User Clustering and Traffic Prediction in a Trunked Radio System

Hao (Leo) Chen

lcheu@cs.sfu.ca

Simon Fraser University

ROADMAP

- Introduction
- E-Comm network
- Traffic data
- User clustering
- Traffic prediction
- Conclusions
- Reference

ROADMAP

- Introduction
- E-Comm network
- Traffic data
- User clustering
- Traffic prediction
- Conclusions
- Reference

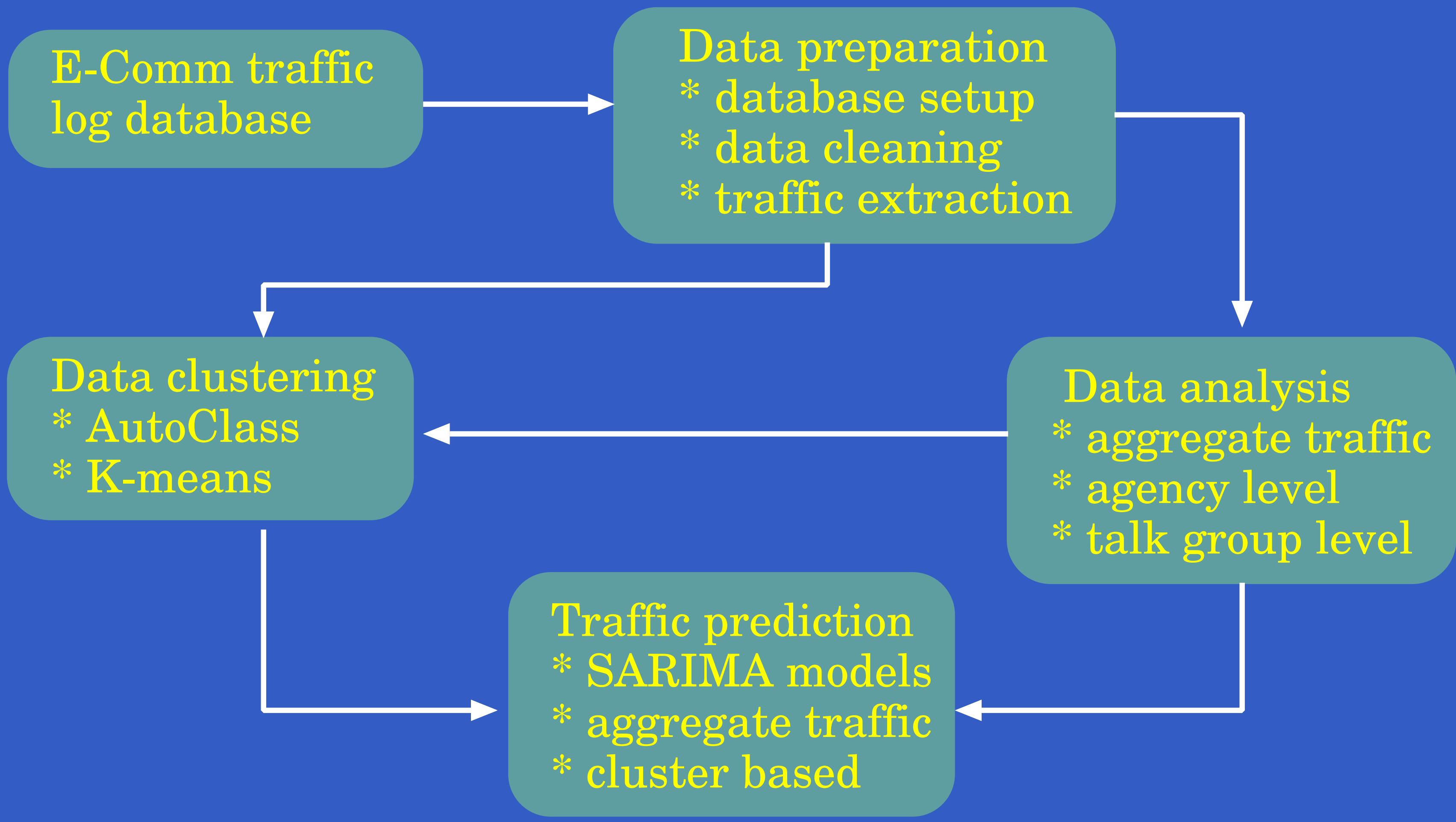
MOTIVATION

- Analysis of traffic from operational wireless networks enables:
  - better understanding of user behavior patterns
  - better quality of service.
- Traffic prediction methods:
  - “Top-down” approach: based on aggregate traffic.
  - “Bottom-up” approach: focuses on individual users.
  - Our approach: user cluster based prediction.

