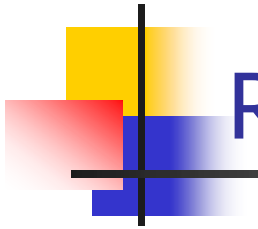
Prediction of Traffic in a Public Safety Network

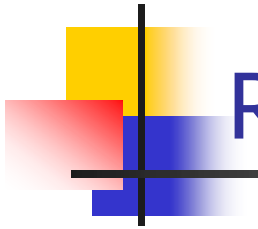
Božidar Vujičić, Hao Chen, and Ljiljana Trajković
{bvujicic, lcheu, ljilja}@cs.sfu.ca

Communication Networks Laboratory
http://www.ensc.sfu.ca/cnl
Simon Fraser University
Vancouver, BC



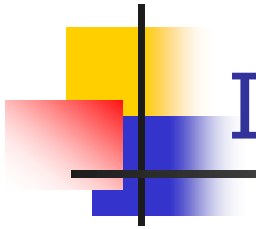
Roadmap

- Introduction
- E-Comm network and traffic data:
  - overview of the E-Comm network
  - data preprocessing and extraction
- Data clustering
- Traffic prediction based on:
  - aggregate traffic
  - user clusters
- Conclusion



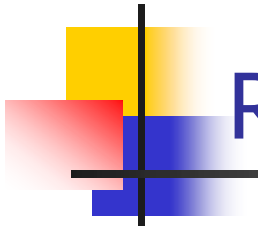
Roadmap

- Introduction
- E-Comm network and traffic data:
  - overview of the E-Comm network
  - data preprocessing and extraction
- Data clustering
- Traffic prediction:
  - based on aggregate traffic
  - cluster based
- Conclusion



Introduction

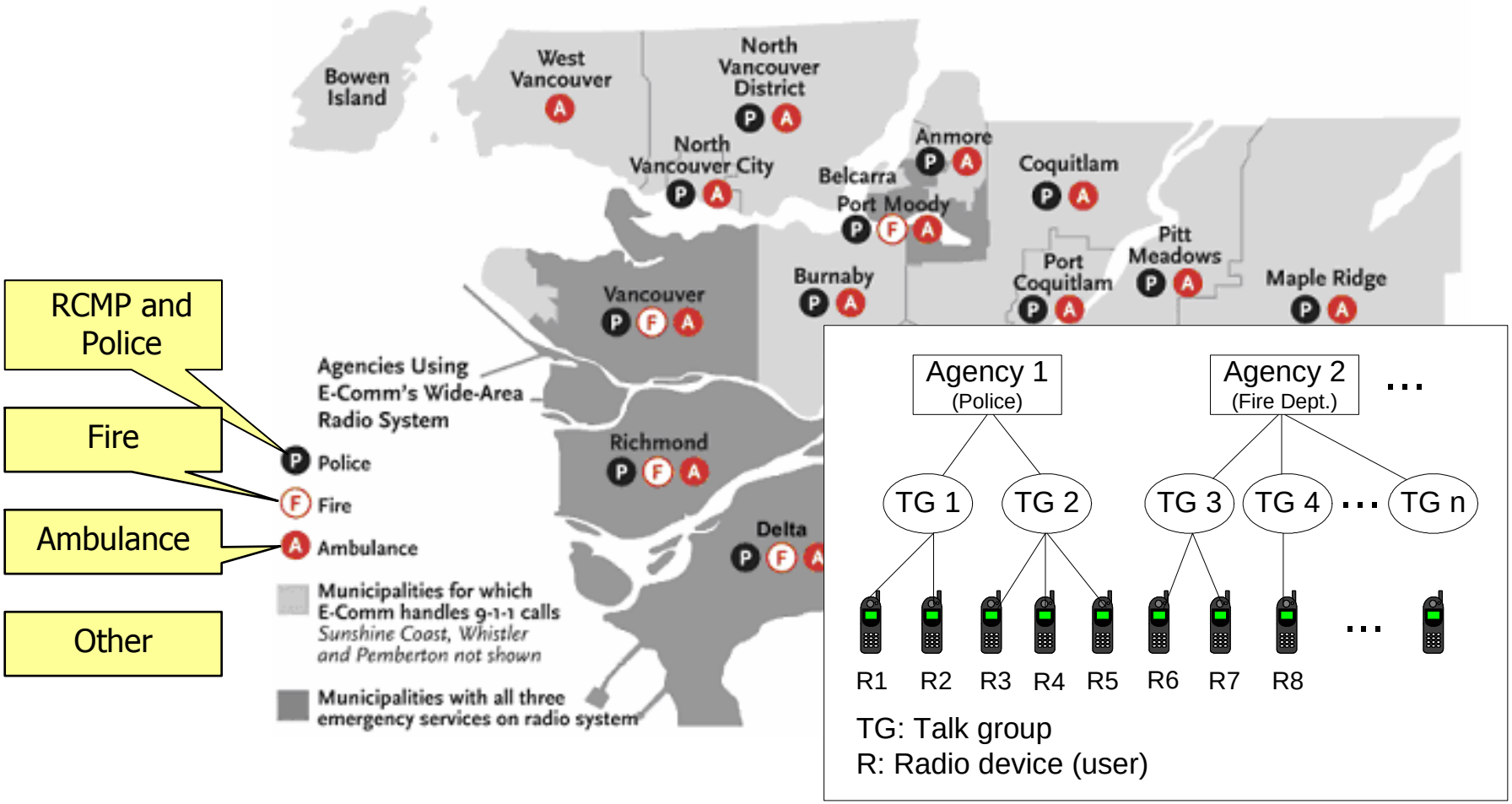
- Analysis of traffic data from operational networks provides insight into behavior of network users:
  - better utilization of network
  - better quality of service
- Data clustering:
  - identify traffic pattern
  - employ K-means algorithm
- Traffic prediction:
  - based on aggregate traffic
  - cluster based prediction



