Anomaly Detection in the WIPER System Using Markov Modulated Poisson Process

Ping Yan Timothy Schoenharl Alec Pawling Greg Madey

Department of Computer Science and Engineering

University of Notre Dame

Notre Dame, IN 46556

pyan, tschoenh, apawling, gmadey at cse.nd.edu

Abstract

Cell phone call activity records the behavior of individuals, which reflects underlying human activity over time. Therefore this data displays multi-level periodicity, such as weekly, daily, hourly, etc. Simple stochastic models that rely on aggregate statistics are not able to differentiate between normal daily variations and legitimate anomalous (and potentially crisis) events. In this paper we describe a framework for unsupervised learning using a Markov modulated Poisson process (MMPP) [15, 6, 13] to model the data and use the posterior distribution to calculate the probability of the existence of anomalies over time. This paper focuses on anomaly detection in the WIPER system. WIPER is a system that helps to detect possible emergencies from cell phone data, provides the corresponding information to emergency planners and responders, and suggests possible actions to mitigate the emergency. One of the most important components of the system is the Detection and Alert System (DAS), which detects anomalies (which could be crisis events) from a cell phone data stream.

1 Introduction

A time series is a sequence of observations that can be measured over consecutive time periods at (often uniform) time intervals [16]. Time series data arise in a variety of domains, especially in economic systems, such as stock and financial data. Time series data also is used in environmental, medical, and

telecommunication data. Patterns of human behavior over time can be observed as a time series if the observed data reflects human behavior and can be measured from a data collection system. The crucial characteristic of a time series is that the data are not independent, their distribution varies over time, and usually displays an underlying trend. Analyzing and understanding the trend of human behavior over time has become an interesting research area [6, 15, 14, 12]. The goal of this paper is to analyze time series data from cell phone call activity, discover underlying human behavior, and use the results to detect potentially anomalous events.

First we want to distinguish outlier detection from anomaly detection in a time series scenario. The classic definition of an outlier is:

> ... an observation that deviates so much from other observations as to arouse suspicion that it was generated from a different mechanism. [5]

However, a single bursty point is not what we are looking for, what interests us most is anomalies, which can be called "special" patterns. Keogh $et\ al.$ define a surprising pattern as:

> ... if the frequency of its occurrence differs substantially from that expected by chance, given some previously seen data. [7]

From previously observed data, a normal behavior pattern is generated. If bursty activity happens within a certain time from the previously observed data, it is classified into the normal behavior pattern. Only patterns which greatly differ from expected data are considered as anomalous. By using this definition, no explicit description of an anomaly is required and normal patterns are inferred from a collection of previously observed data.

In this paper, we use the definition of anomaly from Keogh et al. [7]. We analyze the cell phone network data from two cities within a one-month time period. We collect the call activities for each transaction, which includes the account number, calling time, and location. From this detailed information, the call activities of individuals can be measured over a given time interval. Since this type of measurement contains the aggregated behavior of large amount of individuals, it typically demonstrates a periodicity of human behaviors on many time scales, such as hourly, daily, and weekly.

2 Related Work

Anomaly detection or event detection in time series data has received wide attention in the data-mining community. The process of finding interesting or surprising patterns from a time series dataset has been

studied by several researchers [4, 7, 8, 6, 15, 14, 2, 1].

Guralnik and Srivastava [4] present an iterative algorithm for detecting a change-point from time series data, and uses a maximum likelihood estimation to further segment the time slice if no change-point is observed. The approach does not require *a priori* knowledge of the data, and is independent of regression and model selection method.

Keogh et al. [7] used a suffix tree to encode the frequency of observed patterns, and applied a Markov model to detect surprising patterns. No explicit definition of surprising patterns are required, they are generated from previous observed data.

Kleinberg [8] used an infinite-state automaton to model data stream, and use it to detect the underlying content in a document stream. The bursts are identified as state transition. The author applied the method to analyze e-mail and research archives, and shows the efficiency of the proposed algorithm.

Scott et al. [15, 14] and Ihler et al. [6] used Markov-modulated Poisson processes to analyze human behaviors in web surfing[15], telephone network [14] and freeway traffic [6], and deployed the model to detect anomalies in their systems. Our model is derived from their work, and is used to analyze human behavior on a regular cell phone call-based scenario.

Basu and Meckesheimer [2] analyze the sensor data from airplane to detect the anomalies. Median value from a neighborhood of a data point is used to compare the difference to the observed data value. The proposed methods are fast and have quick response to data stream.

Agarwal [1] applied an empirical Bayes method on daily logs of large scale spoken dialog systems. The method fit a two-component Gaussian mixture to deviations of present time, which can avoid false positive by suppressing the consequence merely caused by sharp changes in the marginal distribution.

3 WIPER System And Data Characteristics

The Wireless Phone-based Emergency Response (WIPER) system is designed to provide emergency planners and responders with an integrated system that will help to detect possible emergencies, as well as to suggest and evaluate possible courses of action to deal with the emergency [11, 10]. Components of the system for detecting and mitigating emergency situations can be added and removed from the system as the need arises. WIPER is designed to evaluate potential plans of action using a series of GIS-enabled Agent-Based Simulations that are grounded on real time data from cell phone network providers. The system relies on the DDDAS concept [2], the interactive use of partial aggregate and detailed real time data to continuously update the system, which ensures that simulations always present timely and pertinent

data. WIPER presents information to users through a web-based interface of several overlaid layers of information, allowing users rich detail and flexibility.