PRIOR DATA ANALYSIS

- User behavior and mobility patterns exhibit daily and weekly patterns [Tang and Baker, 1999].
- User behavior in Cellular Digital Packet Data (CDPD) mobile wireless networks has similar cyclic patterns [Andriantiatsaholiniaina and Trajković, 2002].
- Trunked radio network traffic [Sharp et al., 2004]:
  - call holding time distribution is approximately lognormal
  - call inter-arrival time is closely approximated by an exponential distribution.

OUR RESEARCH



ROADMAP

- Introduction
- E-Comm network
- Traffic data
- User clustering
- Traffic prediction
- Conclusions
- Reference

E-COMM NETWORK

- Regional emergency communications center.
- Covers Greater Vancouver Regional District (GVRD) – 11 systems/cells.
- Provides emergency dispatch/communication services.
- Serves 16 agencies such as RCMP, fire and rescue, police, and ambulance.
- Employs Enhanced Digital Access Communications System (EDACS).

E-COMM NETWORK COVERAGE

GROUP/MULTI-SYSTEM CALLS

- A group call is a standard call made in a trunked radio system.
- EDACS network operators have observed that more than 85% of calls are group calls.
- A multi-system call is a single group call involving more than one system/cell.
- More than 55% of group calls are multi-system calls.

ROADMAP

- Introduction
- E-Comm network
- Traffic data
- User clustering
- Traffic prediction
- Conclusions
- Reference

TRAFFIC DATA

- Raw event log generated from a distributed database system.
- Events generated in the network from March 1^{st} 00:00:00 2003 to May 31^{st} 23:59:59 2003.
- The size of the original data is $\sim$ 6 GBytes, with 44,786,489 record rows for the 92 days of data.
- From the 26 original fields in the database, 9 fields are of interest for our analysis.

DATA SAMPLE

no.	event_utc_at	dur.	sys.	ch.	caller	callee
1	03-03-01 00:00:00.30	1340	1	12	13905	401
4	03-03-01 00:00:00.259	3330	6	3	14663	249
6	03-03-01 00:00:00.489	1350	7	4	13905	401
7	03-03-01 00:00:00.590	2990	6	4	4266	1443
29	03-03-01 00:00:03.620	7550	2	7	13233	249
30	03-03-01 00:00:03.700	2980	9	7	16068	673
31	03-03-01 00:00:03.760	7560	1	3	13233	249
32	03-03-01 00:00:03.830	1580	2	8	13333	245
37	03-03-01 00:00:04.260	7560	7	6	13233	249
38	03-03-01 00:00:04.340	7560	6	6	13233	249

