	h	0	1	0	1	1
	p	0.3833	0.0062	0.3916	0.0644	0.0953
	k	0.0159	0.0287	0.0148	0.0227	0.0202

Significance level σ	0.01	0.04	0.05	0.08	0.09	0.1
02.11.2001, 16:00–17:00: cv	0.0275	0.0237	0.0230	0.0215	0.0211	0.0207
01.11.2001, 00:00–01:00: cv	0.0267	0.0229	0.0223	0.0208	0.0204	0.0201

Call inter-arrival times: best-fitting distributions (cdf)

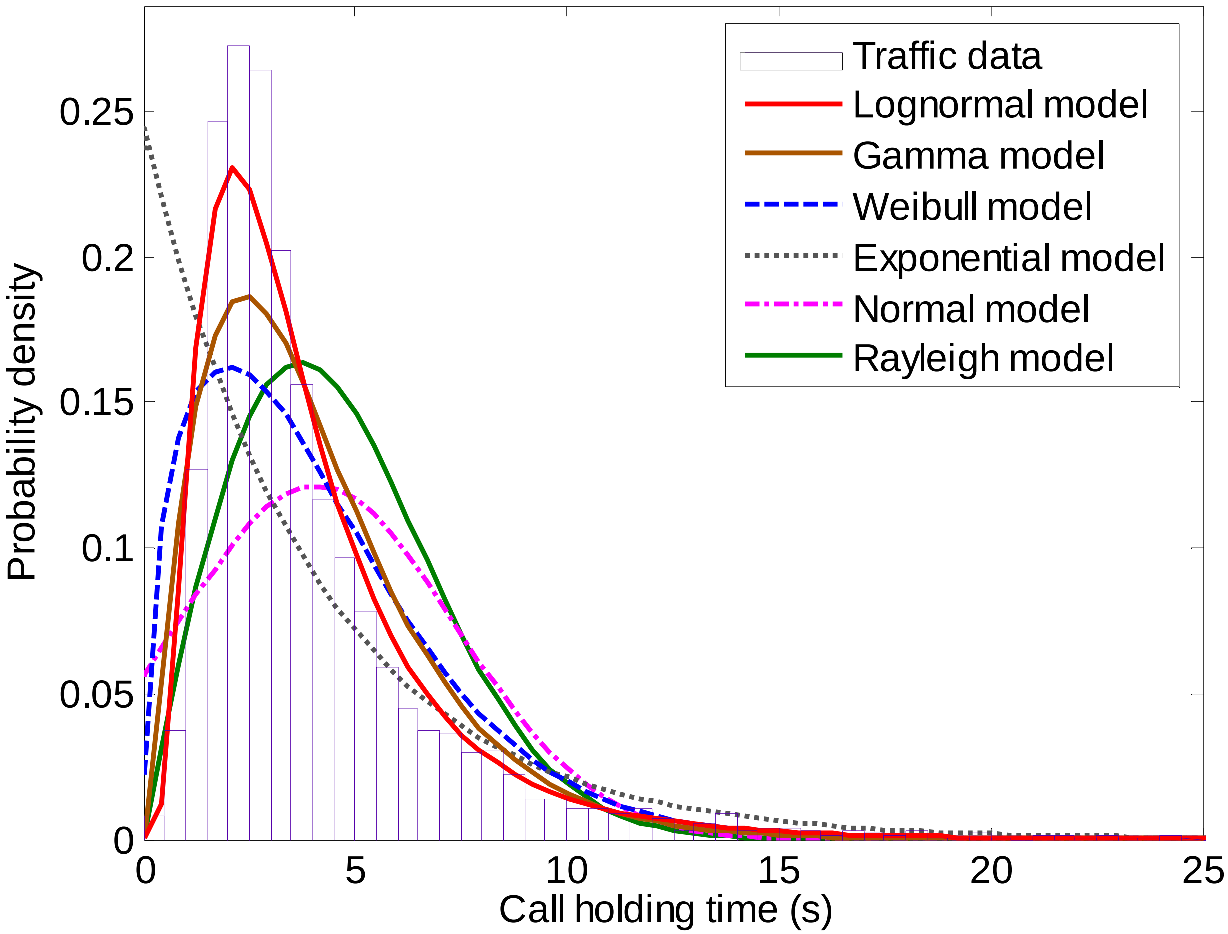


Call inter-arrival times: estimates of H

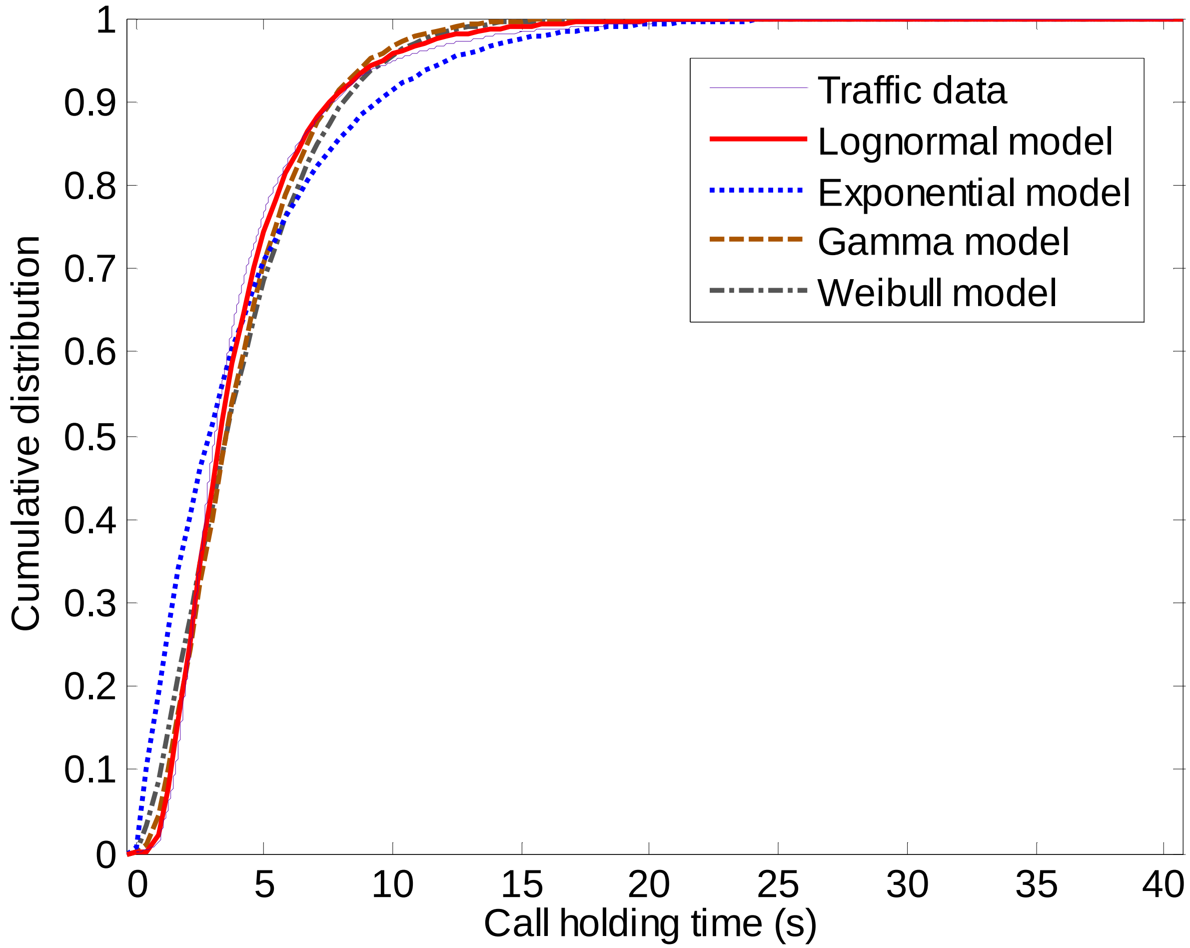
- Traces pass the test for time constancy of α: estimates of H are reliable

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: best-fitting distributions (cdf)



Call holding times: K-S test results (2003 data)

- No distribution passes the test when the entire trace is tested (significance levels = 0.1 and 0.01)
- Lognormal distribution passes test (significance level = 0.01) for:
  - 5-6 sub-traces from 15 randomly chosen 1,000-sample sub-traces
  - passes the test for almost all 500-sample sub-traces
- Test rejects null hypothesis when the sub-traces are compared with candidate distributions:
  - exponential
  - Weibull
  - gamma

Call holding times: estimates of H

- All (except one) traces pass the test for constancy of α
- 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

- 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: preprocessing

- Collected data contain continuous data records from 92 days: March 1st 2003 – May 31st 2003
- Original database: ~6 GBytes, with 44,786,489 record rows:
  - contains event log tables recording network activities
  - aggregated from distributed database of individual network management systems
  - sorted in 92 event log tables, each containing one day's events
- 9 (out of original 26) fields are of interest for our analysis