One of the most important components of the WIPER system is the anomaly detection. In order to analyze and understand the human behavior represented by the cell phone call records, we selected two cities, referred to here as A and B, as our test models. City A is a smaller city with population around 20,000, and the city has 4 cell towers. City B is a bigger city with population around 200,000, and has 31 cell towers. In our collection of call activities, data is aggregated at tower level.

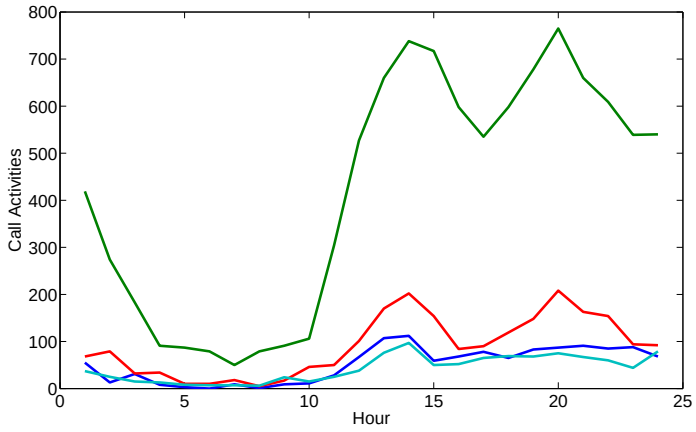
Figure 1 shows the similar behavior of call activities on different towers in the two cites on the same day. Some towers carry more calls than the others, and that is mainly because of location variations. Some towers are located in areas with higher population density, and some are located in a more sparse regions. Our following analysis always uses data from all towers in the region to represent the human behavior of the city.

Figure 2 shows call activity for city A over two weeks, with a total of 273,022 calls. Figure 3 shows call activity for city B over two weeks, with a total of 2,167,693 calls. Figures 2 and 3 show two weeks of calling activity with each day plotted separately and with the same days of the week, such as Monday, Tuesday, ... etc, shown in the same color. The figures demonstrate that there are similar behaviors on Day 1, Day 8 and Day 15 which are all Sundays in the dataset. Also Day 2 and Day 9 (Saturdays) share similar behavior, and all the remaining weekdays have similar behavior. In addition, for a particular day, call activities are always low between midnight and 8 am, and keep growing from 8am to 1-2 pm, but reduce around 3-4 pm in the afternoon, peak around 8-9 pm and then drop till midnight. The trend seems universal over all of the days, despite different days of week. These observed effects (day of week and hour of day effects) motivated our further investigation.

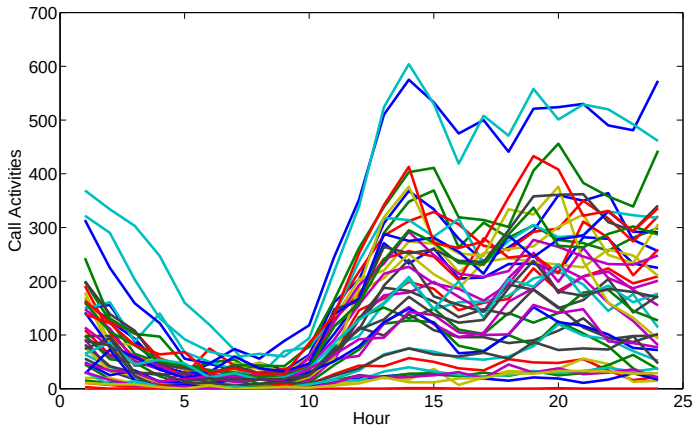
The data set consists of one month of cell phone activity. Each record contains calling account, calling time and location. There is no known emergency event in this time period in these two cities. The dataset only gives the basic call activities for each call, such as the begin/end time of the call, call id and call location.

4 MMPP Modeling

The model applied in this paper is derived from the Markov-Modulated Poisson Processes used by Ihler al. etc. for freeway traffic analysis[6], Scott and Smyth for web surfing behavior analysis [15] and Scott for telephone network intrusion detection [14].



(a) Small City



(b) Big City

Figure 1. Call Activities On Different Towers In The Two Cities On The Same Day

The MMPP is a special case of the doubly stochastic Poisson process [3, 9]. Its rate parameter obeys a Markov process. The Poisson distribution is widely used as a probabilistic model for count data, while the rate of the Poisson process represents the average number of observations in a fixed time period. In the MMPP model, a Markov process is employed to model the rate of the Poisson process, which indicates that the average number of occurrences follows a continuous time Markov chain.