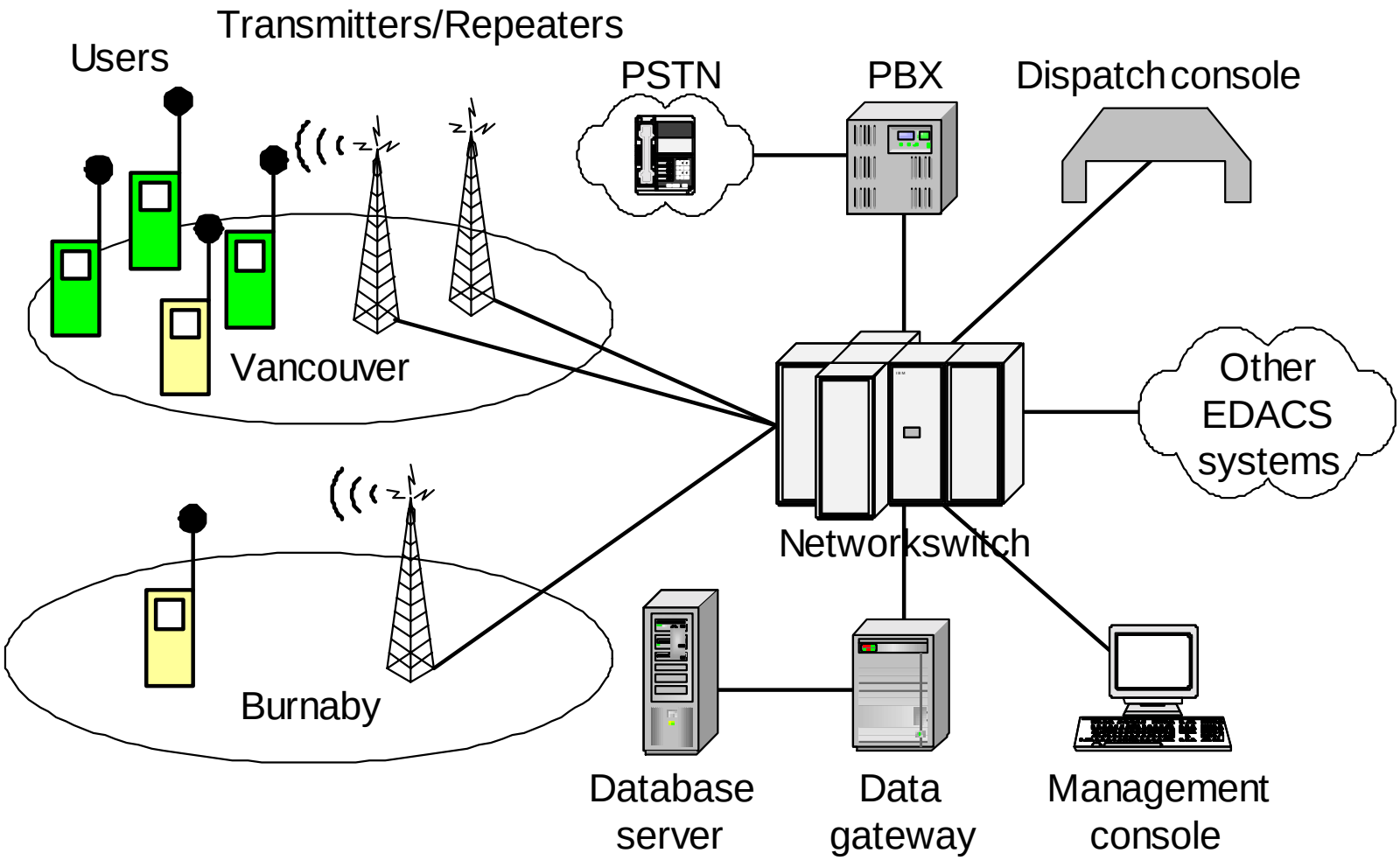
Roadmap

- Introduction
- E-Comm network and traffic data:
  - overview of the E-Comm network
  - data preprocessing and extraction