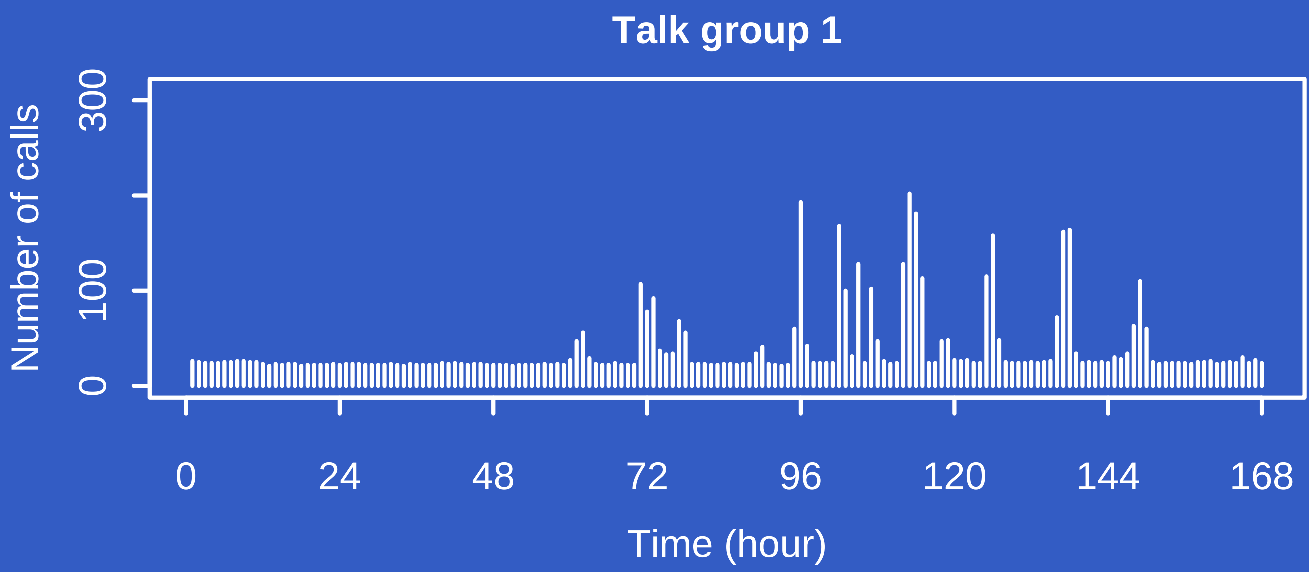
DATA CLEANING/EXTRACTION

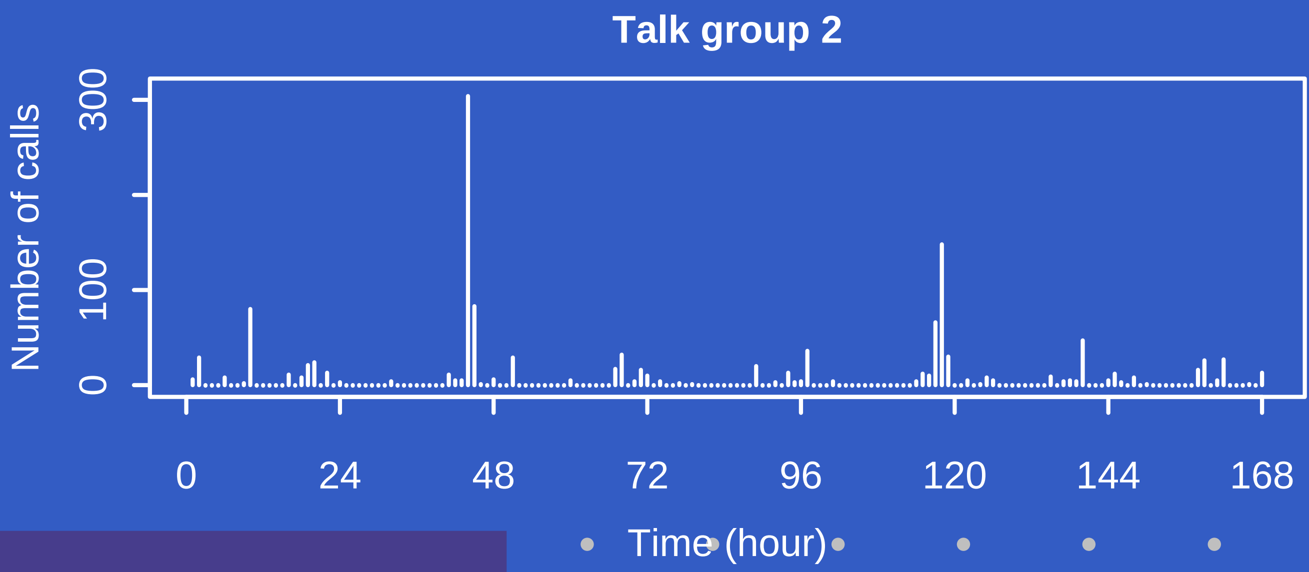
- Data cleaning: reducing database dimension, filtering outliers, removing redundant records.
- Traffic extraction: use single entry to replace multiple records for multi-system calls.
- $\sim$55% records removed after cleaning.
- $\sim$20% records remained after extraction.