In our application, let $N(t)$ refer to the count number of observed call activities at time t over a fixed time interval, where $t \in 1, ..., T$. In order to model $N(t)$, we need to model both normal behavior and abnormal behavior. The normal behavior represents the regular life of individuals, while abnormal behavior corresponds to rarely occurring events indicated by a change in call activity. At this stage, we did not classify the types of different anomalies: all the anomalous events are considered as one

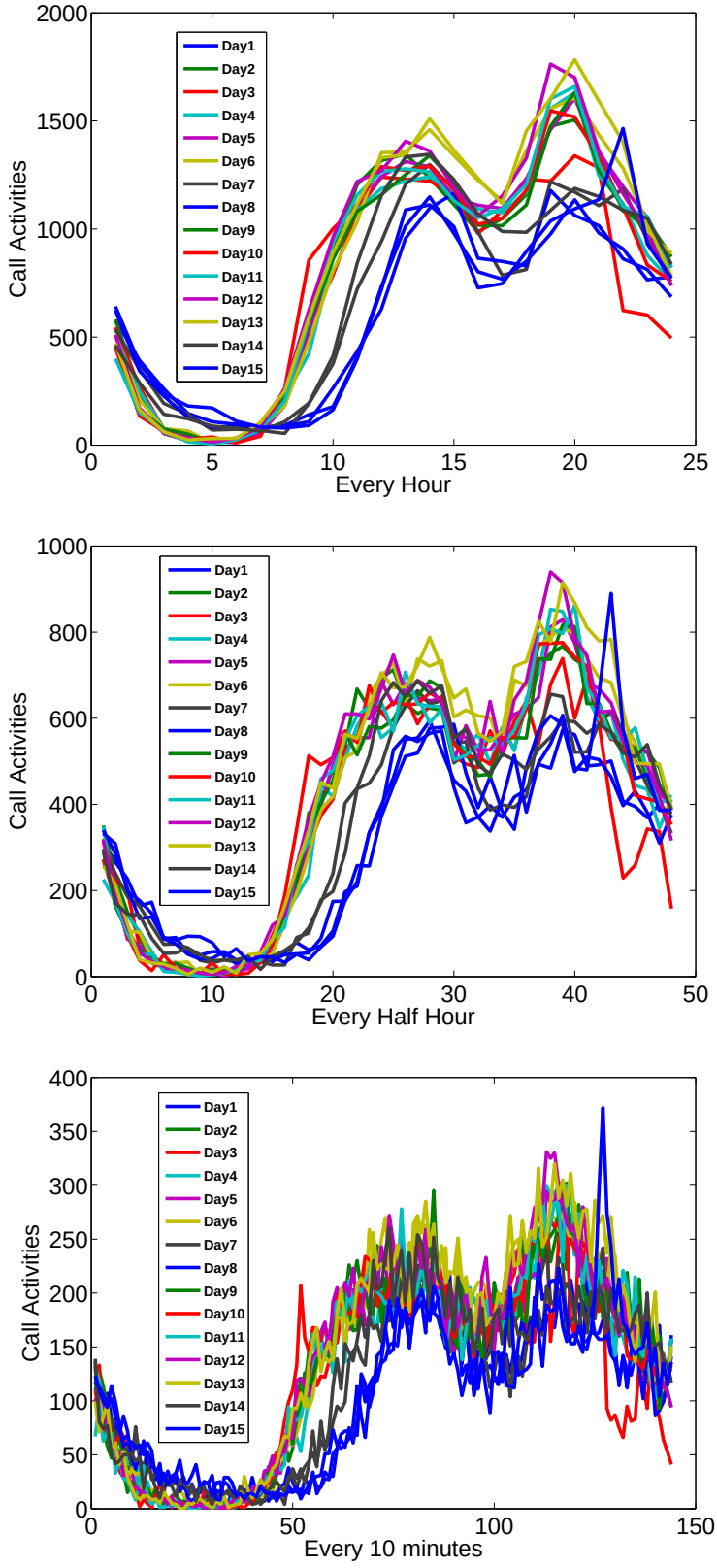


Figure 2. The Small City A's Activities Over a Two-week Time Period

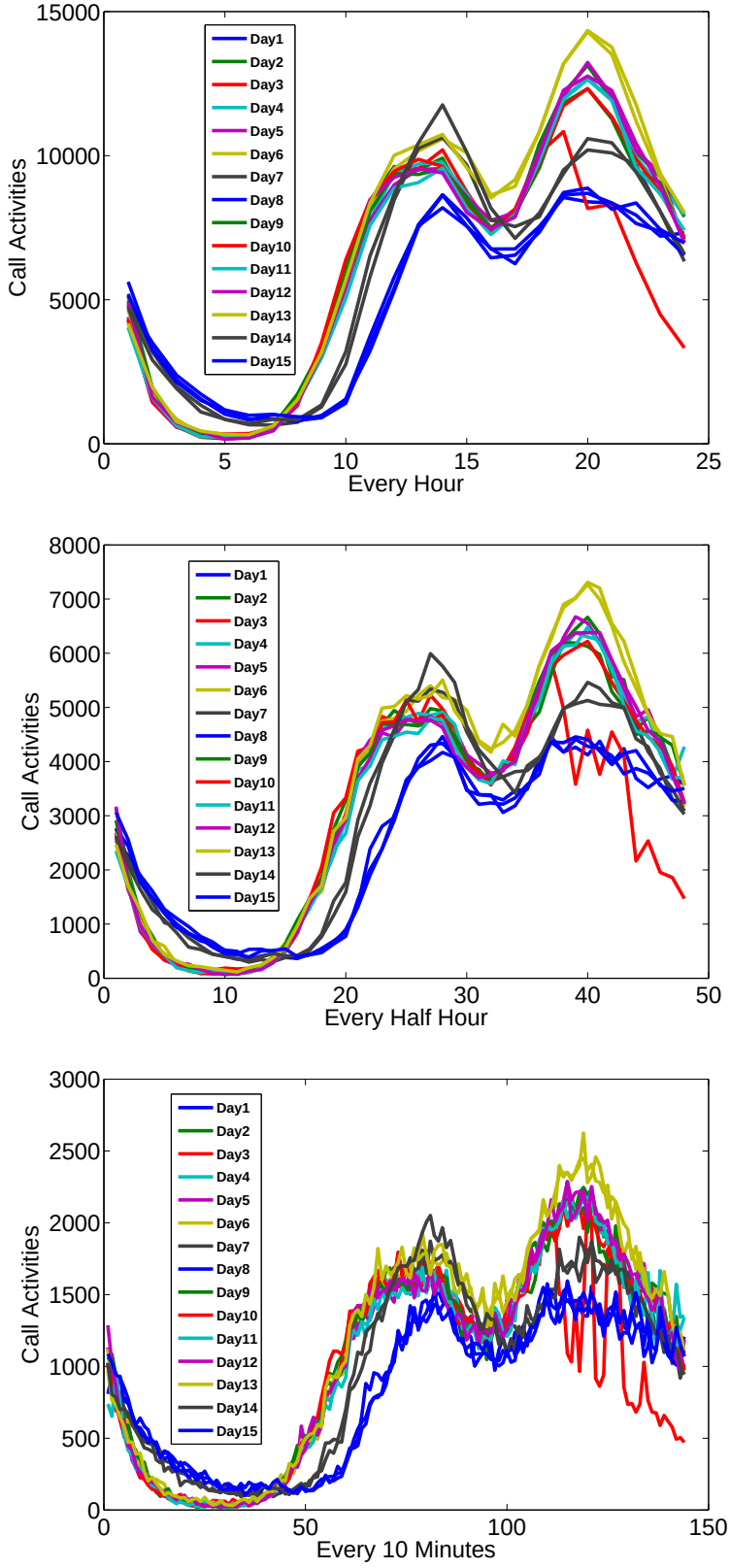


Figure 3. The Large City B's Activities Over a Two-week Time Period

type. Therefore, we use $N_0(t)$ to represent the normal call activities, and $N_A(t)$ to represent all kinds of abnormal call activities caused by known/unknown events. The observed activity $N(t)$ is the process formed by the superposition of unobserved components $N_0(t)$ and $N_A(t)$:

$$N(t) = N_0(t) + N_A(t)$$

Each of these two components can be modulated as a Poisson process, we will give the detailed description in next section.

4.1 Modeling Normal Data

The mass function of the Poisson distribution is given by:

$$P(N;\lambda) = \frac{e^{-\lambda}\lambda^N}{N!} \quad N = 0, 1, ...$$

where λ is the rate parameter representing the average number of occurrences in a fixed time interval. In MMPP, λ is a Markov chain $\lambda(t)$ as a function of time. This is also called a nonhomogeneous Poisson distribution, and $\lambda(t)$ measures the degree of the heterogeneity. The model we apply here is derived from Scott [13] and Ihler [6].

$$\lambda(t) = \lambda_0 \, \delta_{d(t)} \, \eta_{d(t),h(t)}$$

where $d(t) \in [1, 2, ..., 7]$ and associates with Monday(1), ..., Sunday(7), and $h(t)$ indicates the time interval t in, such as 10 minutes, half hour, one hour, etc. Additionally δ and η must meet these constrains:

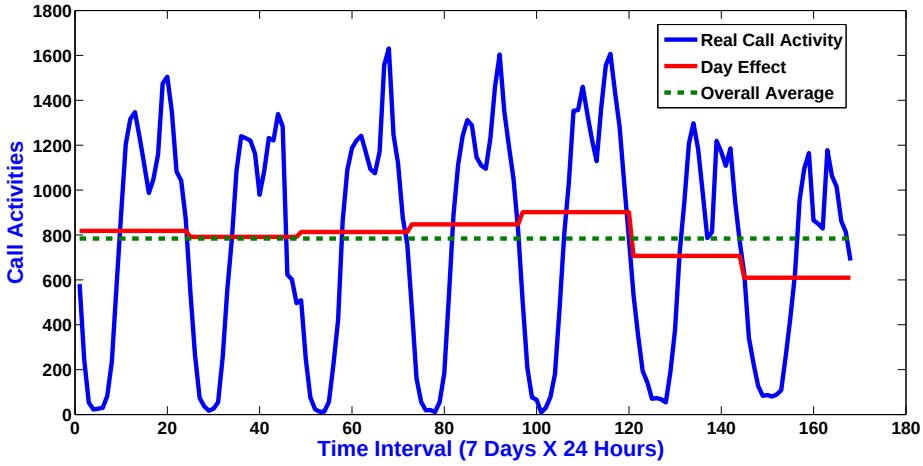
$$\sum_{i=1}^{7} \delta_i = 7$$

and

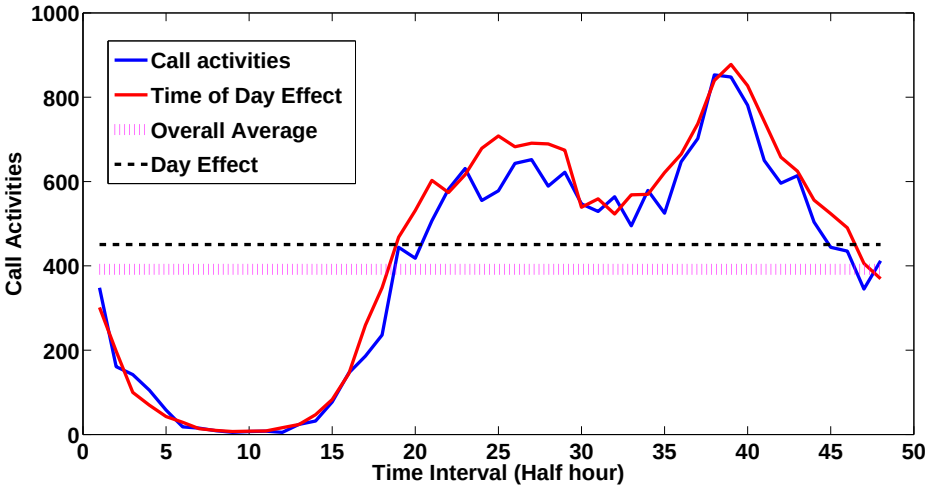
$$\sum_{j=1}^{D} \eta_{i,j} = D, \ \ \forall i$$

Where D is the number of intervals in one day for a given fixed time interval, λ_0 is the average rate of the Poisson process over one week, δ_i is the day effect and $\eta_{i,j}$ is the time of day effect. The day effect δ_i indicates the change in rate over the day of the week, and the time interval effect indicates the change over the time period j on a given day of i. Figure 4(a) and 4(b) demonstrate these two effects.

Figure 4 shows that the simple overall average cannot accurately represent normal call activity. The day of week effect ($\delta_{d(t)}$) can be described more precisely as a representation of the daily behavior. For



(a) Day Effect ($\delta_{d(t)}$): In a one week time period (Monday through Sunday), call activities during the weekday (Monday through Friday) are higher than the weekend day (Saturday and Sunday).



(b) Time of Day Effect ($\eta_{d(t),h(t)}$): Figure 4(a) also shows that different times in each day of the week correspond with different levels of call activity. Call activities are lower during the early morning and early afternoon, and higher in the late morning and night.

Figure 4. Two effects demonstrate the periodic behavior of call activities

example, the weekday displays higher activity than the weekend. However, the day effect does not adequately describe the changes in calling activity over the course of a day. Observation of the data shows that calling activity over the course of a day is periodic in nature, demonstrating less activity from midnight to early morning and higher activity around 1pm and 9pm. In order to compensate for this we add the time of day effect ($\eta_{d(t),h(t)}$) as presented in [13]. The combination of these yields a much more accurate time-dependent value for $\lambda(t)$.

Also suggested by Scott [13], we use conjugate prior distributions for the parameters.

$$\lambda_0 \sim \Gamma(\lambda; a^L, b^L),\ where\ \Gamma(\lambda; a, b) \propto \lambda^{a-1} e^{-b\lambda}$$

$$\frac{1}{7}[\delta_1, \delta_2, ..., \delta_7] \sim Dir(\alpha_1^d, \alpha_2^d, ..., \alpha_7^d,)$$

$$\frac{1}{D}[\eta_{i,1}, \eta_{i,2}, ..., \eta_{i,D}] \sim Dir(\alpha_1^h, \alpha_2^h, ..., \alpha_D^d)$$

$\Gamma(.)$ is the Gamma distribution with mean a/b and variance a/b^2. $Dir(.)$ is a Dirichlet distribution.

4.2 Modeling Anomalous Data

In our applications, we suppose an anomaly is an event that occurs rarely, briefly and randomly. An anomaly causes a change in the call activity over the time period, and is observed in the $N(t)$.

$N_A(t)$ is also a Poisson process whose rate is $\lambda_A(t)$ when there is an anomalous event at time t, and 0 otherwise. In order to modulate the anomalous behavior over time, we can use an unobserved continuous-time Markov process A(t) to determine the existence of any anomalous event at time t, and the probability distribution over A(t) follows the transition probabilities matrix M_A:

$$A(t) = \begin{cases} 1 & \text{an event is occuring at time t} \\ 0 & \text{otherwise} \end{cases}$$

$$M_A = \begin{pmatrix} 1 - A_0 & A_1 \\ A_0 & 1 - A_1 \end{pmatrix}$$

where $1/A_0$ is the expected time interval between events, and $1/A_1$ is the expected length of the event. Our priors for A_0 and A_1 are:

$$A_0 \sim \beta(A, a_0^A, b_0^A) \quad A_1 \sim \beta(A, a_1^A, b_1^A)$$

where $\beta(.)$ is a Beta distribution.

Therefore, the $N_A(t)$ distribution can be written as :

$$N_A(t) \sim \begin{cases} 0 & \text{A(t) = 0} \\ P(N; \lambda_A(t)) & \text{A(t) = 1} \end{cases}$$

and

$$\lambda_A(t) \sim \Gamma(\lambda_A; a^A, b^A),$$

4.3 Expressing MMPP as a Hidden Markov Model (HMM)

The MMPP can be explained as a nonstationary Hidden Markov Model. The observed data are $N(t)$, and the hidden Markov chain is the anomalous event $A(t)$ occurring at time t. After expressing MMPP as HMM, we can use the HMM's recursive procedures to calculate the parameters and posterior distribution of $A(t)$.

