Prediction of Traffic in a Public Safety Network

Božidar Vujičić, Hao Chen, and Ljiljana Trajković
Simon Fraser University
Vancouver, British Columbia, Canada
{bvujicic, lcheu, ljilja}@cs.sfu.ca

Abstract— **Traditional statistical analysis and mining of network data are often employed to determine traffic distribution, to summarize a user's behavior patterns, or to predict future network traffic. We analyze three months of network log data from a deployed public safety trunked radio network. After data cleaning and traffic extraction, we apply the K-means algorithm and identify that three clusters of talk groups best reflect users' behavior patterns represented by the hourly number of calls. We propose a traffic prediction model by applying the classical SARIMA models on clusters of users. The predicted network traffic agrees with the collected traffic data and the proposed cluster-based prediction approach performs well compared to the prediction based on the aggregate traffic.**

I. Introduction

Analysis of traffic data from operational networks provides insight into the behavior of network users. It may lead to better utilization of network resources and better quality of services. Data clustering may be used to identify traffic patterns. Network users are usually classified into user groups according to their geographical location, organizational structure, payment plan, or behavior pattern.

Prior analysis of traffic data from a metropolitan-area wireless network indicated the recurring daily user behavior and mobility patterns [1]. Analysis of traffic from a trunked radio network showed that the call holding time distribution is approximately lognormal [2], while the call inter-arrival times are long-range dependent and could be modeled by both Weibull and gamma distributions [3]. Channel utilization and the multi-system call behavior of trunked radio network have also been analyzed using network simulation tools [4]. A preliminary study of traffic data collected from this radio network was reported in [5].

In this paper, we analyze traffic data collected from a deployed network. We predict network traffic based on the aggregate traffic and based on the clusters of users identified by the K-means algorithm. Experimental results show that the cluster-based prediction produces results comparable to the traditional prediction of network traffic. An advantage of cluster-based prediction is that it may be used for predictions in networks with variable number of users.

The network and the traffic data are introduced in Section II. In Section III, the K-means algorithm is applied to classify talk groups into clusters based on their calling behavior. The aggregate and cluster-based traffic prediction models are given in Section IV. The paper concludes with Section V.

II. E-Comm Network and Traffic Data

The traffic data were collected from the E-Comm network [6]. We examine the database schema and describe the procedure for data cleaning and the traffic data extraction.

A. E-Comm Network Overview

E-Comm is the regional emergency communications center providing emergency dispatch/communication services for a number of agencies in the Greater Vancouver Regional District (GVRD) in Southwest BC, Canada. The E-Comm network is currently used by sixteen agencies, such as police, fire and rescue, and ambulance. Each agency has a number of affiliated talk groups and the entire network serves 617 talk groups.

The E-Comm network is a trunked radio system, employing the Enhanced Digital Access Communications System (EDACS [7]) technology. The E-Comm network architecture consists of 11 cells. Each cell covers one or more municipalities. The basic talking unit in the trunked radio network is a talk group of individual users who need to communicate frequently.

A Group Call is the typical call made in a trunked radio system. A user places a group call by pressing the push-to-talk (PTT) button on the radio device. All users belonging to the same talk group hear the communications in a group call irrespective of their physical locations.

A *Multi-System Call* represents a single group call involving more than one system/cell. If all members of the talk group reside within one system, the group call is a single-system call occupying only one channel in the system. However, when group members are distributed over multiple systems, the group call becomes a multi-system call that occupies one traffic channel in each system.

B. Data Preprocessing and Extraction

The E-Comm database contains event log tables recording the network activities. They are aggregated from the distributed database of the individual network management systems. Data records span from 2003-03-01 00:00:00 to 2003-05-31 23:59:59. They are sorted in 92 event log tables, each containing one day's events. The size of the database is ∼6 Gbytes, with 44,786,489 records.

Not all data fields are useful to our analysis. Certain fields are not populated in the database, while others have

This research was supported by the NSERC Grant 216844-03 and Canada Foundation for Innovation New Opportunities program.