- Data clustering
- Traffic prediction:
  - based on aggregate traffic
  - cluster based
- Conclusion

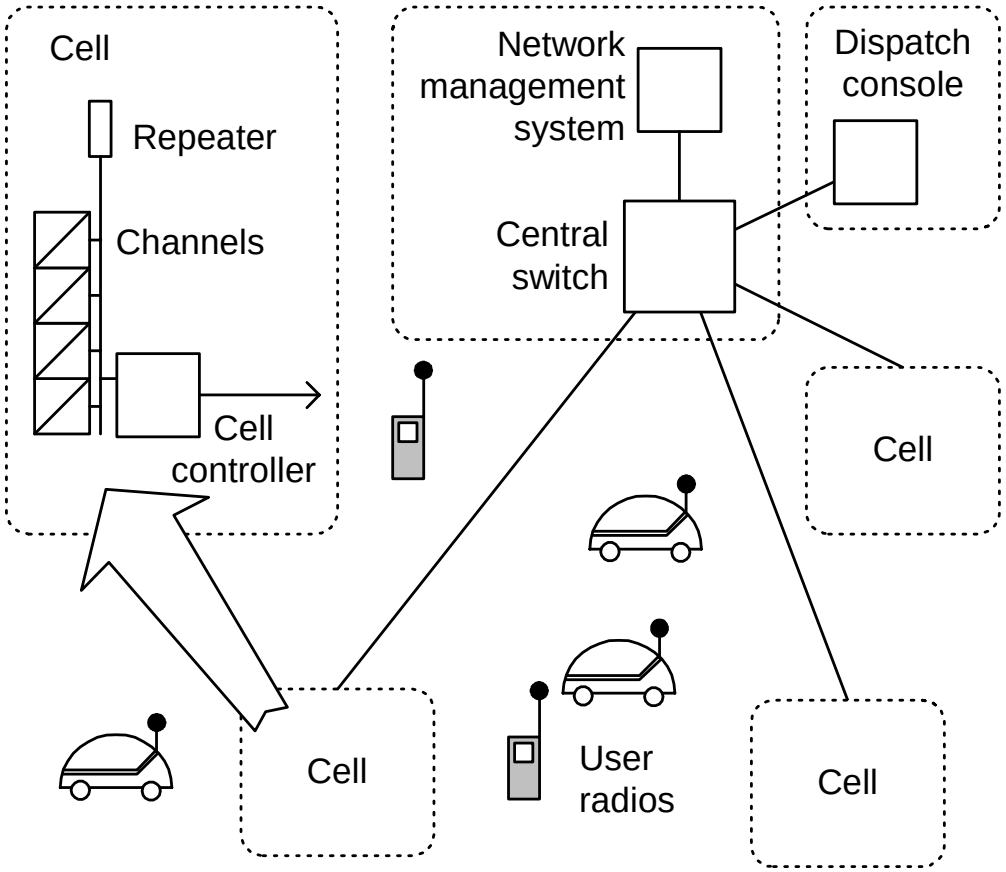
E-Comm network: coverage and user agencies