Traffic data: preprocessing

- 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

Traffic data: sample

Date	Time	Ms	Duration	Sys_id	Chl_id	Caller	Callee	C_type	C_state	Multi
2003-03-20	00:00:01	450	3730	8	4	6155	1801	0	0	0
2003-03-20	00:00:01	469	3730	6	7	6155	1801	0	0	0
2003-03-20	00:00:01	560	3730	3	7	6155	1801	0	0	0
2003-03-20	00:00:01	570	3730	2	7	6155	1801	0	0	0
2003-03-20	00:00:01	640	3730	1	7	6155	1801	0	0	0
2003-03-20	00:00:01	880	5260	9	6	13314	251	0	0	0
2003-03-20	00:00:01	910	5260	7	6	13314	251	0	0	0
2003-03-20	00:00:01	970	5260	6	8	13314	251	0	0	0
2003-03-20	00:00:01	980	2520	7	7	13911	418	0	0	0
2003-03-20	00:00:02	29	5270	4	2	13314	251	0	0	0
2003-03-20	00:00:02	109	5260	2	8	13314	251	0	0	0
2003-03-20	00:00:02	139	5270	1	8	13314	251	0	0	0
2003-03-20	00:00:02	9	2510	6	1	13911	418	0	0	0
2003-03-20	00:00:02	149	2510	2	9	13911	418	0	0	0
2003-03-20	00:00:05	289	3560	8	5	6011	2035	0	0	0
2003-03-20	00:00:05	309	3550	6	3	6011	2035	0	0	0
2003-03-20	00:00:05	389	3560	3	2	6011	2035	0	0	0
2003-03-20	00:00:05	449	3550	2	2	6011	2035	0	0	0
2003-03-20	00:00:05	480	3550	1	9	6011	2035	0	0	0
2003-03-20	00:00:05	550	3440	1	12	7614	945	0	0	0
2003-03-20	00:00:05	550	3440	2	3	7614	945	0	0	0
2003-03-20	00:00:05	949	9780	6	4	15840	418	0	0	0
2003-03-20	00:00:05	959	9780	7	2	15840	418	0	0	0
2003-03-20	00:00:06	679	3040	2	6	13931	471	0	0	0
2003-03-20	00:00:06	709	3040	1	2	13931	471	0	0	0
2003-03-20	00:00:06	130	9780	2	4	15840	418	0	0	0
2003-03-20	00:00:08	109	6640	9	2	13420	251	0	0	0
2003-03-20	00:00:08	179	6630	7	3	13420	251	0	0	0
2003-03-20	00:00:08	200	6640	6	5	13420	251	0	0	0
2003-03-20	00:00:08	270	6630	4	5	13420	251	0	0	0
2003-03-20	00:00:08	329	6640	1	4	13420	251	0	0	0
2003-03-20	00:00:08	340	6640	2	7	13420	251	0	0	0