TABLE I
A SAMPLE OF CLEANED DATA. DATE: 2003-03-01, CALL TYPE = 0, CALL STATE = 0, MULTI-SYSTEM CALL = 0.

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

identical values. From the 26 fields in the database, 9 fields that capture the user's behavior and network traffic are of particular interest to our study: Event_UTC_At, Duration$ms, SystemId, ChannelId, Caller, Callee, CallType, CallState, and MultiSystemCall. A sample of the pre-processed traffic data is shown in Table I. After reducing the database dimension to 9, we removed the redundant records. The records with call_state = 1, which implies the *call drop* event, are redundant because each *call drop* event already has a corresponding *call assignment* event in the database. (Note that the reverse is not true.) We also removed records for the control channel whose traffic data were not available.

If a call is a multi-system call involving several systems, one record for each involved system is created to represent this call in the original event log database. As shown in Table I, based on the caller, callee, and call duration, records 1 and 6 represent one group call from caller 13905 to callee 401, involving systems 1 and 7 and lasting ∼1350 ms. Records 29, 31, 37, and 38 represent a group call from caller 13233 to callee 249, involving systems 2, 1, 7, and 6.

The call duration is sometimes inconsistent because of the transmission latency and glitches in the distributed database system. For example, records 1 (1340 ms) and 6 (1350 ms) in Table I, have 10 ms difference in call duration field although they represent one single group call. We used 50 ms difference in call duration as an empirical choice when combining the multiple records. The result of the data preprocessing is a database with ∼55% fewer records. After the traffic extraction, the number of records in the database was reduced to only 19% of the original records.

III. DATA CLUSTERING

Clustering analysis groups or segments a collection of objects into subsets or clusters so that the resulting intra-cluster similarity is high while the inter-cluster similarity is low. An object can be described by a set of measurements or by its relations to other objects. Network users' behavior may be characterized by the time of the calls, the average call duration, or the number of calls during a certain time interval.

A commonly used metric in the telecommunication industry is the hourly number of calls. It may be regarded as the footprint of a user's calling behavior. Since the talk group is the basic talking unit in the E-Comm network, we use a talk group's hourly number of calls to represent a user's behavior. The collected 92 days of traffic data (2,208 hours) imply that each talk group's calling behavior may be portrayed by the 2,208 ordered hourly numbers of calls.

A general approach to clustering is to view it as a density estimation problem. We assume that data are generated from a mixture model where the probability at each data point is the sum of a mixture of several distributions. We begin by choosing K seeds (means of distributions) and iterate over the estimation and the maximization steps. Distances of each data point from the K seeds are first calculated (estimation step). The mean of each distribution is then moved towards the centroid of the entire data set, weighted by the number of data points in the cluster (maximization step). These steps are repeated until the distributions no longer move. At the end of the process, each point is tied to a certain cluster with the highest probability. In a mixture model M with K clusters $C_i, i = 1, \cdots, K$, the probability of a data point x belonging to the model is:

$$P(x|M) = \sum_{i=1}^{K} W_i * P(x|C_i, M),$$

where W_i is the mixture weight.

One of the most commonly used data clustering algorithms is K-means [8]. The number of clusters K (known a priori) and the object similarity function are two input parameters.

We use the inter-cluster and the intra-cluster distances to assess the overall clustering quality. The inter-cluster distance reflects the dissimilarity between clusters. It is defined as the Euclidean distance between two cluster centroids (the mean value of the objects in a cluster, which can be viewed as the cluster's center of gravity). The intra-cluster distance is the average distance of objects from their cluster centroids, expressing the coherent similarity of data in the same cluster. A large inter-cluster distance and a small intra-cluster distance indicate better clustering. The overall clustering quality indicator is defined as the difference between the minimum inter-cluster and the maximum intra-cluster distances. The greater the indicator, the better the overall clustering quality. Another measure for the clustering quality is silhouette coefficient [8], which is rather independent of the number of clusters K. If $a(x)$ and $b(x)$ are average distances between data point x and other data points in clusters A and B, respectively, then:

$$\text{silhouette coefficient}(x) = (b(x) - a(x))/max\{a(x), b(x)\}.$$

Experience shows that larger values of silhouette coefficient produce better results. Values between 0.7 and 1.0 indicate clustering with excellent separation between clusters.