4.3.1 Forward recursion

Given the complete data $N_0(t), N_A(t)$ and $A(t)$, it is straightforward to draw posterior samples of the parameters $\lambda(t)$ and A_0, A_1. Also we can infer the posterior distribution over each parameter using Markov chain Monte Carlo methods. The likelihood function is:

$$p(N(t)|A(t)) = \begin{cases} P(N(t); \lambda(t)); & \text{A(t) = 0} \\ \sum_{i=0}^{N(t)} P(i, \lambda(t))\textbf{NBIN}(N(t) - i)); & \text{A(t) = 1} \end{cases}$$

For each $t \in 1, 2, ..., T$, the conditional distribution is:

$$p(A(t)|N(t)) = \pi_0 * \sum_{i=0}^{i=t-1} M_A * p(A(t-1)|N(t-1)) * p(N(t)|A(t))$$

where π_0 is the initial distribution of the Markov chain A(t).

4.3.2 Backward recursion

The $p(A(t|N(t))$ from the forward recursion conditions are on all of the observed data. The backward recursion starts with $p'(A(T)|N(T)) = p(A(T)|N(T))$, for each $t \in T, T-1, ..., 1$, $p'(A(t)|N(t)) = M' * p(A(t)|N(t))$. Then we draw samples:

$$AA(t) \sim p(A(t)|A(t+1) = AA(t+1))$$

.

And for a given AA(t) value, we draw samples of $N_0(t)$ and $N_A(t)$ by using:

$$N_0(t) \propto P(i, \lambda(t))\textbf{NBIN}(N(t) - i));$$

$$N_A(t) = N(t) - N_0(t);$$

4.3.3 Parameter Estimation and Posterior Sampling

Sampling A_0, A_1 is straightforward using:

$$A_{ij} = \sum_{\forall t: A(t)=i, A(t+1)=j} 1; where\ i, j \in 0, 1$$

Therefore the posterior distributions:

$$A_0 \sim \beta(A; a_0^A + A_{01}, b_0^A + A_{00})$$

$$A_1 \sim \beta(A; a_1^A + A_{10}, b_1^A + A_{11})$$

5 Implementation

The MMPP system is written in Matlab 7.0.

5.1 Anomaly Detection

One application of our MMPP framework is detecting anomalous events in an observed data sequence. The existence of an anomaly is represented by the process $A(t)$, and therefore we can use the posterior probability $p(A(t)|N(t))$ as an indicator of anomalies.

As described in the previous section, we draw samples in the MMPP process and predicate the posterior marginal distribution of the anomalies. Figure 5 shows the result of using MMPP for modeling calling activity on a two-week time period. The bottom box shows the posterior probability $p(A(t))$ of anomalous events at each time t. Since we do not have the ground truth of the anomalous events from the observed data sequence, we only give the probability of the events from MMPP in our application.

Figure 6 shows a detailed result on a particular day (Sunday, January 29). The sequence bar of posterior probabilities also demonstrates the duration of the event.

5.2 Different Time Interval

So far, we have used 10 minute intervals to demonstrate our framework. The shorter the interval, the more the data we need to process, and more detailed information we obtain. For the same day (January 29, Sunday), figure 7 shows the MMPP results by using 10-minute, 30-minute and 60-minute intervals. In spite of the difference of the time intervals, the trends of the anomaly and time range of the events are similar. When the interval is shorter, we get more detailed information, and also more anomalous events are detected, such as the peak around 5pm, which is detected using 10-minute intervals, but not detected using 30-minute and 60-minute intervals.

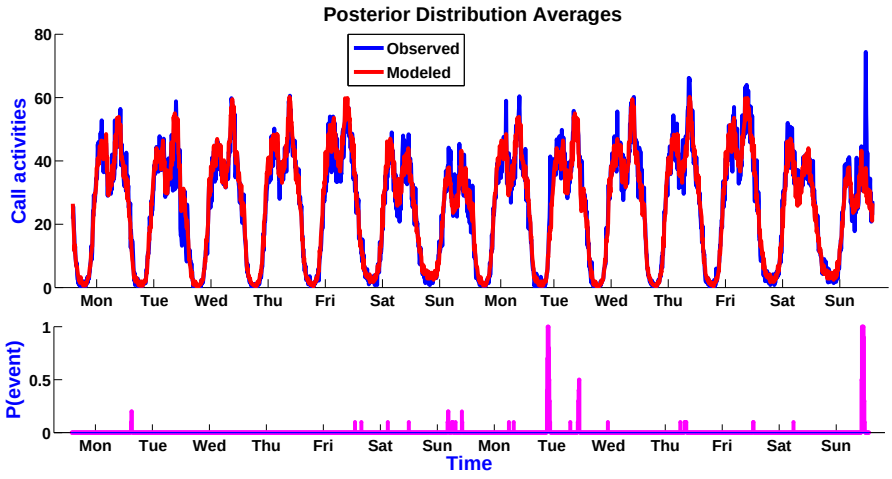


Figure 5. Data for the first two week time period (Jan 16 (Monday) - Jan 29 (Sunday)). The blue curve is generated from the observed data sequence, and the red one is from the sampling data generated by the MMPP model. The posterior probability of anomalous events are shown in the bottom box.