Traffic data: sample

- Traffic data after cleaning and extraction:

Date	Time	Ms	Duration	Caller	Callee	C_type	C_state	Multi	# Sys	System List
2003-03-20	00:00:01	450	3730	6155	1801	0	0	0	5	8,6,3,2,1
2003-03-20	00:00:01	980	2520	13911	418	0	0	0	3	7,6,2
2003-03-20	00:00:01	880	5260	13314	251	0	0	0	6	9,7,6,4,2,1
2003-03-20	00:00:05	550	3440	7614	945	0	0	0	2	1,2
2003-03-20	00:00:05	289	3560	6011	2035	0	0	0	5	8,6,3,2,1
2003-03-20	00:00:05	949	9780	15840	418	0	0	0	3	6,7,2
2003-03-20	00:00:06	810	2350	8022	817	0	0	0	1	1
2003-03-20	00:00:06	819	1590	13902	497	0	0	0	4	10,9,8,4
2003-03-20	00:00:06	440	3030	13931	471	0	0	0	5	10,9,4,2,1
2003-03-20	00:00:08	109	6640	13420	251	0	0	0	6	9,7,6,4,1,2

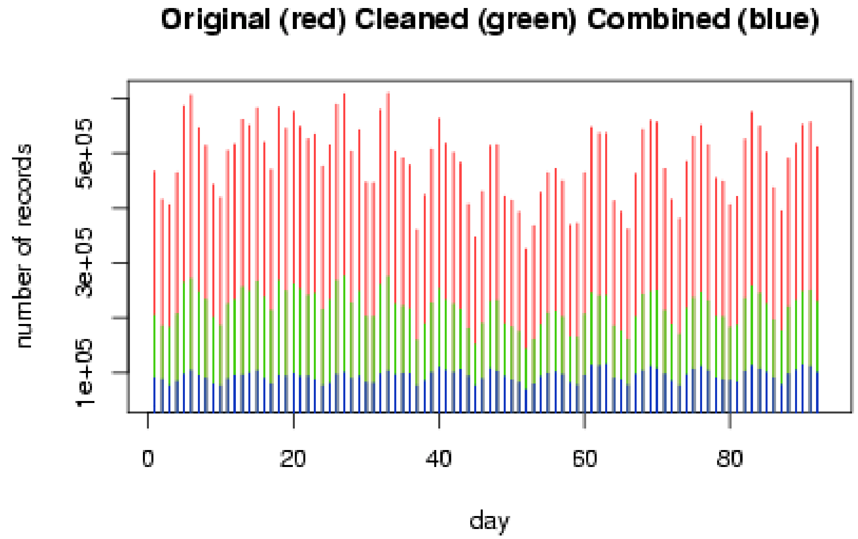
Sample of processed data: 2003-03-01

No	Time (hh:mm:ss)(ms)	Call duration (ms)	System ID	Channel ID	Caller	Callee
1	00:00:00 30	1340	1	12	13905	401
6	00:00:00 489	1350	7	4	13905	401
29	00:00:03 620	7550	2	7	13233	249
31	00:00:03 760	7560	1	3	13233	249
37	00:00:04 260	7560	7	6	13233	249
38	00:00:04 340	7560	6	6	13233	249