The inter-cluster and the intra-cluster distances, the overall quality, and silhouette coefficients for various number of clusters K are shown in Table II. Cluster sizes are: 17, 31, and 569 for $K = 3$; 17, 33, 4, and 563 for $K = 4$; and 13, 17, 22, 3, 34, and 528 for $K = 6$. Based on the overall quality and the silhouette coefficient, K = 3 produces the best clustering results. One week of traffic data for talk groups in each cluster and their distinct calling behavior are shown Fig. 1.

The properties of the three K-means clusters are given in Table III. The first cluster has 17 talk groups, representing the busiest dispatch groups whose main tasks are coordinating

TABLE II

K-MEANS CLUSTERING: CLUSTER DISTANCES.

K	Avg. intra dist.	Avg. inter dist.	Max. intra dist.	Min. inter dist.	Overall clustering quality	Silhouette coeff.
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

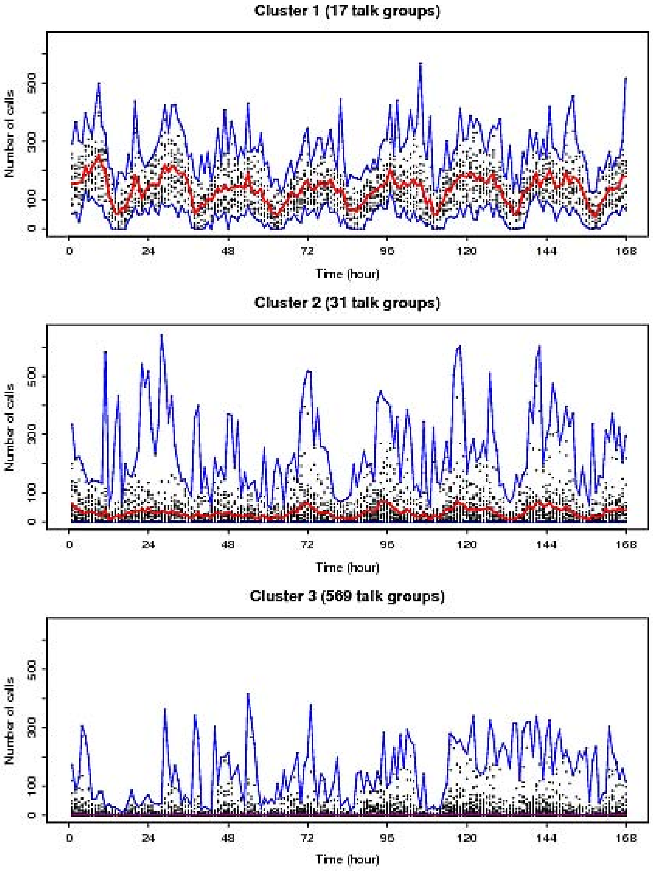

Fig. 1. K-means clustering: number of calls in the three clusters.

and scheduling other talk groups for certain tasks. The second cluster contains 31 talk groups with medium network usage. The last cluster identifies a group of the least frequent network users who made on average no more than 16 calls per hour. These interpretations of clusters have been confirmed by the E-Comm domain experts. Therefore, in the prediction of traffic, we use the three clusters identified by K-means.

IV. TRAFFIC PREDICTION

We compare predictions of network traffic based on aggregate traffic and based on user clusters.