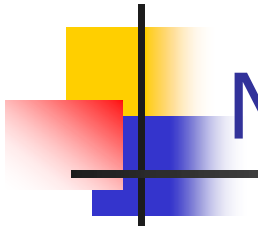


E-Comm network architecture



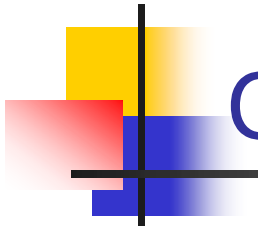
Structure of trunked radio systems





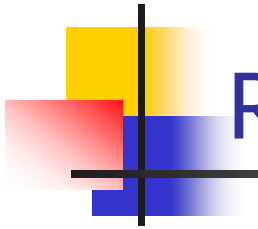
Network characteristics

- EDACS: Enhanced Digital Access Communications Systems
- Simulcast: repeaters covering one cell use identical frequencies
- Trunking: available frequencies in a cell are shared dynamically among mobile users
  - transmission trunking
  - message trunking
- Cell capacity (number of available frequencies in cell):
  - one radio channel occupies one frequency
  - one call occupies one radio channel



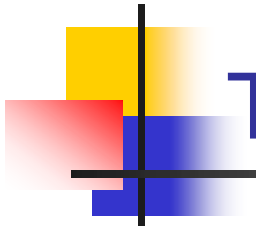
Call establishment

- 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



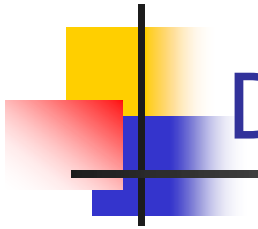
Roadmap

- Introduction
- E-Comm network and traffic data:
  - overview of the E-Comm network
  - data preprocessing and extraction
- Data clustering
- Traffic prediction:
  - based on aggregate traffic
  - cluster based
- Conclusion



Traffic data

- 2001 data set:
  - 2 days of traffic data
    - 2001-11-01 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)

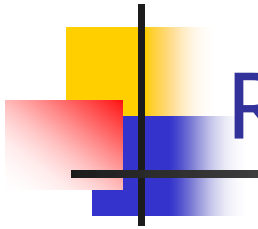


Data preprocessing: 2003 data

- Database size ~6 Gbytes, with 44,786,489 records:
  - 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
- Not all data fields are useful to our analysis
- Number of records was reduced to only 19% of original records after removing irrelevant data fields

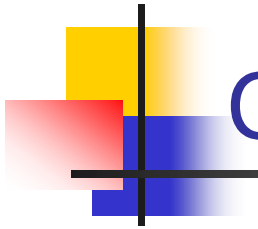
Sample of processed data: 2003-03-01

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



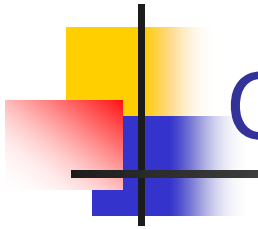
Roadmap

- Introduction
- E-Comm network and traffic data:
  - overview of the E-Comm network
  - data preprocessing and extraction
- Data clustering
- Traffic prediction:
  - based on aggregate traffic
  - cluster based prediction
- Conclusion