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%

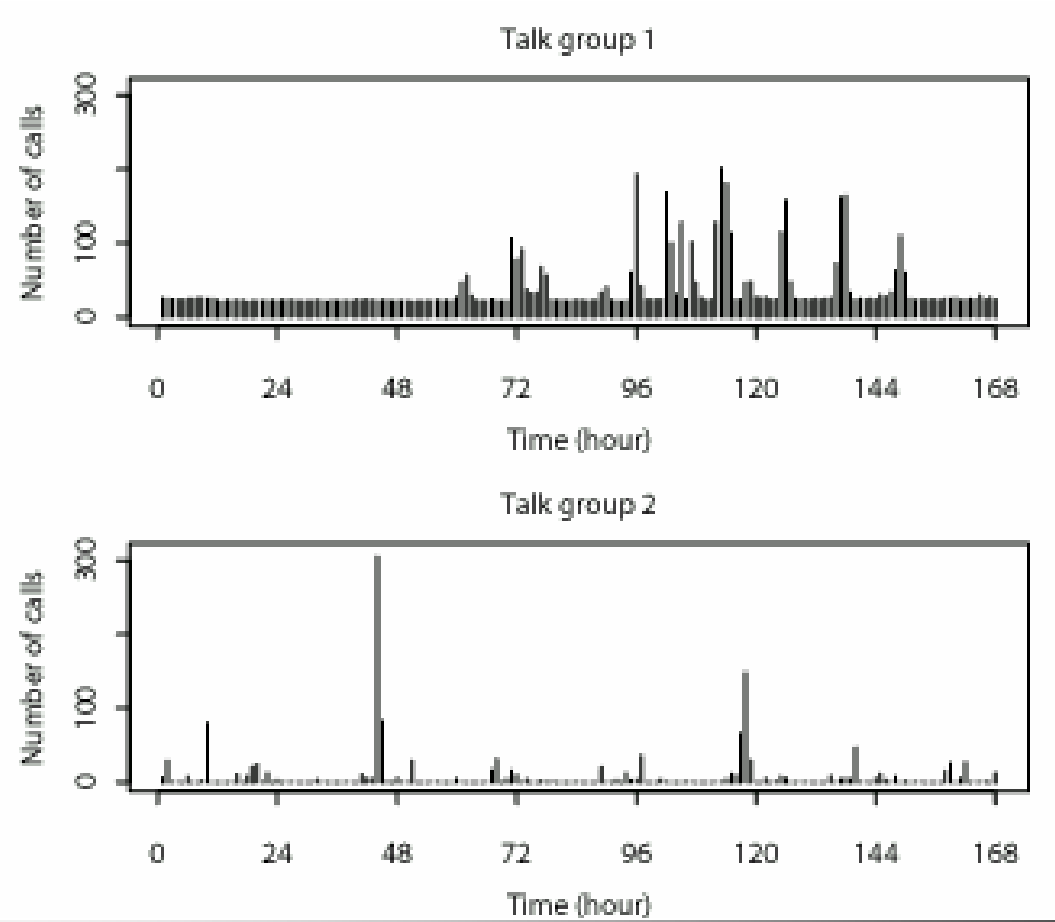


Talk group data

- The hourly number of calls is a common metric employed in telecommunication industry
- It is a footprint of users' calling behavior
- Time scales:
  - finer than an hour are too small to record the calling activity
  - larger than an hour are too coarse to capture user behavior patterns
- Traffic data were collected during 92 days (2,208 hours)
- The 2,208 hourly calls captured each talk group's calling behavior

Calling patterns: talk groups 1 and 2

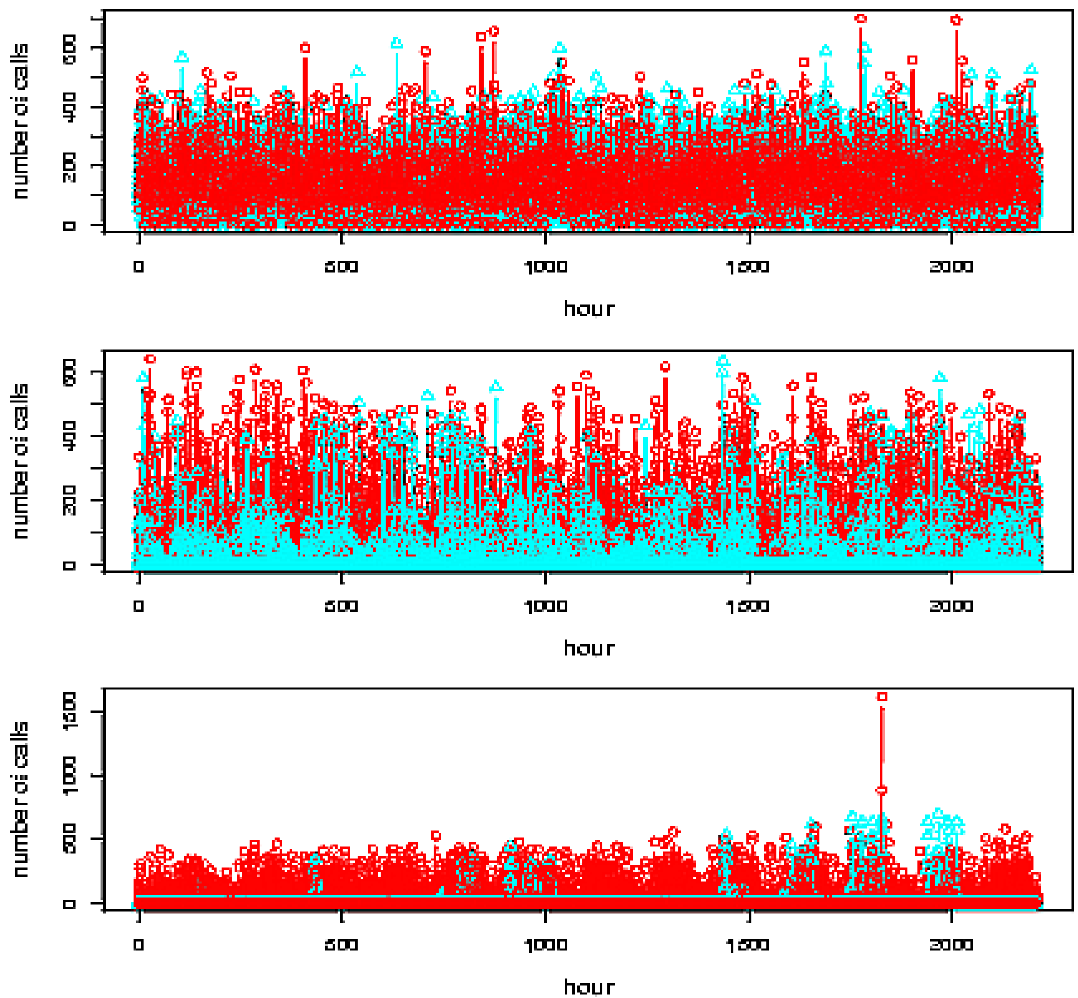
- Calling patterns are used for classification:



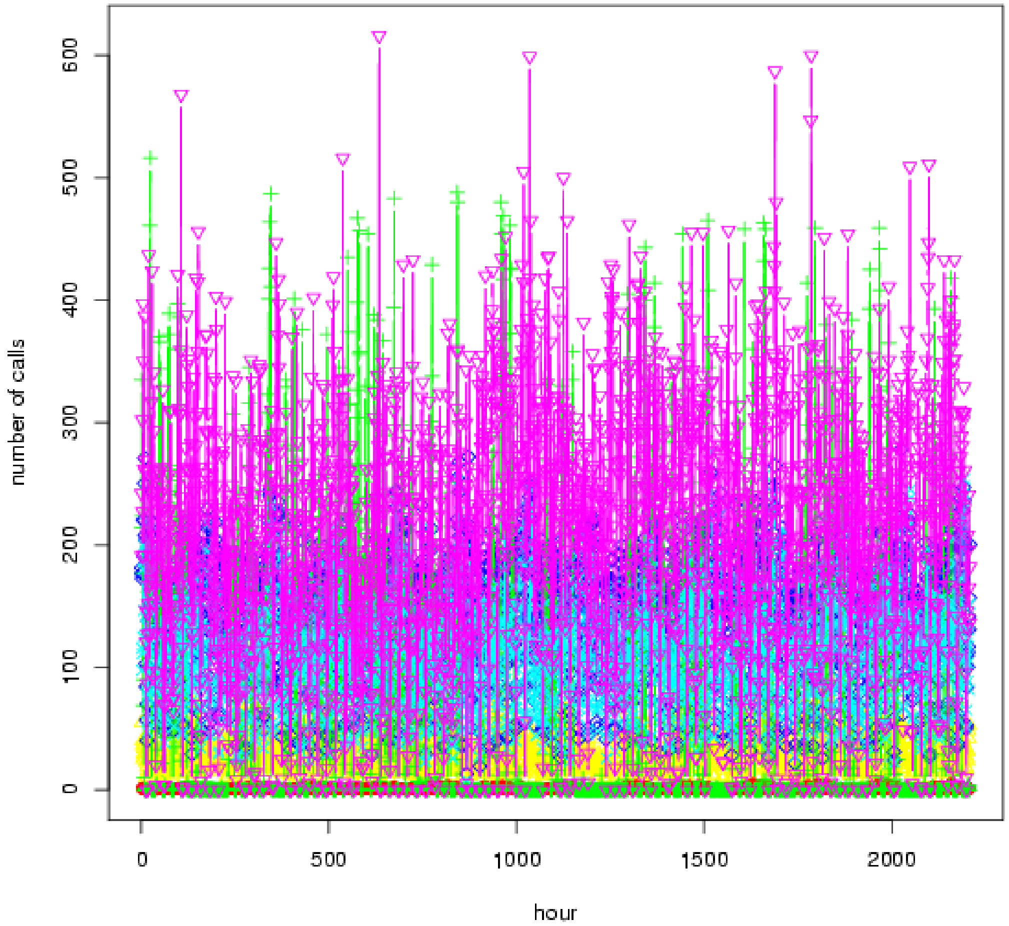
User clustering with K-means: k = 3

- First cluster (heavy network users):
  - 17 talk groups, contributing to 59% of the calls, with an average number of calls ranging from 94 to 208 per hour
  - They are dispatch groups that assign and schedule other talk groups for specific tasks
- Second cluster (average network users):
  - 31 talk groups, contributing to 26% of the calls
- Third cluster (least frequent network users):
  - 569 talk groups, contributing to only 15% of the calls
  - They represent over 90% of all talk groups