A. Traffic Prediction Based on Aggregate Traffic

The Auto-Regressive Integrated Moving Average (ARIMA) model [9] is a general model for forecasting a time series. A seasonal ARIMA (SARIMA) [9] $(p, d, q) \times (P, D, Q)_S$ model captures the seasonal pattern. Parameters (p, P), (d, D), and (q, Q) represent the order of the Auto-Regressive (AR), difference, and Moving Average (MA) model for the original data points and the seasonal pattern, respectively. Parameter S is the seasonal period of the models. SARIMA models have been applied to modeling and predicting traffic from a large scale network [10] and a small scale subnetwork [11]. They may 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, respectively. B is the back-shift operator: $B^i X_t \;=\; X_{t-i}$.

TABLE III

K-MEANS CLUSTERS OF TALK GROUP (NC: NUMBER OF CALLS).

Cluster size	Min. nc	Max. nc	Avg. nc	Total nc	Total nc (%)
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

TABLE IV

SUMMARY OF SELECTION CRITERIA FOR SARIMA MODELS.

$(p, d, q) \times (P, D, Q)_S$	m	*nmse*	AIC	AIC_C	BIC
$(2, 0, 9) \times (0, 1, 1)_{24}$	1680	0.379	22744.6	22744.9	22826.8
$(2, 0, 1) \times (0, 1, 1)_{168}$	1680	0.174	23129.8	23129.8	23161.9
$(1, 0, 1) \times (0, 1, 1)_{168}$	1680	0.175	23145.1	23145.1	23170.8
$(2, 0, 9) \times (1, 1, 1)_{24}$	1680	0.525	25292.1	25292.4	25382.1
$(1, 0, 2) \times (1, 1, 1)_{24}$	1680	0.411	25332.6	25332.6	25371.2
$(2, 0, 1) \times (0, 1, 1)_{24}$	1680	0.408	25360.5	25360.6	25392.6
$(3, 0, 1) \times (0, 1, 1)_{24}$	1680	0.404	25361.2	25361.2	25399.7

The E-Comm network traffic possesses both daily and weekly cyclic patterns. Both 24-hour and 168-hour (one week) intervals are selected as seasonal period parameters. The order of the SARIMA models is selected based on the time series plot of traffic data and the autocorrelation and partial autocorrelation functions. In order to check validity of the parameter selection for SARIMA models, we employed the Akaike's information criterion AIC, the Akaike's information criterion corrected $AICC$ [12], and the Bayesian information criterion BIC [13]. Based on the 1,680 training data, models $(2, 0, 9) \times (0, 1, 1)_{24}$ and $(2, 0, 1) \times (0, 1, 1)_{168}$ have the smallest criterion values. Hence, they are selected as model candidates. The order $(0, 1, 1)$ is commonly used for the seasonal part (P, D, Q) because the cyclical seasonal pattern is usually a random-walk and may be modeled as an MA process after one-time differencing. The model's goodness-of-fit is validated using the null hypothesis test that includes time plot analysis and the autocorrelation function of the model residual. The summary of parameter selection criteria is shown in Table IV.

Four models with parameters fitted for the E-Comm network traffic and the aggregate traffic prediction results are shown in Table V. The model performance is tested for several groups of data (A, B, C). We forecast future n traffic data based on m past traffic data samples. Normalized mean square error ($nmse$) is used to measure prediction quality by comparing the deviation between predicted and observed data. The $nmse$ of the forecast is equal to the ratio of the normalized sum of

TABLE V

PREDICTION RESULTS 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
B3	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

the variance of the forecast to the squared bias of the forecast. Smaller values of $nmse$ indicate better prediction model.

Comparisons of rows A1 with A2, B1 with B2, and C1 with C2, indicate that Model 1 ($(2,0,9) \times (0,1,1)_{24}$) gives better prediction results than Model 2 ($(2,0,1) \times (0,1,1)_{24}$). Furthermore, for all three groups of training data, Model 3 ($(2,0,9) \times (0,1,1)_{168}$) and Model 4 ($(2,0,1) \times (0,1,1)_{168}$) with the 168-hour period always lead to better prediction than Model 1 and Model 2 with the 24-hour period. The 24-hour period models assume that the traffic is relatively constant for a weekday, while the 168-hour period models take into account traffic variations during a week. To predict traffic on a Wednesday based on Tuesday's data is not as accurate as predicting Wednesday's traffic based on the data of previous Wednesdays. However, the computational cost of identifying and forecasting 168-hour period models is often over 100 × CPU utilization required for the 24-hour period models.