CALLING BEHAVIOR PATTERN

- The basic talking unit in the E-Comm network is the talk group and the basic behavior is making a call.
- An important calling behavior pattern in the voice network is the number of calls.
- Hourly number of calls is used to represent the calling behavior pattern of talk groups.
- The collected 92 days of traffic data (2,208 hours) permitted each talk group's calling behavior pattern to be captured by the 2,208 hourly number of calls.

SAMPLE OF CALLING PATTERNS





ROADMAP

- Introduction
- E-Comm network
- Traffic data
- User clustering
- Traffic prediction
- Conclusions
- Reference

CLUSTERING ALGORITHMS

- Clustering analysis groups or segments a collection of objects into 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.
- AutoClass [Cheeseman and Stutz, 1996] and K-means [Kaufman and Rousseeuw, 1990] algorithms are used to classify calling patterns of talk groups.

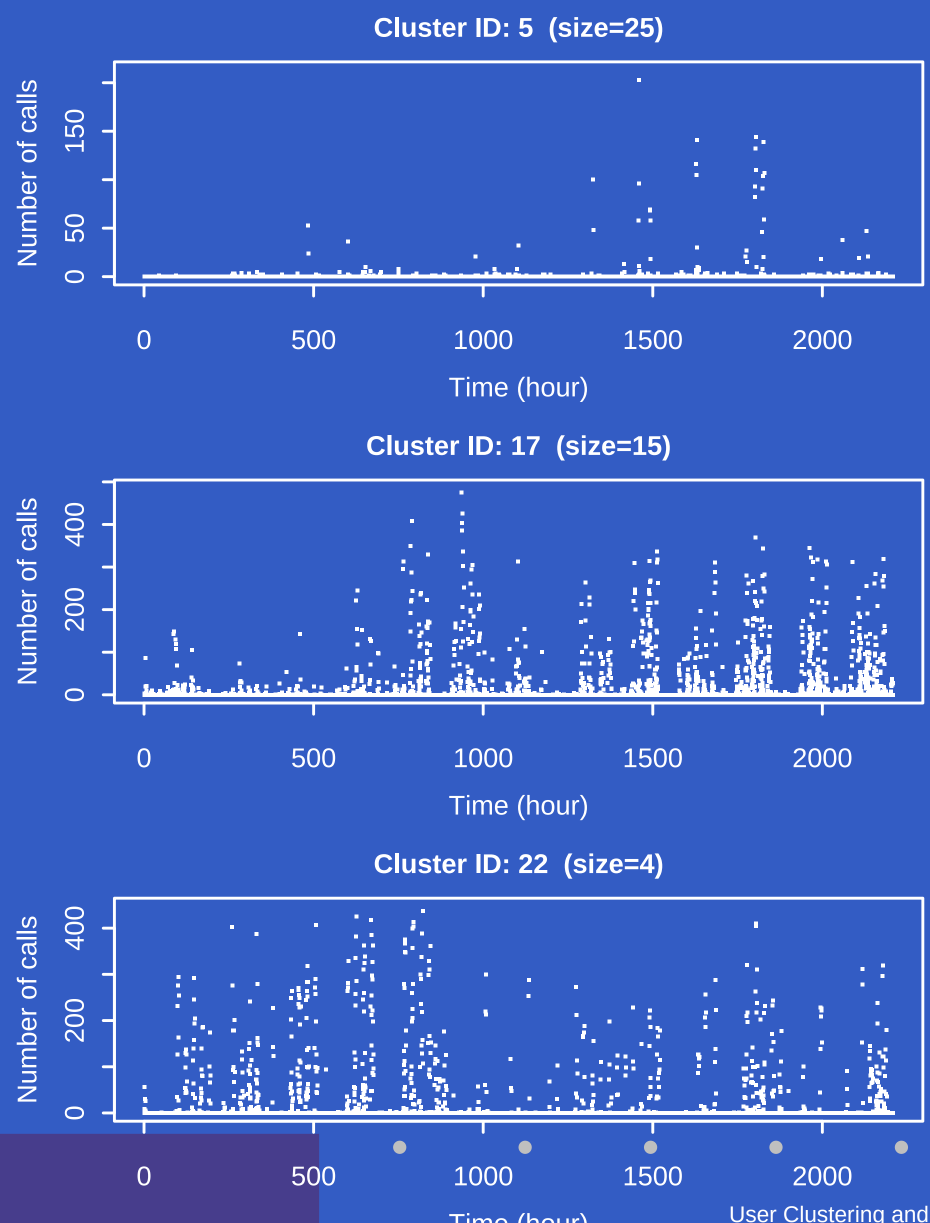
ALGORITHM: AutoClass

- An unsupervised classification tool based on the classical finite mixture model.
- Begins by creating a random classification and then manipulates it into a high probability classification through local changes.
- Repeats the process until it converges to a *local maximum*.