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



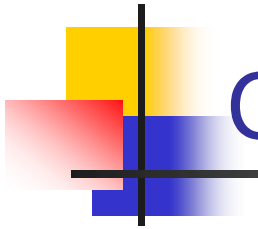
Clustering quality

- 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



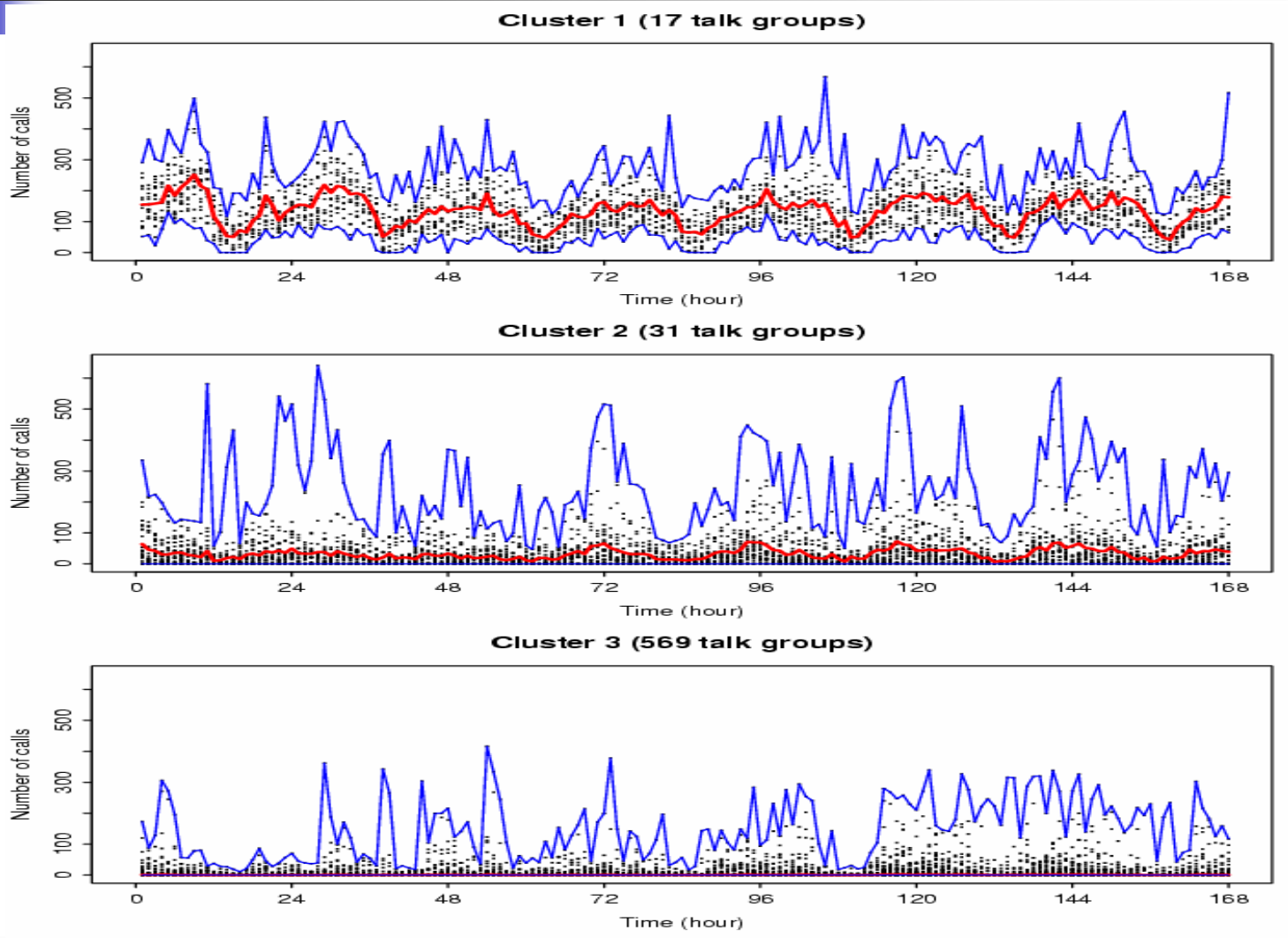
Clustering results

- 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

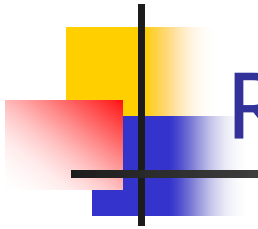
K	Average intra-cluster distance	Average inter-cluster distance	Maximum intra-cluster distance	Minimum inter-cluster distance	Overall clustering quality	Silhouette coefficient
3	1882.14	4508.38	2971.76	1626.40	-1345.36	0.7756
4	1863.00	3889.12	2971.76	1556.68	-1415.07	0.7684
6	2059.67	3284.52	3299.43	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



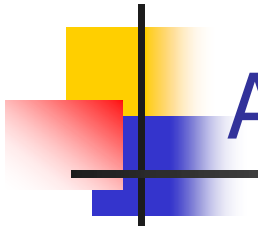
K-means clusters of talk groups

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



Roadmap

- Introduction
- E-Comm network and traffic data
  - overview of the E-Comm network
  - data preprocessing and extraction
- Data clustering
- Traffic prediction:
  - based on aggregate traffic
  - cluster based
- Conclusion



ARIMA models

- 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

SARIMA models

- Seasonal ARIMA (SARIMA) model:

$$(p,d,q)\times(P,D,Q)_S$$

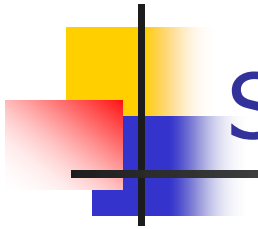
  - captures seasonal pattern
- SARIMA additional model parameters:
  - seasonal period parameter: S
  - seasonal autoregressive parameter: P
  - number of seasonal differencing passes: D
  - seasonal moving average parameter: Q

SARIMA models

- SARIMA $(p,d,q)\times(P,D,Q)_S$ model:

$$\phi(B^S)\phi(B)(1-B^S)^D(1-B)^d X_t = \theta(B^S)\theta(B)Z(t)$$

  - AR and MA parts: $\phi(B)$ and $\theta(B)$
  - seasonal AR and seasonal MA parts: $\phi(B^s)$ and $\theta(B^s)$
  - B = the back-shift operator: $B^i X_t = X_{t-i}$