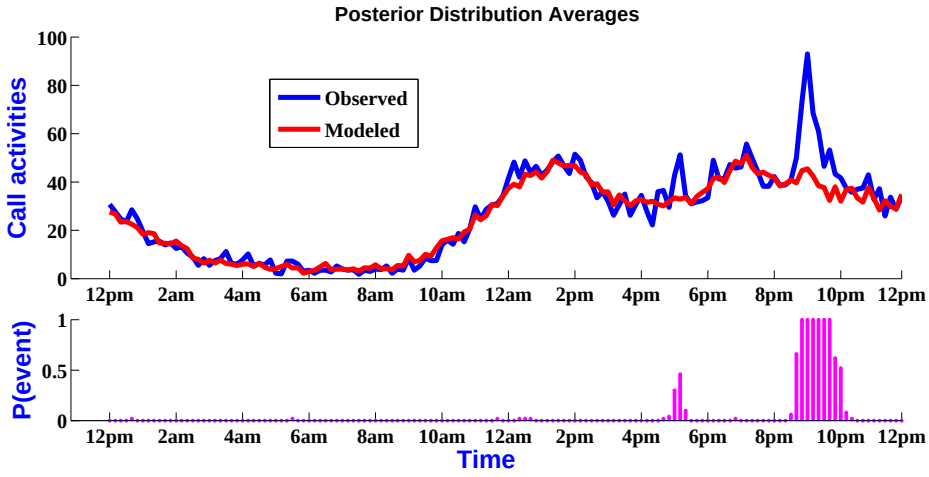
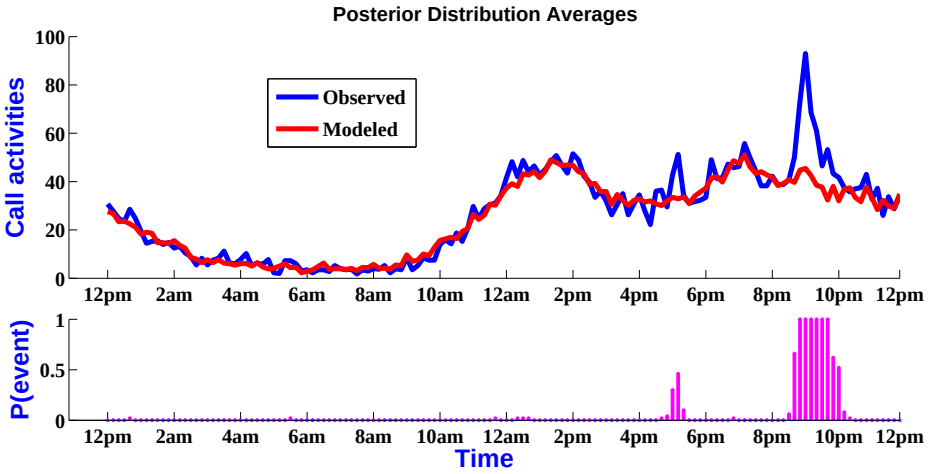


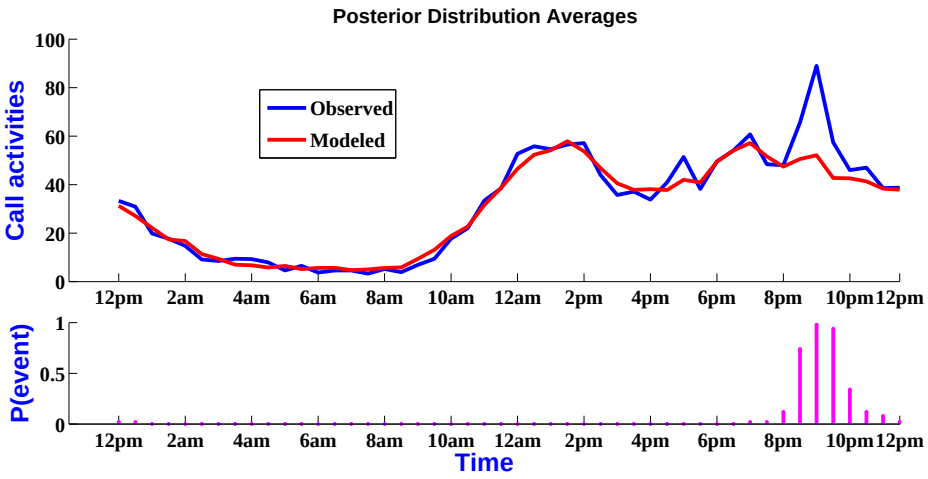
Figure 6. Data for January 29 (Sunday). The blue curve is generated from the observed data sequence, and the red one is from the sampling data generated by the MMPP model. The posterior probability of anomalous events are shown in the bottom box.