- Starts over again and continues for a specified number of tries.

ALGORITHM: K-means

- 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.
- Intra-cluster similarity is measured with respect to the mean value of the objects in a cluster.
- K-means is well-known for its simplicity and efficiency.
- Own implementation and $pam()$ function in R system.

AutoClass: CLUSTERS PLOT



K-means: CLUSTER RESULTS

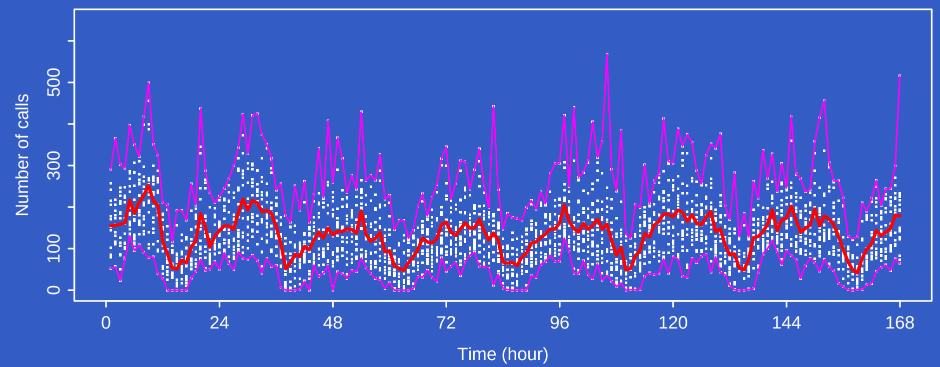
- We tested the performance of K-means for: K = 3, 6, and 16.
- The Euclidean distance was used as the distance function to measure the similarity among talk groups.
- Overall quality is defined as the minimum inter-cluster distance minus the maximum intra-cluster cluster distance.
- 3 is the best number of clusters, in terms of inter-cluster, intra-cluster distance, and overall quality.

K-means: CLUSTER RESULTS

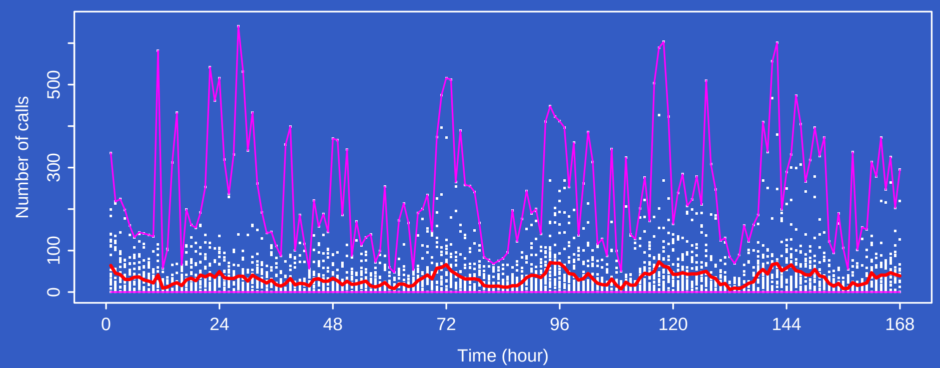
Num (K)	Sizes	Avg. intra	Avg. inter	Max. intra	Min. inter	Overall quality
3	17,31 569	1882.14	4508.38	2971.76	1626.4	-1345.36
6	13,17 22,3 34,528	2059.67	3284.52	3299.43	594.21	-2705.21
9	...	1020.08	3520.04	3065.25	808.28	-2256.96
12	...	1372.67	3582.98	3278.14	731.26	-2546.88
16	...	983.63	1815.79	3571.27	248.19	-3323.07