User clusters with K-means: k = 3



User clusters with K-means: k = 6



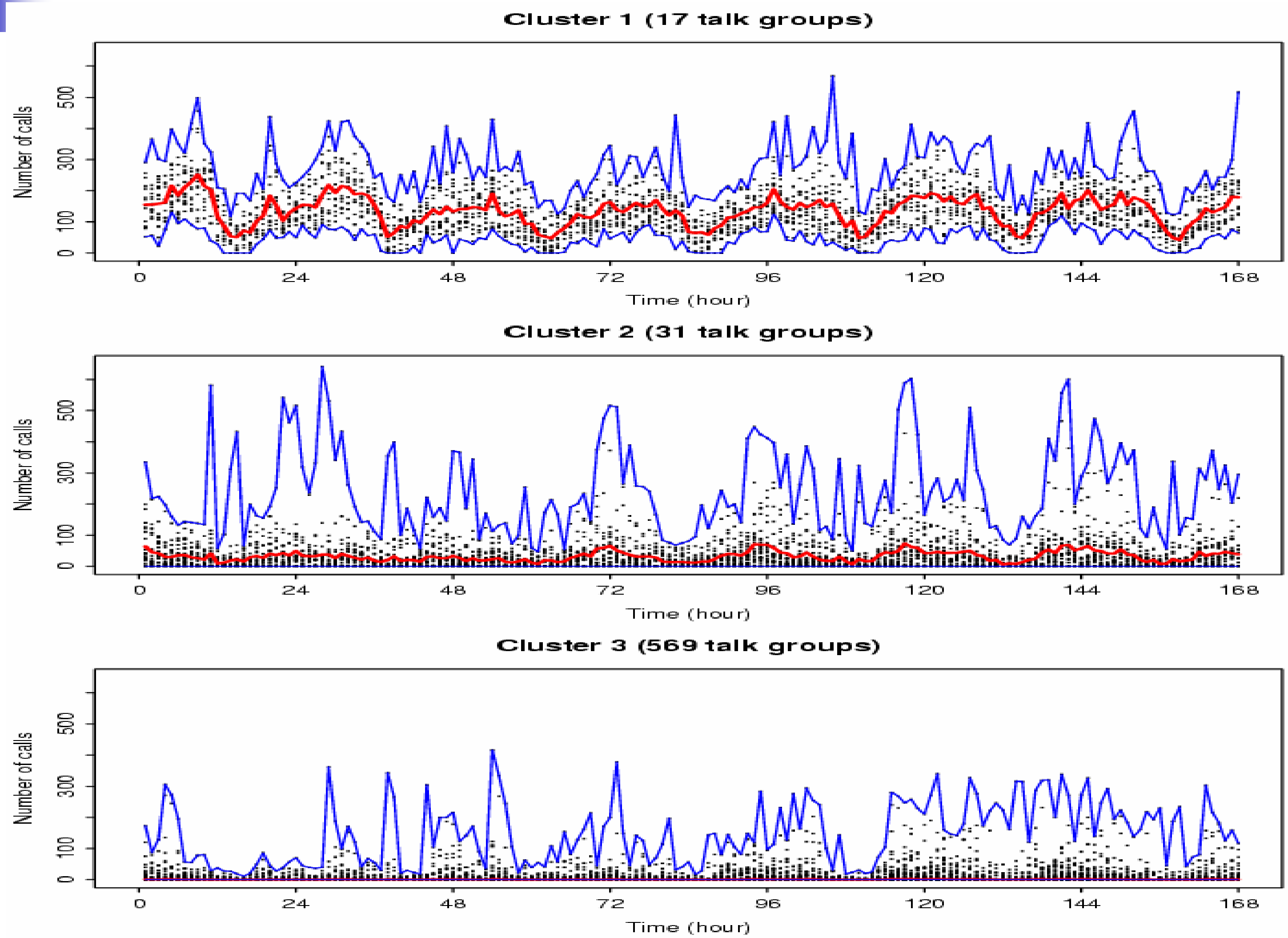
Clustering results

- 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 clustering: cluster distances and silhouette coefficient

K	Average intra-cluster distance	Average inter-cluster distance	Maximum intra-cluster distance	Minimum inter-cluster distance	Overall clustering quality	Silhouette coefficient
3	1882.14	4508.38	2971.76	1626.40	-1345.36	0.7756
4	1863.00	3889.12	2971.76	1556.68	-1415.07	0.7684
6	2059.67	3284.52	3299.43	594.21	-2705.21	0.7640
9	1020.08	3520.04	3065.25	808.28	-2256.96	0.7492
12	1372.67	3582.98	3278.14	731.26	-2546.88	0.7435
16	983.63	1815.79	3571.27	248.19	-3323.07	0.7337
20	1355.80	2458.39	3604.33	314.49	-3289.84	0.7386

K-means clustering: number of calls in the three clusters



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

Prediction based on aggregate traffic

- The aggregate network traffic consists of all network users' traffic
- The R system was used to identify, estimate, and verify the SARIMA model for the aggregate users' traffic
- Both 24-hour (one day) and 168-hour (one week) intervals were selected as seasonal period parameters
- Based on m past traffic data samples, we forecast the future n traffic data samples
- The prediction quality was measured using the normalized mean square error nmse:

$$nmse(a,b) = \sum_{i=m+1}^{m+n} \frac{(a_i - b_i)^2}{a_i^2},$$

- where: a_i is the observed and b_i is the predicted data

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

SARIMA models: summary of selection criteria

(p,d,q) x (P,D,Q)s	m	nmse	AIC	AICc	BIC
(2,0,9) x (0,1,1)24	1680	0.379	22744.6	22744.9	22826.8
(2,0,1) x (0,1,1)168	1680	0.174	23129.8	23129.8	23161.9
(1,0,1) x (0,1,1)168	1680	0.175	23145.1	23145.1	23170.8
(2,0,9) x (1,1,1)24	1680	0.525	25292.1	25292.4	25382.1
(1,0,2) x (1,1,1)24	1680	0.411	25332.6	25332.6	25371.2
(2,0,1) x (0,1,1)24	1680	0.408	25360.5	25360.6	25392.6
(3,0,1) x (0,1,1)24	1680	0.404	25361.2	25361.2	25399.7

Prediction: based on the aggregate traffic

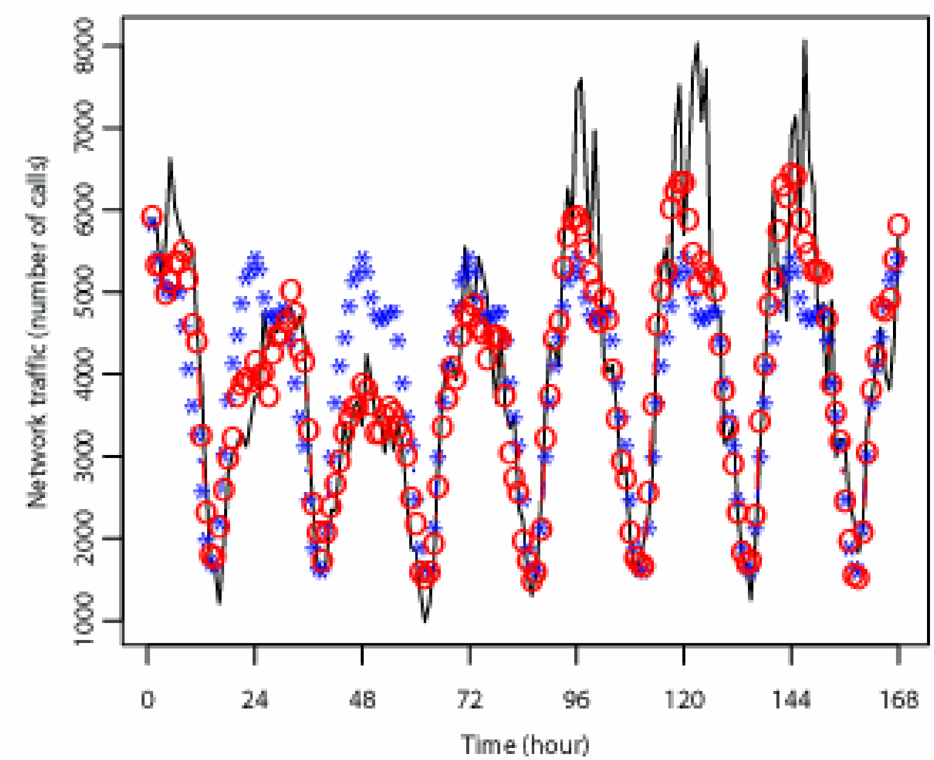
No.	p	d	q	P	D	Q	S	m	n	nmse
A1	2	0	9	0	1	1	24	1512	672	0.3790
A2	2	0	1	0	1	1	24	1512	672	0.3803
A3	2	0	9	0	1	1	168	1512	672	0.1742
A4	2	0	1	0	1	1	168	1512	672	0.1732
B1	2	0	9	0	1	1	24	1680	168	0.3790
B2	2	0	1	0	1	1	24	1680	168	0.4079
B3	2	0	9	0	1	1	168	1680	168	0.1736
B4	2	0	1	0	1	1	168	1680	168	0.1745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