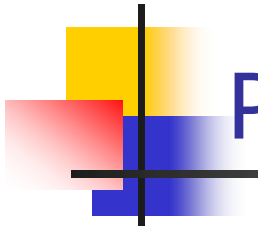


SARIMA models: selection criteria

- 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

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

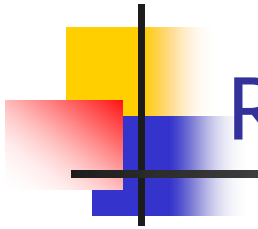
Prediction: based on the aggregate traffic

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) x $(0, 1, 1)_{24 \text{ and } 168}$
  - SARIMA (2, 0, 1) x $(0, 1, 1)_{24 \text{ and } 168}$
- Comparisons:
  - rows A1 with A2, B1 with B2, and C1 with C2
  - SARIMA (2, 0, 9) × $(0, 1, 1)_{24}$ gives better prediction results than SARIMA (2, 0, 1)×$(0, 1, 1)_{24}$
- 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



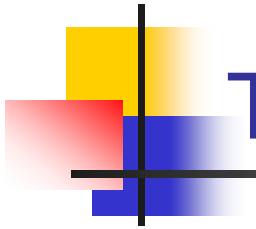
Roadmap

- Introduction
- E-Comm network and traffic data:
  - overview of the E-Comm network
  - data preprocessing and extraction
- Data clustering
- Traffic prediction:
  - based on aggregate traffic
  - cluster based
- Conclusion

Prediction based on user clusters model (2, 0, 1)×(0, 1, 1)

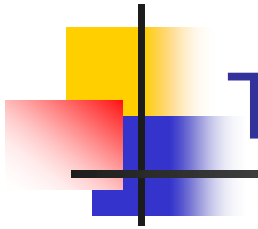
Test no.	S	m	n	nmse cluster 1	nmse cluster 2	nmse cluster 3	nmse aggregate	nmse cluster	nmse optimized
1	24	240	24	0.323	0.548	0.308	0.254	0.241	n/a
2	24	240	48	0.394	0.712	0.445	0.343	0.332	n/a
3	24	1200	72	1.774	1.976	0.270	0.884	0.886	0.846
4	24	1200	96	1.319	0.866	0.260	0.611	0.613	0.610
5	24	1200	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

Models forecast future n traffic data based on m past traffic data samples



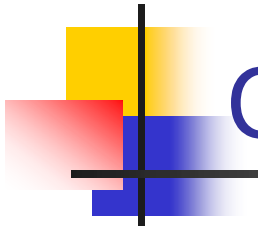
Traffic prediction with user clusters

- 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
  - Test 9: 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



Conclusion

- We analyzed network traffic data from an operational network
- By applying data mining techniques on traffic data, we discovered user clusters based on patterns of calling behavior expressed by hourly number of calls
- 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

References

- 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.
- N. K. Groschwitz and G. C. Polyzos, "A time series model of long-term NSFNET backbone traffic," in *Proc. ICC*, May 1994, vol. 3, pp. 1400–1404.
- 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.
- 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.
- 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. 14–223.
- N. Cackov, J. Song, B. Vujičić, S. Vujičić, and Lj. Trajković, "Simulation and performance evaluation of a public safety wireless network: case study," *Simulation*, vol. 81, no. 8, pp. 571-585, Aug. 2005.
- H. Chen and Lj. Trajković, "Trunked radio systems: traffic prediction based on user clusters," in *Proc. ISWCS 2004*, Mauritius, Sept. 2004, pp.76–80.

References

- E-Comm, Emergency Communications for SW British Columbia Incorporated. (2005, May). [Online]. Available: http://www.ecomm.bc.ca.
- EDACS Trunking Information. (2004, May). [Online]. Available: http://www.trunkedradio.net/ trunked/ edacs/ EDACS Whitepaper.pdf.

- L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: Wiley-Interscience, 1990.
- G. E. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day, 1976.
- N. H. Chan, *Time Series: Applications to Finance*. New York, NY: 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.

Erlang traffic models

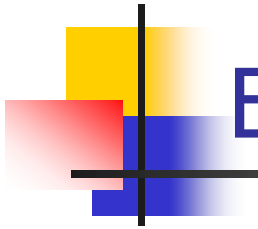
Erlang B

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

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



Erlang traffic 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