K-means CLUSTER PLOT

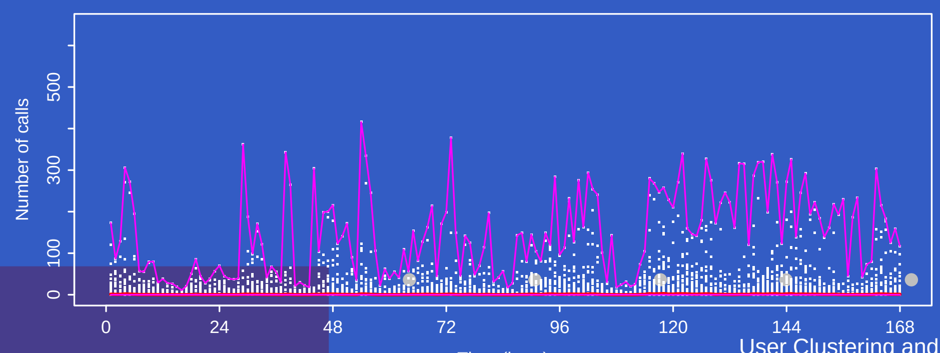
Cluster 1 (17 talk groups)



Cluster 2 (31 talk groups)



Cluster 3 (569 talk groups)



K-means CLUSTER PROPERTIES

Cluster size	Min. N.C.	Max. N.C.	Avg. N.C.	Total N.C.	Total N.C. (%)
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

(N.C.: Number of calls)

ROADMAP

- Introduction
- E-Comm network
- Traffic data
- User clustering
- Traffic prediction
- Conclusions
- Reference

ARIMA MODELS

- The Autoregressive Integrated Moving Average (ARIMA) models were developed by Box and Jenkins in 1976.
- ARIMA notation (ARIMA (p, d, q)).
- Autoregressive model: AR(p)
  $X_t \;=\; \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + Z_t$.
- Moving average model: MA(q)
  $X_t \;=\; Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$.
- Number of differencing. (D)

SARIMA MODELS

- Seasonal ARIMA: ARIMA plus seasonal period.
- A SARIMA $(p, d, q) \times (P, D, Q)_S$ model can be represented as:

$$\phi(B^s)\phi(B)(1 - B^s)^D(1 - B)^d X_t \;=\; \theta(B^s)\theta(B)Z_t,$$

  where $\phi(B)$ and $\theta(B)$ represent the AR and MA parts, $\phi(B^s)$ and $\theta(B^s)$ represent the seasonal AR and seasonal MA parts.

- B is the back-shift operator $(B^i X_t \;=\; X_{t-i})$.

ARIMA MODEL BUILDING

- Model identification
  - (p, d, q, P, D, Q, S)
- Model estimation
  - $\phi(x)$, $\theta(x)$
- Model verification
  - residual analysis

SARIMA MODELS & NMSE

- SARIMA models $(2,0,1) \times (0,1,1)_{24}$ and $(2,0,1) \times (0,1,1)_{168}$ were selected to predict the future n hours traffic data, based on m hours past traffic data.
- Normalized mean square error $nmse$ was used to measure the prediction quality:

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

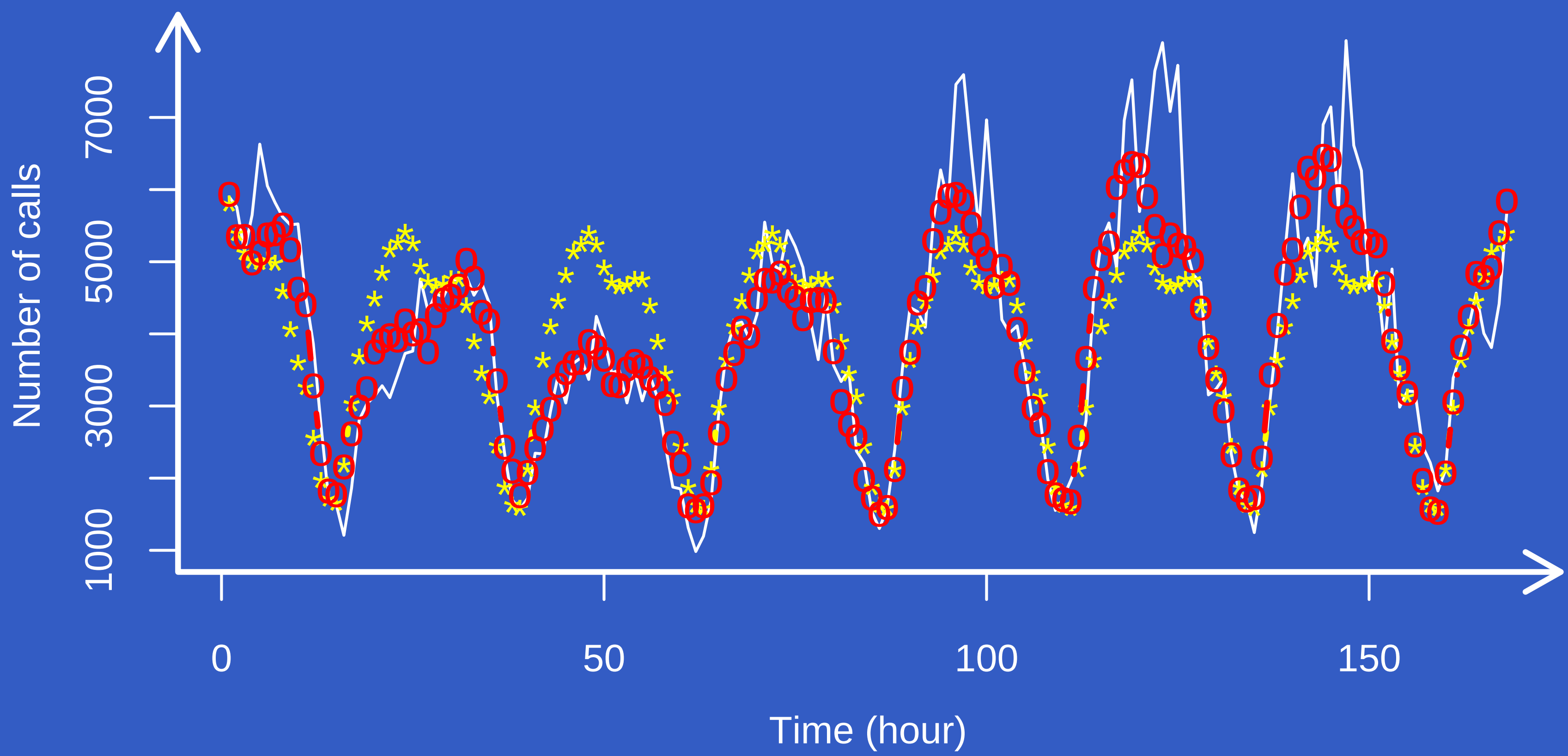
where a_i is the observed, b_i is the predicted data, and $\bar{a}$ is the mean value of a_i.

PREDICTION RESULTS

p	*d*	*q*	*P*	*D*	*Q*	*S*	*m*	*n*	*nmse*
2	0	1	0	1	1	24	1,920	24	0.1941
3	0	1	0	1	1	24	1,920	24	0.1907
2	0	1	0	1	1	24	1,680	168	0.4079
3	0	1	0	1	1	24	1,680	168	0.4081
2	0	1	0	1	1	168	1,920	24	0.0969
3	0	1	0	1	1	168	1,920	24	0.1012
2	0	1	0	1	1	168	1,680	168	0.1745
3	0	1	0	1	1	168	1,680	168	0.1748

PREDICTION VISUALIZATION

- Comparison of $(2,0,1) \times (0,1,1)_{24}$ to $(2,0,1) \times (0,1,1)_{168}$ (m:1680, n:168)



CLUSTER BASED PREDICTION

- Talk groups were partitioned into three clusters.
- SARIMA models $(2,0,1) \times (0,1,1)_{24}$ and $(2,0,1) \times (0,1,1)_{168}$ are applied for each cluster to predict the traffic.
- Predict of network traffic by aggregating the traffic predicted from three clusters of users.
- Optimize the prediction for “bad” cluster prediction.