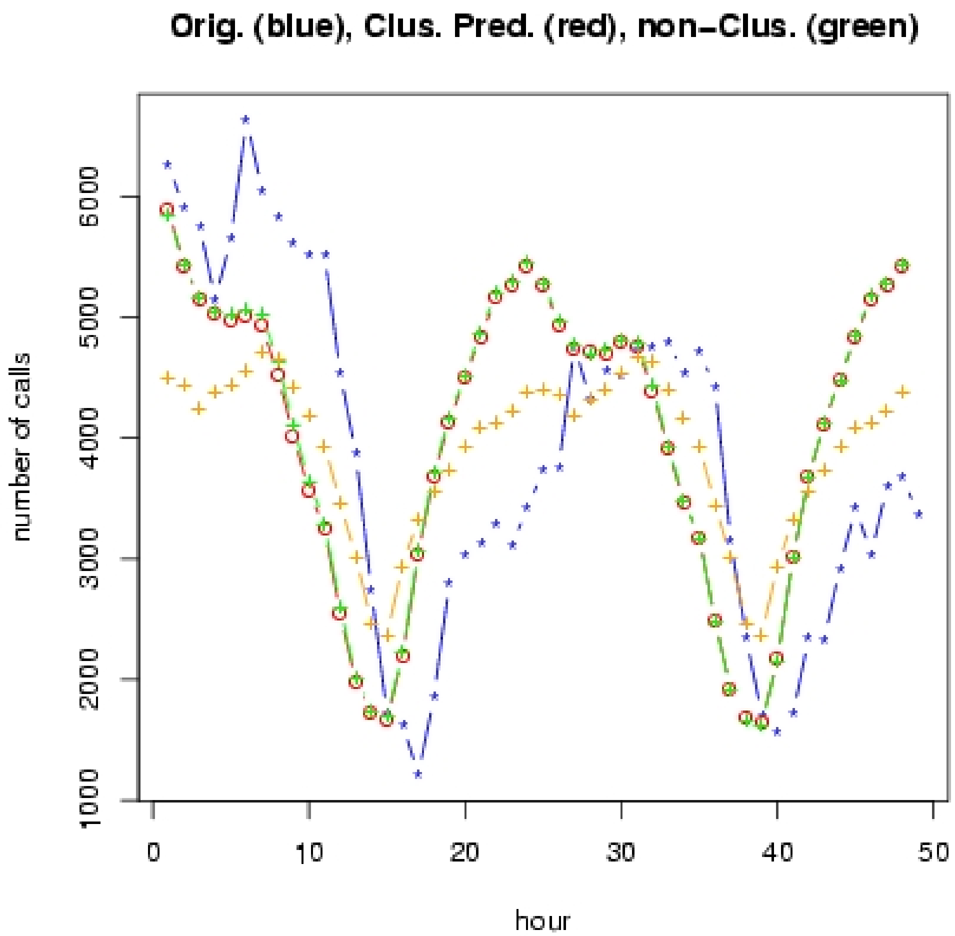
Prediction with user clustering

- Raw network log data collected over 92 days: March 1st 2003 – May 31st 2003
- Footprint of network usage for talk groups: the hourly number of calls
- AutoClass and the K-means algorithm were used to classify network talk groups into clusters
- The behavior of each user cluster was predicted using Seasonal Autoregressive Integrated Moving Average (SARIMA)
- We used aggregation to predict the overall network behavior

Traffic prediction based on user clusters

- The developed aggregate users based prediction assumes that the adopted model is static: the number of network users and their behavior pattern are constant in time
- This assumption does not hold when planning further network expansions and cannot be used to forecast network traffic
- We employed a user clusters based prediction approach to predict the network traffic by accumulating the prediction results from user clusters
- In large networks with many users, it is impractical to predict individual users' traffic and then aggregate the predicted data
- With user clusters, traffic prediction is reduced to predicting and aggregating users' traffic from few clusters

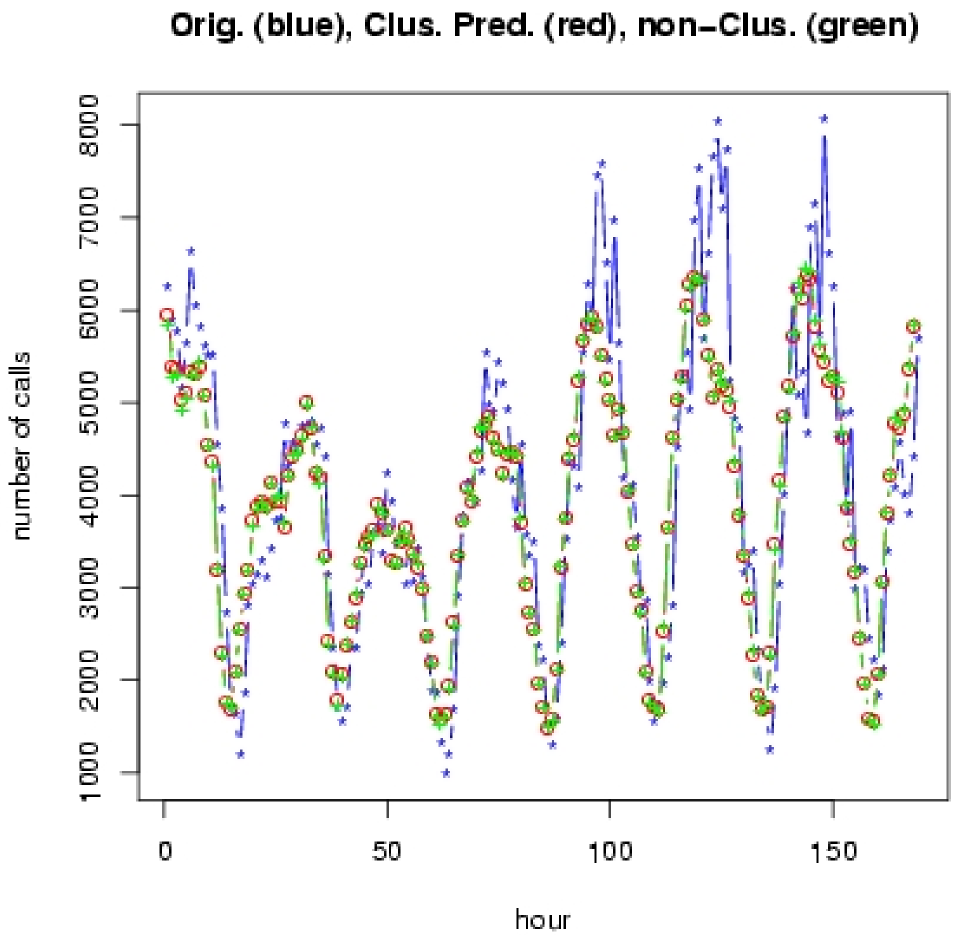
Prediction of 48 hours of traffic based on 1,680 past hours