B. Cluster Based Traffic Prediction

A key assumption of the prediction based on the aggregate traffic is the constant number of network users and steady behavior patterns. However, this assumption does not hold in the case of network expansions. Hence, we propose here a cluster-based approach to predict the overall network traffic by aggregating traffic predicted for individual clusters.

Comparison of prediction based on three clusters and prediction based on aggregate traffic is shown in Table VI. For each of the three clusters of talk groups, we employed SARIMA models $(2,0,1) \times (0,1,1)_{24}$ and $(2,0,1) \times (0,1,1)_{168}$ to predict traffic based on various number of training data. Predictions for the three clusters are then combined to predict the overall network traffic. Note that the $nmse > 1.0$ for clusters 1 (tests 3, 4, and 9) and for cluster 2 (test 3) implies that the prediction results are worse than prediction based on the mean value of the past data. If the mean value prediction is adopted for clusters 1 and 2 in Test 3, and cluster 1 in Test 4, we obtain better prediction results shown in the column "$nmse$ optimized" (optimized cluster-based prediction). (The non-optimized cluster-based prediction performs worse than the aggregate-traffic-based prediction.) Test 1, 2, 7, 8, 10, and 11 show that prediction based on clusters performs better than the prediction based on aggregate traffic. In our tests, 57% of the cluster-based predictions perform better than the aggregate-traffic-based prediction with SARIMA model $(2,0,1) \times (0,1,1)_{168}$.

TABLE VI

SUMMARY OF THE PREDICTION BASED ON USER CLUSTERS. MODEL $(2,0,1) \times (0,1,1)$. CLUSTERS ARE LABELED CL.1, CL.2, AND CL.3.

	S m n	$nmse$			$nmse$	$nmse$	$nmse$
#		**cl.1**	**cl.2**	**cl.3**	**aggr.**	**cl.**	**opt.**
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

Additional advantage of the cluster-based prediction is the ability to predict network traffic with variable number of users as long as the new user groups could be placed into the existing user clusters. The computational cost of forecasting the network traffic is reduced to the number of clusters times the prediction cost for one cluster.

V. CONCLUSIONS

We analyzed network traffic data from an operational network. By applying the data mining techniques on the traffic data, we discovered user clusters based on the patterns of calling behavior expressed by the hourly number of calls. The proposed cluster-based prediction produces comparable results to prediction based on the aggregate traffic. It is applicable to networks with variable number of users where the prediction based on aggregate traffic could not be applied.

REFERENCES

[1] 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.

[2] D. Sharp, N. Cackov, N. Lasković, Q. Shao, and Lj. Trajković, "Analysis of public safety traffic on trunked land mobile radio systems," *JSAC*, vol. 22, no. 7, pp. 1197–1205, Sept. 2004.

[3] 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.

[4] 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.

[5] H. Chen and Lj. Trajković, "Trunked radio systems: traffic prediction based on user clusters," in *Proc. ISWCS 2004*, Mauritius, Sept. 2004, pp. 76–80.

[6] E-Comm: Emergency Communications for SW British Columbia [Online]. Available: http://www.ecomm.bc.ca.

[7] EDACS Explained [Online]. Available: http://www.trunkedradio.net/trunked/edacs/EDACS_Whitepaper.pdf.

[8] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: Wiley-Interscience, 1990.

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

[10] 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.

[11] N. H. Chan, *Time Series: Applications to Finance*. New York, NY: Wiley-Interscience, 2002.

[12] K. Burnham and D. Anderson, *Model Selection and Multimodel Inference*, 2nd ed. New York, NY: Springer-Verlag, 2002.

[13] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978.