PREDICTION RESULTS

no	(p,d,q)	(P,D,Q)	S	m	n	nmse
1	(2,0,1)	(0,1,1)	24	1680	48	1.1954
2	(2,0,1)	(0,1,1)	24	1680	48	2.4519
3	(2,0,1)	(0,1,1)	24	1680	48	0.3701
*	(2,0,1)	(0,1,1)	24	1680	48	0.6298
A	(2,0,1)	(0,1,1)	24	1680	48	0.6256
O	(2,0,1)	(0,1,1)	24	1680	48	0.4231

PREDICTION RESULTS (cont.)

no	(p,d,q)	(P,D,Q)	S	m	n	nmse
1	(2,0,1)	(0,1,1)	168	1,920	24	0.2241
2	(2,0,1)	(0,1,1)	168	1,920	24	0.3818
3	(2,0,1)	(0,1,1)	168	1,920	24	0.1163
*	(2,0,1)	(0,1,1)	168	1,920	24	0.0969
A	(2,0,1)	(0,1,1)	168	1,920	24	0.1175

PREDICTION SUMMARY

- $SARIMA(2,0,1) \times (0,1,1)_{24}$ model
  - 14% prediction based on cluster traffic beats prediction based on aggregate traffic.
  - 87% optimized prediction based on cluster traffic beats prediction based on aggregate traffic.
- $SARIMA(2,0,1) \times (0,1,1)_{168}$ model
  - 59% prediction based on cluster traffic beats prediction based on aggregate traffic.
  - None of the optimized prediction based on cluster traffic beats prediction based on aggregate traffic.

ROADMAP

- Introduction
- E-Comm network
- Traffic data
- User clustering
- Traffic prediction
- Conclusions
- Reference

CONCLUSIONS

- We analyzed traffic data collected from an operational trunked radio network.
- We used the K-means algorithm and AutoClass to classify network users into user clusters.
- We predicted network traffic using the SARIMA model based on aggregate user traffic and based on three user clusters.
- Some user cluster based prediction perform better than the aggregate traffic based prediction.

CONCLUSIONS - contributions

- Analyzed real-world data and problems.
- Applied clustering algorithms on real data.
- Proposed cluster based prediction method.
- Compared cluster based prediction with traditional prediction method.
- Paper published on International Symposium on Wireless Communication Systems 2004 (http://www.ieeevtc.org/iswcs04).

FUTURE WORK

- Test various clustering algorithms.
- Compare with other prediction models (HMM, FARIMA).
- Integrate with simulation tool (WarnSim).

ROADMAP

- Introduction
- E-Comm network
- Traffic data
- User clustering
- Traffic prediction
- Conclusions
- Reference

REFERENCE: TRAFFIC ANALYSIS

[1] D. Tang and M. Baker, “Analysis of a metropolitan-area wireless network,” in *Proc. MOBICOM 1999*, Seattle, WA, USA, Aug. 1999, pp. 13–23.

[2] L. A. Andriantiatsaholiniaina and Lj. Trajković, “Analysis of user behavior from billing records of a CDPD wireless network,” in *Proc. IEEE Workshop on Wireless Local Networks (WLN) 2002*, Tampa, FL, Nov. 2002, pp. 781–790.

[3] 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, Special Issue on Quality of Service Delivery in Variable Topology Networks*, vol. 22, no. 7, pp. 1197–1205, Sept. 2004.

REFERENCES: CLUSTERING

[4] 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.

[5] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

[6] J. W. Han and M. Kamber, *Data Mining: Concepts And Techniques*. San Francisco: Morgan Kaufmann Publishers, 2001.

REFERENCES: PREDICTION

[7] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1976.

[8] The R Project for Statistical Computing [Online]. Available: http://www.r-project.org.

[9] 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)*, vol. 3, New Orleans, LA, May 1994, pp. 1400–1404.

[10] Y. W. Chen, “Traffic behavior analysis and modeling sub-networks,” *International Journal of Network Management*, John Wiley & Sons, vol. 12, 2002, pp. 323–330.

THE END

THANKS !