Comparisons: model $(2,0,1)x(0,1,1)_{24}$

* observation
* prediction without clustering
o prediction with clustering
+ optimized prediction with clustering

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



Comparisons: model $(1,0,1)\mathrm{x}(0,1,1)_{168}$
* observation
* prediction without clustering
o prediction with clustering

Traffic prediction with user clusters: examples (2,0,1) x (0,1,1)

Cluster	S	m	n	nmse
1	168	1,920	24	0.2241
2	168	1,920	24	0.3818
3	168	1,920	24	0.1163
*	168	1,920	24	0.0969
A	168	1,920	24	0.1175
1	24	1,920	24	0.2508
2	24	1,920	24	0.2697
3	24	1,920	24	0.3020
*	24	1,920	24	0.1941
A	24	2,920	24	0.2052
1	24	1,680	168	0.5477
2	24	1,680	168	0.6883
3	24	1,680	168	0.2852
*	24	1,680	168	0.4079
A	24	1,680	168	0.4093

Prediction results with user clusters

- For each group, rows 1, 2, and 3: traffic prediction results for user clusters 1, 2, and 3
- Row *: the aggregate user traffic prediction obtained without clustering the users
- Row A: the aggregate prediction of network traffic based on the three user clusters
- The performance of the clusters based prediction (nmse: 0.1175) is comparable to the best prediction based on aggregate traffic (nmse: 0.0969)
- 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

Prediction based on user clusters model $(2, 0, 1)\times(0, 1, 1)$

Test no.	S	m	n	nmse cluster 1	nmse cluster 2	nmse cluster 3	nmse aggregate	nmse cluster	nmse optimized
1	24	240	24	0.323	0.548	0.308	0.254	0.241	n/a
2	24	240	48	0.394	0.712	0.445	0.343	0.332	n/a
3	24	1200	72	1.774	1.976	0.270	0.884	0.886	0.846
4	24	1200	96	1.319	0.866	0.260	0.611	0.613	0.610
5	24	1200	120	0.840	0.703	0.245	0.463	0.467	n/a
6	24	1200	144	0.665	0.647	0.236	0.396	0.399	n/a
7	168	1008	336	0.616	0.466	0.190	0.285	0.260	n/a
8	168	1008	504	0.439	0.446	0.190	0.237	0.224	n/a
9	168	1176	24	3.401	0.747	0.168	0.365	0.507	0.436
10	168	1512	504	0.348	0.375	0.155	0.180	0.178	n/a
11	168	1680	24	0.367	0.444	0.115	0.132	0.129	n/a
12	168	1680	48	0.380	0.467	0.095	0.114	0.116	n/a

Traffic prediction with user clusters

- nmse > 1.0 for cluster 1 (tests 3, 4, and 9) and for cluster 2 (test 3) implies that prediction is worse than prediction based on the mean value of past data
- Mean value prediction leads to better prediction results shown in column "nmse optimized" (optimized cluster-based prediction) for:
  - Test 3: clusters 1 and 2
  - Test 4: cluster 1
- Prediction based on clusters performs better than the prediction based on aggregate traffic:
  - Tests 1, 2, 7, 8, 10, and 11

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

Traffic prediction: summary

- We analyzed traffic data collected from an operational trunked radio network
- By applying data mining techniques (K-means algorithm and AutoClass) on traffic data, we discovered user clusters based on patterns of calling behavior expressed by hourly number of calls
- Network traffic was predicted using the SARIMA model based on aggregate user traffic and based on user clusters
- Proposed cluster-based prediction produces comparable results to prediction based on aggregate traffic
- It is applicable to networks with variable number of users where prediction based on aggregate traffic could not be applied

Road map

- Introduction
- Networks and traffic data:
  - data collection
  - statistical analysis
  - traffic prediction
- Case study:
  - wireless network: E-Comm
- Conclusions and references

Conclusions

- We re-used simulation tools and analytical methods to analyze traffic data from the E-Comm network:
- Network:
  - network performance was evaluated using simulation tools (OPNET and WarnSim)
- Traffic characterization and modeling:
  - models of inter-arrival and call holding times were developed
- Users:
  - clustering algorithms (K-means and AutoClass) were employed to classify network users into user clusters
- Traffic prediction:
  - SARIMA models were used to predict network traffic based on aggregate user traffic and based on three user clusters

References: downloads

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

- 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.