6 Discussion and Conclusion

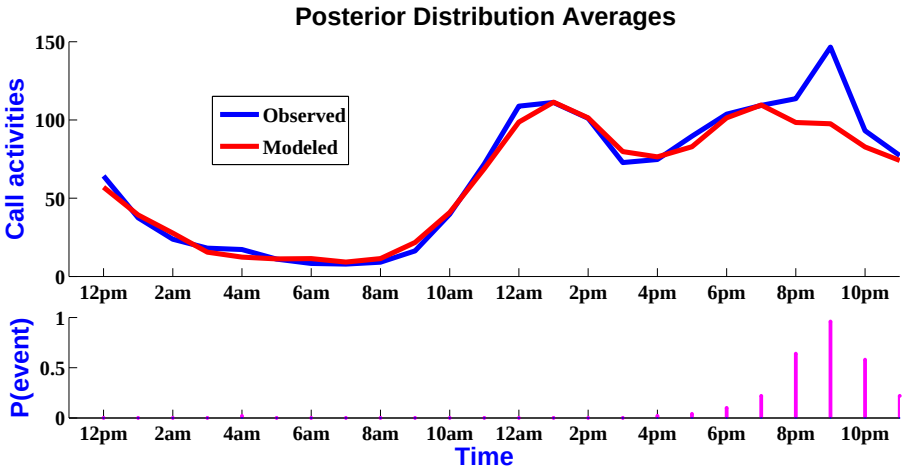
Data from the cell phone network reflects the behaviors and activities of human society, which occur on daily and weekly schedules, with variations. Out of this periodic, varying data, Emergency and Crisis Responders and Planners would like to be able to detect crisis events, however, simple stochastic



(a) Time Interval = 10 minutes



(b) Time Interval = 30 minutes



(c) Time Interval = 60 minutes

Figure 7. MMPP Results vs. Time Interval

models that rely on aggregate statistics are not able to differentiate between normal daily variations and legitimate anomalous (and potentially crisis) events. The Markov-Modulated Poisson Process provides a method of modeling call activity that varies on several periodic scales and allows anomalous events to be differentiated from random variations.

In the examples we show, the MMPP model yields a posterior probability of anomalous events. Careful examination of the plot of the posterior probability in conjunction with expected and observed activity show the MMPP model to be tolerant of small variations in the activity level but able to distinguish anomalous behavior both in short duration (events lasting less than a minute) and longer duration events (lasting up to 30 minutes).

7 Future Work

This paper and previous work by other authors in the area demonstrate the power and effectiveness of the Markov-Modulated Poisson Process model for the detection of anomalous events in time series data that varies according to periodic effects on multiple time scales. Now that its effectiveness has been shown, it remains to implement the MMPP model as part of a real time system that operates on streaming data. This implementation can be relatively straightforward, but we envision the need to develop a more flexible model that can continually be updated to reflect concept drift in the data stream.

The final anomaly detection model will be incorporated into a service that will be integrated with the WIPER system. As a component in the WIPER system, it will monitor a real time stream of call data and anomalies will be flagged and conveyed to the end user via a web-based console.

Acknowledgements

The research work is supported by the National Science Foundation, the DDDAS Program, under Grant No. CNS-050348. The authors would also like to thank Steven L. Scott and Alexander T. Ihler for their support regarding the MMPP implementation, and useful discussions about this area.

References

[1] D. Agarwal. Detecting anomalies in cross-classified streams: a bayesian approach. *Knowledge and Information Systems*, 11(1):29 – 44, 2007.

[2] S. Basu and M. Meckesheimer. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, 11(2):137 – 154, 2007.

[3] D. R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society, Series B*, (2):129 – 164, 1955.

[4] V. Guralnik and J. Srivastava. Event detection from time series data. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–42, 1999.

[5] D. M. Hawkins. *Identification of outliers*. Chapman and Hall New York, 1980.

[6] A. T. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 207–216, 2006.

[7] E. Keogh, S. Lonardi, and B. Chiu. Finding surprising patterns in a time series database in linear time and space. In *Proceedings of The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*, July 2002.

[8] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, 2002.

[9] E. G.-P. na and L. E. Nieto-Barajas. Bayesian nonparametric inference for mixed poisson processes. *Bayesian Statistics 7*, pages 163–179, 2003.

[10] T. Schoenharl, R. Bravo, and G. Madey. Wiper: Leveraging the cell phone network for emergency response, 2007.

[11] T. Schoenharl, G. Madey, G. Szabó, and A.-L. Barabási. Wiper: A multi-agent system for emergency response. In *Proceedings of the Third International ISCRAM Conference*, May 2006.

[12] M. Schonlau and M. Theus. Detecting masquerades in intrusion detection based on unpopular commands. *Information Processing Letters*, 76(1-2):33–38, 2000.

[13] S. Scott. *Bayesian methods and extensions for the two state Markov modulated Poisson process*. PhD thesis, Department of Statistics, Harvard University, 1998.

[14] S. L. Scott. A bayesian paradigm for designing network intrusion systems. *Computational Statistics and Data Analysis*, 45(1):69 – 83, 2004.

[15] S. L. Scott and P. Smyth. The markov modulated poisson process and markov poisson cascade with applications to web traffic data. *Bayesian Statistics 7*, pages 671–680, 2003.

[16] W. S. Wei. *Time Series Analysis: Univariate and Multivariate Methods (Second Edition)*. Addison-Wesley, 2006.